AIME-Con 2025

Artificial Intelligence in Measurement and Education Conference (AIME-Con)

Volume 1: Full Papers

October 27-29, 2025

The AIME-Con organizers gratefully acknowledge the support from the following sponsors.

Platinum

Gold

Silver

Supporters

duolingo english test

The future of language assessment is here

The Duolingo English Test is a computer adaptive test powered human-in-the-loop AI and supported by rigorous validity research. The test measures speaking, writing, reading, and listening skills, providing a deeper insight into English proficiency.

Built on the latest language assessment science

- Accessible by design, supporting test takers wherever they are for just $70
- Built on rigorous research and industry- leading security
- Integrates the latest assessment science and AI for accurate results
- Accepted by over 5,800 programs worldwide

englishtest.duolingo.com

Evidence-based approach to AI in Measurement & Learning

At the intersection of artificial intelligence and educational measurement, Pearson stands as your trusted partner—delivering clarity, confidence, and innovation in every assessment moment.

Why Pearson?
- AI-Enhanced Accuracy: Using automated scoring and predictive analytics to provide insights that are accurate, fair, and timely.
- Future-Ready Solutions: Platforms that evolve with policy, pedagogy, and technology.
- Personalized Learning Journeys: Multi-lingual access and adaptive item generation to support each student's unique growth trajectory.
- Ethical AI Practices: Commitment to data security, transparency, explainability, and bias mitigation.
- Collaborative Innovation: Partnering with educators, researchers, and technologists to shape the future of assessment.

Human-Centric AI	Pearson believes AI's highest purpose is to elevate and empower human capabilities.
Assessment as a Learning Continuum	We reimagine assessments not as endpoints, but as integral parts of the learning journey.
AI as an Environment	Pearson is exploring how this shift impacts our approach to assessment—ensuring our tools are adaptive and future-ready.
Balancing Vision and Capabilities	We deliver reliable solutions today while building toward the future of AI in education.

Pearson

Order copies of this and other NCME proceedings from:

National Council on Measurement in Education (NCME)
520 S. Walnut St. Box 2388
Bloomington, IN 47402
USA
Tel: +1-812-245-8096
ncme@ncme.org

ISBN 979-8-218-84228-4

Preface

Introduction

The inaugural NCME-sponsored Artificial Intelligence in Measurement and Education Conference (AIME-Con) brought together an interdisciplinary community of experts working at the intersection of artificial intelligence (AI), educational measurement, assessment, natural language processing, learning analytics, and technological development. As AI continues to transform education and assessment practices, this conference provided a critical platform for fostering cross-disciplinary dialogue, sharing cutting-edge research, and exploring the technical, ethical, and practical implications of AI-driven innovations in measurement and education. By bringing together experts from varied domains, the conference fostered a rich exchange of knowledge to enhance the collective understanding of AI's impact on educational measurement and evaluation.

Conference Theme - Innovation and Evidence: Shaping the Future of AI in Educational Measurement

The NCME-Sponsored AIME-Con focused on how rigorous measurement standards and innovative AI applications can work together to transform education. With sessions spanning summative large-scale assessment, formative classroom assessment, automated feedback, and informal learning tools, this conference fostered both the advancement and evaluation of AI technologies that are effective, reliable, and fair.

The National Council on Measurement in Education

The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications of educational measurement to benefit society. A professional organization for individuals involved in assessment, evaluation, testing, and other aspects of educational measurement, our members are involved in the construction and use of standardized tests; new forms of assessment, including performance-based assessment; program design; and program evaluation. Learn more about NCME, including our goals and our leadership, at www.ncme.org. We are grateful to the NCME.

NCME Special Interest Group on Artificial Intelligence in Measurement and Education

The AIME SIGIMIE seeks to advance the theoretical and applied research into AI of educational measurement by bringing together data scientists, psychometricians, education researchers, and other interested stakeholders. The SIGIMIE will discuss current practices in using Generative AI, approaches to evaluate their precisionaccuracy, and areas where more foundational research is required into the way we test and measure educational outcomes. This group seeks to create a strong professional identity and intellectual home for those interested in the use of AI in many areas, including automated scoring, item evaluation, validity studies, formative feedback, and generative AI for automated item generation.

Proposal Requirements and Review Process for Full Papers

AIME-Con invited submission of "Full Papers", which were submissions of up to six pages (excluding references, tables, and figures), prepared using the ACL LaTeX or Word templates. These papers presented completed research or theoretical work intended for inclusion in the published conference proceedings. Submissions included a title ($\leq$ 12 words), a brief abstract ($\leq$ 50 words), a designated topic of interest, and the full paper. **Submissions were blinded for peer review.**

Submissions were evaluated by members of the review committee using a rubric that evaluated the following dimensions:

- **Relevance and community impact:** pertinence to the AI in measurement and education community, and potential contribution to current discussions and challenges in the field
- **Significance and value:** scholarly merit or practical importance of the work, and potential impact on theory, practice, or policy
- **Methodological rigor:** coherence and appropriateness of the proposed methods, techniques, and approaches; and soundness of the overall research design
- **Quality of expected outcomes:** whether the proposed analysis and interpretation methods are appropriate, and the potential contribution to knowledge in the field
- **Feasibility and timeline:** the realistic likelihood that the proposed work can be completed by the conference date

For the purposes of this conference, "AI" was defined broadly to include rule-based methods, machine learning, natural language processing, and generative AI/large language models. Reviewers provided constructive feedback and overall recommendations to ensure that accepted sessions reflected both scholarly merit and practical value to the AI in measurement and education community.

Organizing Committee

NCME Leadership

Amy Hendrickson, Ph.D. (President)
Rich Patz, Ph.D. (Executive Director)

Conference Chairs

Joshua Wilson, University of Delaware
Christopher Ormerod, Cambium Assessment
Magdalen Beiting Parrish, Federation of American Scientists

Proceedings Chair

Nitin Madnani, Duolingo

Proceedings Committee

Jill Burstein, Duolingo
Polina Harik, NBME

Program Committee

Conference Chairs

Joshua Wilson, University of Delaware
Christopher Ormerod, Cambium Assessment
Magdalen Beiting Parrish, Federation of American Scientists

Reviewers

Ketan , University of Massachusetts, Amherst
Hope Adegoke, University of North Carolina
Tazin Afrin, NBME
Ernest Amoateng, Western Michigan University
Kylie Anglin, University of Connecticut
Sergio Araneda, Caveon
Meirav Attali, Fordham University
Nurseit Baizhanov
Lee Becker, Pearson
Beata Beigman Klebanov, ETS
Ummugul Bezirhan, Boston College
Janet Shufor Bih Epse Fofang, University of Pittsburgh
Peter Bodary, University of Michigan School of Kinesiology
Brad Bolender, Finetune by Prometric
Jill Burstein, Duolingo
Hye-Jeong Choi, HumRRO
Jinmin Chung, Univ. of Iowa
Christina Cipriano, Yale University
Lisa Clark, City University of New York
Victoria Delaney, San Diego State University
Onur Demirkaya, Riverside Insights
Scott Elliot, SEG Measurement
Andrew Emerson, National Board of Medical Examiners
Mingyu Feng, WestEd
Taiwo Feyijimi, University of Georgia
Carla Firetto, Arizona State University
Jonathan Foster, University at Albany
Samantha Goldman, The University of Kansas
Chad Green, Loudoun County Public Schools
Joe Grochowalski, College Board
Yi Gui, The University of Iowa
Aysegul Gunduz, University of Alberta
Hongwen Guo, ETS Research Institute
Yage Guo, Center for Applied Linguistics
Gulsah Gurkan, Pearson
Suhwa Han, Cambium Asessment
Michael Hardy, Stanford University
Qiwei He, Georgetown University
Alexander Hoffman, AleDev Research & Consulting
Ruikun Hou, Technical University of Munich

Ruiping Huang, University of Illinois Chicago
Yue Huang, Measurement Incorporated
Hiu Ching Hung, Friedrich-Alexander-Universität Erlangen-Nürnberg
HUIMIN JIAO
Jamie Jirout, University of Virginia
Ji Yoon Jung, Boston College
Olasunkanmi Kehinde, Norfolk State University
YoungKoung Kim, The College Board
Becky King, University of Pittsburgh
Miryeong Koo, University of Illinois at Urbana-Champaign
Aakash Kumar, Texas A&M University
Alexander Kwako, Cambium Assessment
Brandon LeBeau, WestEd
Hansol Lee, Stanford University
Arun Balajiee Lekshmi Narayanan, University of Pittsburgh
Hongli Li, Georgia State University
Tianwen Li, University of Pittsburgh
Li Liang
Boyuan LIU, Department of Educational Psychology, The Chinese University of Hong Kong
Chen Liu, UC Merced
Will Lorie
Susan Lottridge, Cambium Assessment
Max Lu, Harvard University
Yi Lu, Federation of State Boards of Physical Therapy
Wenchao Ma, University of Minnesota
Henry Makinde, University of North Carolina - Greensboro
Mike Maksimchuk, Kent Intermediate School District
Salih Mansur, Touro University of New York
Jamie Mikeska, ETS
Mubarak Mojoyinola, University of Iowa
Wesley Morris, Vanderbilt University
Tim Moses, Buros Center for Testing
William Muntean, National Council of State Boards of Nursing
Mariel Musso, University of Granada- CONICET
Supraja Narayanaswamy, Acelero Inc.
Lynn Nguyen, Fruitions eTutoring
Tram-Anh Tran Nguyen, University of Massachusetts, Amherst
Chunling Niu, The University of the Incarnate Word
Kai North, Cambium Learning Group, Inc.
Teresa Ober, ETS
Maria Oliveri, Purdue University
Christopher Ormerod, Cambium Assessment
Jay Parkes, University of New Mexico
Hallie Parten, University of Virginia
Katie Pedley, Pearson
Benjamin Pierce, University of Pittsburgh
Andrew Potter, Arizona State University
Sonya Powers, Edmentum
Ricardo Primi, Universidade São Francisco
Sarah Quesen, WestEd
Ruchi Sachdeva, Pearson

Fariha Hayat Salman, American University in Dubai
Lydia Scholle-Cotton, Queen's University (Kingstion, ON, Canada)
Qingzhou Shi, Northwestern University
Jinnie Shin, University of Florida
Anthony Shiver, Law School Admission Council
Stephen Sireci, University of Massachusetts Amherst
Anastasia Smirnova, San Francisco State University
Xiaomei Song, Case Western Reserve University School of Medicine
Kayden Stockdale, Virginia Tech
Caitlin Tenison, ETS
Danielle Thomas, Carnegie Mellon University
Zewei Tian, University of Washington
Nhat Tran, University of Pittsburgh
FELIPE Valentini, Graduate School of Psychology, Universidade São Francisco
Marcus Walker, National Commission on Certification of Physician Assistants
Cole Walsh, Acuity Insights
Huanxiao Wang, University of Pennsylvania
Yun-Han Weng, Ohio State University
Joshua Wilson, University of Delaware
Sirui Wu, University of British Columbia
Hyesun You, University of Iowa
Meltem Yumsek Akbaba, Ministry of National Education, Turkey
Diego Zapata-Rivera, ETS
Dake Zhang, Rutgers University
Jiayi (Joyce) Zhang, University of Pennsylvania
Liang Zhang, University of Georgia
Ting Zhang, American Institutes for Research
Lauren Zito, WGU Labs

Table of Contents

Input Optimization for Automated Scoring in Reading Assessment

Ji Yoon Jung **Ummugul Bezirhan** **Matthias von Davier**
TIMSS & PIRLS International Study Center at Boston College
{jiyoon.jung, bezirhan, vondavim}@bc.edu

Abstract

This study examines input optimization for enhanced efficiency in automated scoring (AS) of reading assessments, which typically involve lengthy passages and complex scoring guides. We propose optimizing input size using question-specific summaries and simplified scoring guides. Findings indicate that input optimization via compression is achievable while maintaining AS performance.

1 Introduction

Automated scoring (AS) has a rich history in educational measurement (Lottridge et al., 2023), dating back to the 1960s when the primary focus was on scoring multiple-choice responses or implementing machine-supported scoring based on pattern matching or manual feature selection. The rapid advances in natural language processing (NLP), machine learning, and computational power have led to significant developments in large language models (LLMs). Integrating LLMs, such as OpenAI's GPT models or META's Llama, into AS expands the applicability and scalability of AS in educational assessment.

However, applying LLMs to the AS of reading assessments presents unique challenges in processing long inputs, including extended reading passages and complex scoring guides (SGs). Given that the cost of using LLMs through APIs depends on the number of input, cached, and output tokens (OpenAI, 2025), extensively long prompts can lead to inflated costs for each API call. Moreover, previous study indicated that long prompts can cause a "lost in the middle" effect, where LLMs struggle to appropriately use the most relevant context embedded within the extensive input (Liu et al., 2023). This limitation persists, particularly for smaller models operated locally.

To address the challenge of processing long inputs, we propose input optimization to improve the scalability and efficiency of AS in international large-scale assessments (ILSAs).

2 Background

Very long inputs can slow LLMs' inference processes and increase energy use due to the increased the number of tokens that need to be processed. Prior research showed that LLMs do not robustly utilize information in long input contexts and may ignore parts of the given context, generating incorrect outputs (Liu et al., 2023). Crucially, extended input lengths lead to a linear increase in both computational costs and energy demands (Poddar et al., 2025).

Text compression shrinks textual data while preserving crucial information, improving storage and computational efficiency, and enhancing the performance of LLMs (Rahman et al., 2024; Wang et al., 2024). Compression can be achieved through either soft or hard prompts. Soft prompts are continuous vectors, enabling LLMs to address long and complex input by distilling critical information into a smaller number of special tokens (Li et al., 2024; Wang et al., 2024). Yet, soft prompts are less interpretable by humans and are often highly customized to specific tasks. Their reusability or transferability across different tasks can be constrained (Su et al., 2022).

In contrast, hard prompts comprise discrete words and tokens, making them easily understandable by humans. This readability and transparency allow humans to review, debug, and modify prompts by facilitating effective human-machine interaction (Chang et al., 2024; Wen et al., 2023). Hard prompts can be especially powerful when prompts need human interpretation or are integrated into a text-based interface (Wen et al., 2023; Jiang et al., 2023). Zhang et al. (2024) found that hard prompts yield superior performance for

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 1–8
October 27-29, 2025

summarization compared to soft prompts in human evaluations.

Despite the demonstrated usefulness of text compression techniques, they have not been widely integrated into AS for reading assessments in ILSAs, such as the Progress in International Reading Literacy Study (PIRLS). Optimizing long input through compression in reading assessments can contribute to improving AS scalability and cost- and computational efficiency in ILSAs. This paper examines how advances in hard prompt-based input optimization can be integrated into AS in PIRLS, which involves a substantial volume of multilingual responses.

3 Method

3.1 Dataset

The PIRLS, administered every five years since 2001, assesses the reading comprehension skills of fourth-grade students across 50-60 countries worldwide. In PIRLS 2021, approximately 50% of countries (27 countries) used computer-based assessments. The assessment framework categorizes reading comprehension into four cognitive processes: focus on and retrieve; straightforward inferences; interpret and integrate; and evaluate and critique (Mullis & Martin, 2019). For this study, we selected five one-point constructed response (CR) items from the PRILS 2021 digital assessment (digital PIRLS). The selected items represent three cognitive processes: one from focus on and retrieve, two from straightforward inferences, and two from interpret and integrate.

These items are "trend" items, kept secure for their reuse in future assessment cycles (Fishbein et al., 2024). We provide general descriptions of these items (Table 1) as this research is part of the preparatory work for AS in PIRLS 2026, where these items will be used. We selected four reading passages with varying difficulty levels: easy (passages B and D), medium (passage A), and difficult (passage C).

Item	Passage	Process	*n*
1	A	Focus on and retrieve	2687
2	B	Straightforward inferences	2951
3	A	Straightforward inferences	2643
4	C	Interpret and integrate	2589
5	D	Interpret and integrate	2452

Table 1: PIRLS trend items used in the study

The dataset included multilingual responses from the 27 participating countries in digital PIRLS 2021, covering 29 languages. While approximately 50% of participating countries used computer-based assessments in PIRLS 2021, the data still contained on average, 2,664 multilingual responses per item (see Appendix A). We used a randomly selected 20% subset for each country given the scope, computational and budgetary limitations.

3.2 PIRLS Scoring Template

We proposed a generalized PIRLS scoring template for AS (see Appendix B), comprising four key elements: (1) instruction, (2) reading passage, (3) question, and (4) SG, as detailed in Table 2. We used GPT-4.1 (i.e., gpt-4.1-2025-04-14) for our AS

Component	Content
Instruction	Comprehensive guidance on AS
Reading passage	A written text serving as the stimulus
Question	A question consisting of one or two sentences
Scoring guide (SG)	Rubric for scoring an item, including descriptions and examples

Table 2: PIRLS scoring template components

implementation, applying parallel processing for efficiency. This template used zero-shot chain-of-thought (CoT), a technique that enhances LLM performance through step-by-step reasoning without requiring specific examples (Kojima et al., 2022; Yuan et al., 2024). Zero-shot COT offers the advantage of easy generalization to other items due to its independence from specific examples.

Instruction: The instruction component offers comprehensive guidance on translating student responses, applying the SG, validating scores, and constructing output.

Reading Passage: The second component, a reading passage, could be presented as either the original passage or a question-specific summary. Original passages provide the complete information as presented to students, whereas summaries include question-relevant details while preserving overall context.

Question: The third component, the item's question, was directly input into the scoring template.

Scoring Guide (SG): The SG could be either the original SG or a simplified version. Simplified SGs were designed to mitigate challenges arising from ambiguous structure or meaning in the original SGs, which may lead to less accurate output from LLMs. Prior studies (Keluskar & Bhattacharjee, 2024; Kamath et al., 2024) indicate that rephrasing or clarifying sentences in prompts can significantly improve LLM output quality.

3.3 Input Optimization

Question-specific Summary: The passage summarization prompt shown in Figure 1 was used to generate question-specific summaries that retain all essential information needed to answer the question while maintaining the overall flow. Query-based text summarization aids users in accessing specific information within lengthy texts, enabling LLMs to provide efficient access to relevant content (Yu & Han, 2022; Zhang et al., 2025). This zero-shot CoT prompt can be applied across various items, requiring only the [[question]] input to be modified.

> Summarize the passage for a fourth-grade student, including the overall flow and all necessary information to correctly answer the question: [[question]]
>
> 1. Read the Passage: Carefully read the passage to understand the main events and details.
> 2. Summarize: Create a summary that includes the overall flow, and the necessary information related to the question.
> 3. Final Output
>   - The output should be a coherent paragraph summarizing the passage.
>   - Avoid new section headings.

Figure 1: Passage summarization prompt

Simplified SG: Original SGs for one-point items in PIRLS 2021 consist of two parts: a description with examples of acceptable responses, and a description with examples of unacceptable responses. For the simplified SG, we utilized GPT-4.1 to improve the readability of acceptable response descriptions from the original SGs. This involved rephrasing or reconstructing sentences and removing examples, guided by the SG modification prompt (Figure 2). For unacceptable response descriptions, we adopted a standard description: "Assign this score if the response does not explicitly include the key content described in the [Score: 1] criteria." Replacing the original item-specific descriptions.

Additionally, we incorporated notes reflecting the general guidelines of the PIRLS Scoring Guides: "(1) Minor irrelevant details are permissible only if the response explicitly includes the key content required for [Score: 1] and the details do not contradict the [Score: 1] criteria. (2) Character names may vary depending on the language used; such variations should not affect scoring."

> Improve the language in the current scoring guide.
>
> # Steps
> 1. Review the Scoring Guide: Carefully read the existing scoring guide to grasp its content and scoring criteria.
> 2. Refine Language: Enhance the language for clarity while keeping the intended meaning of the original scoring guide.
> 3. Final Output: Produce the final output in plain text.
>
> # Output Format
> - Use bullet points if they improve readability.
> - Maintain the given structure: "**[Score: X]**: Assign this score if …"
> - Avoid new section headings or providing examples.

Figure 2: SG modification prompt

3.4 AS with PIRLS Scoring Template

We ran two separate AS models using the PIRLS scoring template: a baseline model and an optimized AS model with compression (Opt-AS). The baseline model used the original reading passages and SGs, while the Opt-AS model integrated question-specific summaries and simplified SGs. For each item, a single summary and simplified SG were created and consistently applied to all responses. Following the Opt-AS, custom Python scripts were utilized to automatically identify and correct mis-formatted outputs to ensure a consistent format.

3.5 Evaluation Metrics

We evaluated AS performance using four metrics: compression ratio, exact agreement (EA), and Cohen's Kappa (κ).

Compression ratio quantifies the efficiency of our input optimization by comparing the token count of optimized inputs to that of original inputs. We specifically focused on the token reduction in reading passages and SGs, where lower values indicate higher compression. For SGs, notes reflecting the general guidelines of the PIRLS Scoring Guides were excluded from the compression ratio calculation.

$$R = \frac{Token\ count\ of\ optimized\ input}{Token\ count\ of\ original\ input} \quad (1)$$

EA, a commonly used metric in AS, is calculated as the percentage of exact matches between human and machine scores.

Cohen's Kappa (Cohen, 1960) measures inter-rater reliability by considering chance agreement, and is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

where p_o is the observed agreement among raters, and p_e denotes the expected probability of chance agreement. The Kappa ranges from 0 (agreement due to chance) to 1 (perfect agreement).

We computed processing time and estimated costs for Opt-AS using Python scripts. Cost estimates were based on the number of input and output tokens, following the GPT-4.1 API pricing (OpenAI, n.d.): $2.00 per million input tokens and $0.80 per million output tokens. One million tokens are approximately equivalent to 750,000 words.

4 Results

Compression Ratio: Tables 3 and 4 present token counts and compression ratios. On average, passages were compressed to 20.22% of the original length, while SGs were reduced to 46.47% of their original size.

	Baseline		Opt-AS	
Item	Passage	SG	Passage	SG
1	724	112	168	67
2	581	119	117	93
3	724	152	155	65
4	1045	163	168	79
5	640	261	143	71
Avg.	743	161	150	75

Table 3: Token count for passage and SG

Item	Passage	SG
1	23.20%	59.82%
2	20.14%	78.15%
3	21.41%	42.76%
4	16.08%	48.47%
5	22.34%	27.20%
Avg.	20.22%	46.47%

Table 4: Compression ratio

EA & Kappa: Our Opt-AS model demonstrated comparable performance to the baseline model, achieving an average EA of 95.16% and kappa of 0.8852. Notably, for Item 1, the Opt-AS model yielded a lower kappa of 0.8482 compared to the baseline (0.9308). This discrepancy can be attributed to Item 1 being a very easy item, resulting in highly imbalanced data where 91.9% of responses received a human score of 1. Despite this, Opt-AS maintained strong precision and recall values of 98.55% and 98.34%, respectively (see confusion matrices in Appendix C).

Item	Baseline		Opt-AS	
	EA	κ	EA	κ
1	98.78%	0.9308	97.18%	0.8482
2	96.13%	0.9203	96.35%	0.9231
3	94.35%	0.8609	94.47%	0.8750
4	93.64%	0.8706	93.48%	0.8768
5	93.27%	0.8570	93.50%	0.8511
Avg.	95.16%	0.8852	94.94%	0.8723

Table 5: EA & Kappa

Processing Time & Cost: The average processing time and cost per item using Opt-AS were approximately 6 minutes and $3.09, respectively (see Table 6). In contrast to the extensive resources required for human rater training and scoring (Ward & Bennett, 2012), this reflects a highly efficient use of time and cost. Moreover, our Opt-AS reduced costs by nearly 50% relative to the baseline model, which incurred approximately $6 per item and required around 7 minutes of processing time.

Item	Processing Time	Cost ($)
1	00:06:05	3.170
2	00:07:17	3.390
3	00:05:59	2.755
4	00:06:32	3.210
5	00:06:19	2.907
Avg.	00:06:26	3.087

Table 6: Processing time & cost

5 Discussion

Our findings indicate that input optimization significantly reduces the complexity of AS in reading assessments. Aligned with prior research (Jiang et al., 2023; Xu & Lapata, 2022), Opt-AS leverages compression techniques to optimize input size, substantially shortening text length while preserving critical information. This optimization effectively lowers computational costs without compromising AS performance, even on low-resource languages such as Arabic, Croatian, and Maltese. Given the considerable cost and time involved in scoring over 12,000 multilingual written responses per CR item in PIRLS, and the shift to fully digital assessment for all participating countries in PIRLS 2026 (von Davier & Kennedy, 2024), Opt-AS offers a cost-effective, energy-efficient, and scalable scoring solution in a computer-based assessment context.

Despite these promising results, this study has limitations. First, due to its exploratory nature, the analysis was conducted on a randomly selected 20% sample. While this sample was representative, future research should assess the generalizability of our approach using the full PIRLS dataset across a broader range of CR items. Next, further investigation into AS consistency is necessary. Although GPT-4.1's temperature was set to 0 to minimize variability, validating the consistency of both AS and human scoring remains important. One potential method is to use sentence embedding techniques to cluster semantically similar responses, allowing for a systematic evaluation of scoring consistency across both scoring methods.

6 Conclusion

This study provides compelling evidence for the effectiveness of input optimization for AS in multilingual reading assessments. Our Opt-AS approach maintained robust performance within the PIRLS framework, concurrently saving time, cost, and computational burden. The streamlined AS enhances operational efficiency and scalability across a multitude of assessment items and countries. Ultimately, well-implemented AS systems promise to deliver timely, accurate, and reliable reporting to participating countries, supporting more informed educational policy decisions.

References

Chang, K., Xu, S., Wang, C., Luo, Y., Liu, X., Xiao, T., & Zhu, J. (2024). Efficient prompting methods for large language models: A survey. arXiv preprint arXiv:2404.01077.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Fishbein, B., Yin, L., & Foy, P. (2024). PIRLS 2021 User Guide for the International Database (2nd ed.). Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/data

Jiang, H., Wu, Q., Lin, C. Y., Yang, Y., & Qiu, L. (2023). Llmlingua: Compressing prompts for accelerated inference of large language models. arXiv preprint arXiv:2310.05736.

Kamath, G., Schuster, S., Vajjala, S., & Reddy, S. (2024). Scope ambiguities in large language models. Transactions of the Association for Computational Linguistics, 12, 738-754.

Keluskar, A., Bhattacharjee, A., & Liu, H. (2024, December). Do LLMs Understand Ambiguity in Text? A Case Study in Open-world Question Answering. In 2024 IEEE International Conference on Big Data (BigData) (pp. 7485-7490). IEEE.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Li, Z., Liu, Y., Su, Y., & Collier, N. (2024). Prompt compression for large language models: A survey. arXiv preprint arXiv:2410.12388.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172.

Lottridge, S., Ormerod, C., & Jafari, A. (2023). Psychometric considerations when using deep learning for automated scoring. Advancing natural language processing in educational assessment, 15.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). PIRLS 2021 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

OpenAI. (n.d.). API pricing. Retrieved May 16, 2025 from https://openai.com/api/pricing/

Poddar, S., Koley, P., Misra, J., Podder, S., Ganguly, N., & Ghosh, S. (2025). Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models. arXiv preprint arXiv:2502.05610.

Rahman, C. M., Sobhani, M. E., Rodela, A. T., & Shatabda, S. (2024, September). An Enhanced Text Compression Approach Using Transformer-based Language Models. In 2024 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6). IEEE.

Su, Y., Wang, X., Qin, Y., Chan, C. M., Lin, Y., Wang, H., ... & Zhou, J. (2021). On transferability of prompt tuning for natural language processing. arXiv preprint arXiv:2111.06719.

von Davier, M. & Kennedy, A., Editors (2024). PIRLS 2026 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center https://doi.org/10.6017/lse.tpisc.tr2103.kb4199

Wang, C., Yang, Y., Li, R., Sun, D., Cai, R., Zhang, Y., & Fu, C. (2024, May). Adapting llms for efficient context processing through soft prompt compression. In Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning (pp. 91-97).

Ward, W. C., & Bennett, R. E. (2012). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Routledge.

Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. Advances in Neural Information Processing Systems, 36, 51008-51025.

Xu, Y., & Lapata, M. (2022). Document summarization with latent queries. Transactions of the Association for Computational Linguistics, 10, 623-638.

Yuan, X., Shen, C., Yan, S., Zhang, X., Xie, L., Wang, W., ... & Ye, J. (2024). Instance-adaptive zero-shot chain-of-thought prompting. arXiv preprint arXiv:2409.20441.

Zhang, W., Huang, J. H., Vakulenko, S., Xu, Y., Rajapakse, T., & Kanoulas, E. (2025). Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In International Conference on Pattern Recognition (pp. 89-104). Springer, Cham.

Zhang, Y., Liu, Y., Yang, Z., Fang, Y., Chen, Y., Radev, D., ... & Zhang, R. (2023). Macsum: Controllable summarization with mixed attributes. Transactions of the Association for Computational Linguistics, 11, 787-803.

A Appendices

A. Sample Size by Country

Country	Item 1	Item 2	Item 3	Item 4	Item 5
A	410	524	406	342	449
B	76	82	77	73	68
C	226	252	219	230	212
D	111	119	107	104	100
E	69	70	67	n/a	56
F	72	74	69	64	58
G	126	138	121	127	120
H	102	112	100	142	82
I	60	58	61	60	47
J	80	89	79	79	75
K	85	90	84	69	77
L	107	118	107	99	100
M	67	67	67	63	52
N	46	46	45	43	45
O	80	88	79	79	72
P	83	90	82	76	77
Q	93	92	90	90	79
R	78	87	79	76	70
S	86	91	86	119	77
T	77	82	77	109	70
U	75	80	74	39	64
V	100	104	99	137	80
W	70	73	69	63	57
X	75	79	73	76	63
Y	76	81	76	76	66
Z	82	89	79	79	76
AA	75	76	71	75	60
Total	2687	2951	2643	2589	2452

Table. Sample size by country

B. PIRLS Scoring Template

Evaluate multilingual responses from an international reading assessment for fourth-grade students.

Steps

1. Translation: Translate the student's response into English.
2. Scoring: Score the response according to the given scoring guide.
3. Validation: Determine if the translation could be "hallucinated" where the text appears linguistically correct but fails to capture the intended meaning.
   - If the translation is inaccurate, re-translate and re-score the response.
   - If the original text is untranslatable and nonsensical, keep the original text and assign a score of 0.
4. Output Construction: Compile the result into a JSON object, with either the translated text or the original text (if untranslatable) and the assigned score.

Output Format

The output should be formatted in JSON as follows:
{"[English translation or original text]": "Score: [score]"}

Passage: **[[Original reading passage or question-specific summary]]**

Question: **[[Item's question]]**

Scoring Guide:
Evaluate responses based on the following criteria.

- [Score: 1]: Assign this score if **[[description]]**
- [Score: 0]: Assign this score if the response does not explicitly include the key content described in the [Score: 1] criteria.

Notes

- Minor irrelevant details are permissible only if the response explicitly includes the key content required for [Score: 1] and the details do not contradict the [Score: 1] criteria.
- Character names may vary depending on the language used; such variations should not affect scoring.

C. Confusion Matrices from Optimized AS

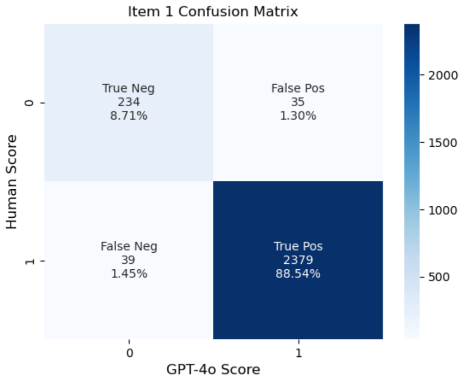


Figure 1. Item 1 confusion matrix

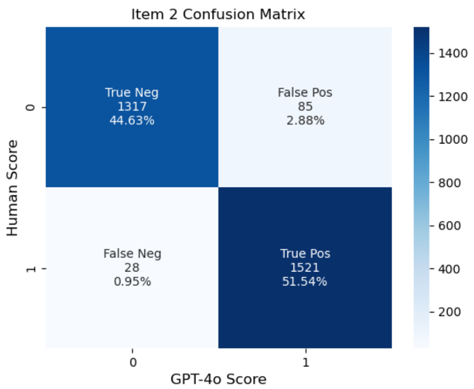


Figure 2. Item 2 confusion matrix

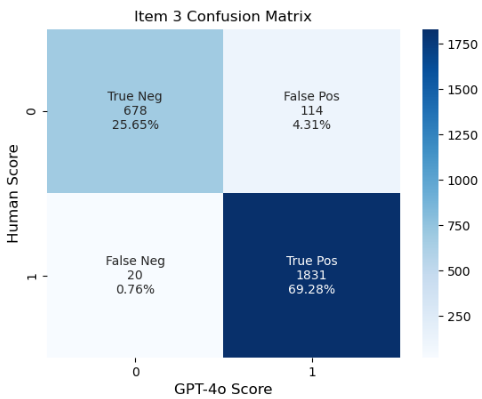


Figure 3. Item 3 confusion matrix

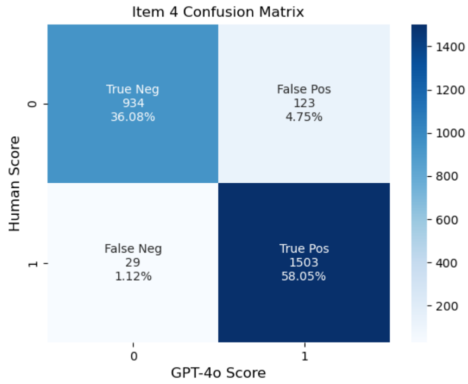


Figure 4. Item 4 confusion matrix

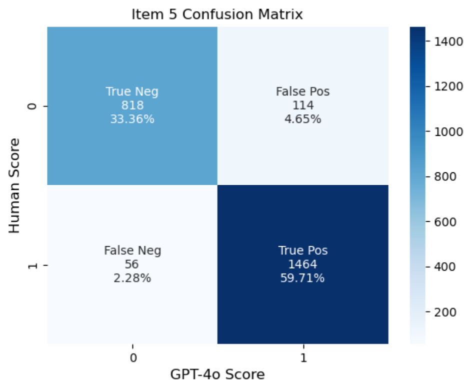


Figure 5. Item 5 confusion matrix

Implementation Considerations for Automated AI Grading of Student Work

Zewei (Victor) Tian[1], Alex Liu[1], Lief Esbenshade[1], Shawon Sarkar[1]
Zachary Zhang[2], Kevin He[2], Min Sun[1]

[1]University of Washington, [2]Hensun Innovation

Abstract

This study explores the classroom implementation of an AI-powered grading platform in K–12 settings through a co-design pilot with 19 teachers. We combine platform usage logs, surveys, and qualitative interviews to examine how teachers use AI-generated rubrics and grading feedback. Findings reveal that while teachers valued the AI's rapid narrative feedback for formative purposes, they distrusted automated scoring and emphasized the need for human oversight. Students welcomed fast, revision-oriented feedback but remained skeptical of AI-only grading. We discuss implications for the design of trustworthy, teacher-centered AI assessment tools that enhance feedback while preserving pedagogical agency.

1 Introduction

The integration of artificial intelligence (AI) into K-12 education has shown promise but also comes with new challenges (Wang et al., 2024). AI-powered educational platforms can offer tools to create instructional materials as well as to provide grading and feedback for assessments. Such tools purport to streamline workflows and provide rapid, individualized feedback. However, concerns arise regarding the alignment with pedagogical goals, the preservation of teacher agency, and mixed impacts on learners. This study engaged 19 teachers in a co-design pilot study for Colleague AI, an online AI-powered education platform for teachers and students that provides AI-based classroom functionality. In this study we focus on the AI grading and feedback functionality and provide generalizable information about how teachers envision the successful implementation of such a tool. By combining quantitative usage data with thematic analyses of teacher interviews and surveys, we examine the conditions under which AI-powered grading practices can augment instructional expertise of the educators. We situate our findings within the broader context of standard-based grading (SBG) and formative feedback theory, elaborating the opportunities AI tools can offer with actionable insights for developers, educators, school leaders and other stakeholders who are committed to empowering education through the assistance of AI without compromising instructional integrity.

1.1 Historical Development of Automated Grading Systems

Automated grading systems have a rich history spanning over nearly a century. In the 1940s, IBM introduced tabulating and test-scoring machines to accelerate scoring, reporting, and computing of assessments (Lorge, 1942). This system was considered to be helpful in saving teachers time and processing student data more efficiently (Benham, 1962), and marks a significant early step toward automated grading. In the 1990s and entering into the 21st century, the introduction of learning management systems (LMS) brought automated grading into the spotlight, together with other functionalities around managing and distributing assessments. With education practices shifting to the digital realm, automated grading systems were driven to improve and adapt. Advanced technological innovations like natural language processing (NLP) and computer vision also assisted in the development of automated grading (Jocovic et al., 2024; Ramesh and Sanampudi, 2022). However, as K-12 education adopts standard-based grading (SBG), assessments, especially formative assessments, require more complicated and comprehensive grading practices. Adding on to that, automated grading primarily focused on handling multiple choice questions while scoring open-ended questions like essays still remains a challenge (Ramesh and Sanampudi, 2022). In this context, the emergence of Large Language Model (LLM) Artificial Intelligence (AI) systems shows a potential next step in the integration of pedagogical frameworks and automated

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 9–20

grading systems, and enables the automated grading process to be adopted on various assessment forms (Chu et al., 2025; Li et al., 2025; Liew and Tan, 2025).

1.2 The Value of Formative Assessment and Feedback in K-12 Education

Effective assessment in K-12 education measures student learning while catalyzing continued growth. However, translating these principles into classroom practice faces practical challenges. This section examines theoretical foundations and empirical evidence supporting formative assessment in K-12 settings, while acknowledging systemic barriers that prevent educators from implementing these practices at scale.

1.2.1 Rubric and standard-based grading in K-12

K–12 education has increasingly adopted standards-based grading (SBG) systems that align with state learning standards, shifting the focus from accumulating points to demonstrating mastery of specific competencies (Guskey and Bailey, 2001; Muñoz and Guskey, 2015). By reporting student performance in terms of proficiency levels—such as "emerging," "developing," "proficient," and "advanced"—SBG provides educators and families with a clearer picture of where learners stand relative to defined objectives (O'Connor, 2007). SBG rubrics feature criteria appropriate to an assessment's purpose and describe these criteria across a continuum of performance levels, ensuring that each standard is assessed with both clarity and precision (Brookhart, 2018). When rubrics are crafted in alignment with state or district standards, they serve as the bridge between curricular goals and day-to-day classroom tasks (McTighe and Wiggins, 2013). In K–12 settings, rubrics serve multiple purposes. First, they clarify expectations for students by defining what knowledge and skills constitute "proficient" and "exemplary" work; knowing these distinctions helps students set concrete targets and engage in self-assessment (Andrade, 2005; Chowdhury, 2018). Second, rubrics provide consistent grading criteria for teachers, reducing subjectivity and inter-rater variability. Rubric-based scoring enhances reliability across different instructors and class sections (Jonsson and Svingby, 2007). Finally, rubrics facilitate communication with parents about student progress: when teachers share rubric scores or performance descriptors, families gain concrete insight into their child's strengths and areas for growth, enabling more focused conversations about how to support learning at home (Chowdhury, 2018; Popham, 2011).

1.2.2 Timing and effectiveness for young learners

Research shows that feedback timing critically affects K–12 learning (Ruiz-Primo and Li, 2013). A meta-analysis reports: "feedback is one of the most powerful influences on learning and achievement" (Hattie and Timperley, 2007). Immediate feedback prevents misconceptions from becoming reinforced, which is particularly important for young learners building foundational skills. Students who receive immediate feedback during tasks retain information better and correct errors faster than those given delayed feedback (Ajogbeje, 2023). These effects are evident across subjects like math and science, where rapid corrective guidance maintains motivation and supports mastery (Dihoff et al., 2004; Mandouit and Hattie, 2023). However, in many K–12 classrooms, practical constraints make providing immediate feedback difficult to sustain. Providing formative feedback to an entire class requires teachers to collect, analyze, and respond to each student's work—a process that research shows is hard to implement at scale and sustain over time (Hopfenbeck et al., 2023). Moreover, a 2024 RAND survey of K–12 educators found that inconsistent access to formative-assessment tools—such as LMS-integrated grading, handheld response devices, or classroom response systems—forces many teachers to rely on paper-based workflows and delay feedback until weekly or biweekly grading cycles (Doan et al., 2024). A 2025 survey of 254 K-12 teachers found that although most value immediacy, workload and inconsistent access to digital tools prevent real-time feedback delivery. Without embedded systems (e.g., response-clickers or automated grading), teachers default to batch feedback, reducing impact (Jin et al., 2025). As a result, feedback often arrives days after submission, by which point students have moved on to new material, weakening the corrective value and allowing misconceptions to persist until the next evaluation cycle.

1.2.3 Separating Formative Feedback from Evaluative Grades in K-12

Separating formative feedback from evaluative grades is essential in K–12 education to prioritize learning and development over ranking. In a classic experimental study, the result showed that sixth-grade students who received detailed comments without grades demonstrated higher intrinsic motivation and better task performance than peers who received grades or grades paired with comments (Butler and Nisan, 1986). Another study found that when grades accompany comments, students tend to focus on the grade itself and disregard substantive feedback (Black and Wiliam, 1998). When feedback is decoupled from grades, teachers can devote attention to describing specific strengths, identifying misconceptions, and suggesting corrective steps without students fixating on scores (Brookhart and Oakley, 2022; Wiliam, 2011).

1.2.4 Recent Research on Automated Graders and Real-Time Feedback in K-12

Recent AI advancements have begun to extend assessment capabilities in K–12 contexts, but implementation in K-12 schools has typically lagged behind higher education. For example, M-Powering Teachers is an automated feedback tool that utilizes natural language processing to analyze verbal classroom interactions and subsequently provides formative feedback to teachers. In a randomized controlled trial with over 1,100 instructors in an online computer science course, the tool increased instructors' use of "uptake" practices (i.e., acknowledging and building on student ideas) by 13 percent (Demszky et al., 2024). This result suggests promise for providing feedback to K-12 teachers to improve their classroom practices with AI-assisted analysis of their teaching This also applies to other activities like administering assessments in the classroom. AI-assisted grading systems are being developed to analyze assessments and provide standard-based rubric (Tian et al., 2025), which then will be used to generate grades and feedback aligned with the standards. These tools recognize the unique needs of K-12 education, including age-appropriate feedback and alignment with Common Core and state standards. However, limitations persist in K-12 contexts. Systematic scoping reviews note that AI tools often assume mature organizational structures and language conventions, which younger learners have not yet mastered (Lindsay et al., 2023; Yan et al., 2024). Moreover, K–12 educators express concerns that AI-mediated feedback may not sufficiently address younger students' socio-emotional needs or align with grade-level curricula—barriers that slow adoption in elementary schools (Castro et al., 2025; Lin and Van Brummelen, 2021).

2 Sample & Methods

For this study, we ran a seven week co-design pilot study with twenty-one teachers from four public school districts and one independent school in the Puget Sound region of Washington state to test the use of an AI powered learning platform's student facing classroom features. Nineteen teachers participated in implementing and testing the assessment feature with their classrooms. Teachers participated in weekly discussion sessions where they received guidance about the platform, discussed how they might use the platform in their classrooms, and provided feedback about how they used the platform. Teachers completed weekly surveys about their platform usage. Two weeks of the pilot study focused on assessment grading. For this study we focus on the usage of the assessment grading functionality. In the pilot study, we interacted directly with teachers as they tested the platform in their classrooms. Students were not the subject of the study, and researchers did not directly interact with students. The study was approved by the University of Washington Institutional Review Board.

During this phase of the study teachers were asked to implement two assessments in their classroom using the platform. Implementation of an assessment comprised several steps. First, teachers defined the purpose, type, and content of the assessment that they would give to their students. Teachers were instructed to only give assessments that fit with their classroom goals and that fit with their regular teaching practice. Then, teachers were given the option to use the AI platform to design a rubric to accompany their assessment and assist in providing feedback to their students. Once students completed the assessment, teachers had the option to allow students to view AI generated feedback and resubmit their assignment (i.e. a formative use) or to allow students only a single submission (i.e. a summative use). Finally, teachers reviewed the AI generated feedback - teachers were able to see the AI generated feedback whether or not they chose to allow students to view it - and returned their own feedback and grades to their students.

Teachers submitted surveys on how the implementation went. Of the 19 teachers who participated in the assessment tool portion of the study, 13 submitted feedback survey forms detailing how they used the platform to implement assessments.

The implementations covered a range of class subject areas including programming/science courses (30%), Math classes (25%), Spanish language classes (15%) and ELA classes (30%). Classes were divided between grades 8 through 12. See Figure 1 for the full breakdown. Some teachers reported trying the tool in multiple class sections, because individual teachers are the focal unit of the study, we have weighted the responses such that each teacher counts equally (e.g. if teacher A reported a single math class and teacher B reported 2 English classes, we would report that study comprised half math and half English classes).

Figure 2 summarizes the type and purpose of the assessments given. Over half (56%) of teachers who repsonded to the survey used the AI platform to administer an in-class formative assessment, and almost half (49%) had 'short-answer' type questions in the assessment. Although only 13 teachers completed the survey, 19 teachers did implement at least one assessment in their classrooms. In total, assessments were created in 33 unique classrooms with 936 student works submitted.

In addition to requesting structured feedback in surveys on the implementation of the AI Grading tool, we applied thematic analysis to qualitative data sources including open-ended survey responses, group discussions, and individual interviews. We employed ground theory to thematic coding (Braun and Clarke, 2006) and identified recurring experiences, affordances, barriers, and recommendations from teachers' perspectives. The established codebook (Appendix A) contains 7 parent code and 18 child code illustrating teachers' and students' user experiences from pilot teachers' perspectives.

We also examined platform log data to understand the scope of the classroom implementation of the AI Grading tool, recording the number of assignments created, the number of student submissions made, whether students resubmitted their assignment and whether the teacher used the platform to return feedback to students.

3 Results

3.1 Platform Log Data Analysis

The platform log dataset includes assessment logs from 33 unique classrooms created by 19 teachers. On average, each classroom implemented approximately 1.76 assessments. From the platform-generated assessment logs from 58 assessments, we conducted usage analysis to capture how AI grading and feedback features were implemented across subjects and school sites. The logs included information on total student enrollment in the classroom, number of submissions, AI-graded assessments, and resubmission counts.

3.1.1 Submission Patterns and Engagement

Submission rates varied widely, with a mean submission rate of 54.8% (SD = 27.9%). While some classrooms achieved full participation, others showed near 0 submission rates, indicating variability in how assessment activities were adopted across contexts. This variation reflects both instructional choice and logistical constraints (e.g., class type, student access, timing).

3.1.2 AI Grading Coverage and Automation

AI systems graded the majority of submitted assessments. In over 75% of classrooms, more than 80% of submitted student work received AI-generated scores. The median AI grading coverage was 92.2%, with many classrooms achieving near-total automation. Both teachers and students can initiate AI grading to generate feedback and evaluation. This high rate of automated grading illustrates the system's capacity to streamline evaluation workflows at scale.

3.1.3 Student Resubmission Behavior

Resubmissions, which may indicate iterative learning or clarification efforts, were relatively infrequent but nontrivial. On average, 8.7% of students submitted work more than once, with a maximum observed rate of 66.7% in one classroom. While not ubiquitous, this behavior suggests some teachers and students leveraged the platform's capacity for revision and feedback loops.

Metric	**Value**
Unique Classrooms	33
Average Assessments per Classroom	1.76
Mean Submission Rate	54.8%
Median AI Grading Coverage on Submitted Works	92.2%
Average Resubmission Rate	8.7%

Table 1: Summary of Assessment Metrics.
Note: Metrics are based on platform logs from 58 classroom-level assessment records across middle and high school implementations.

AI-powered grading was widely implemented across classrooms, with most student work receiv-

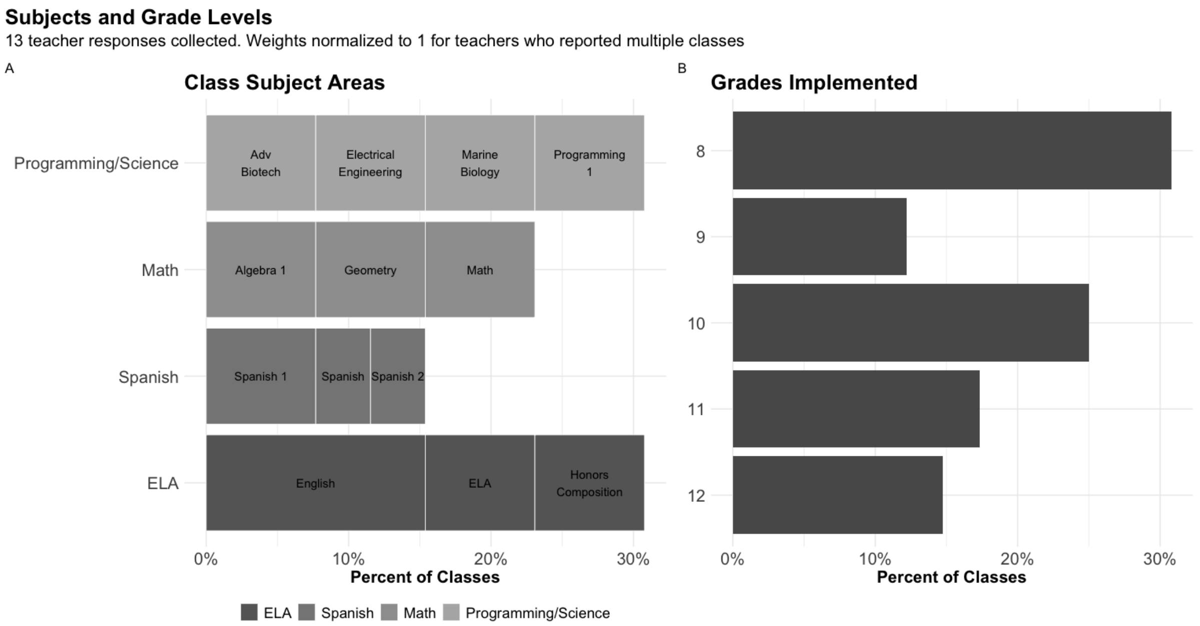


Figure 1

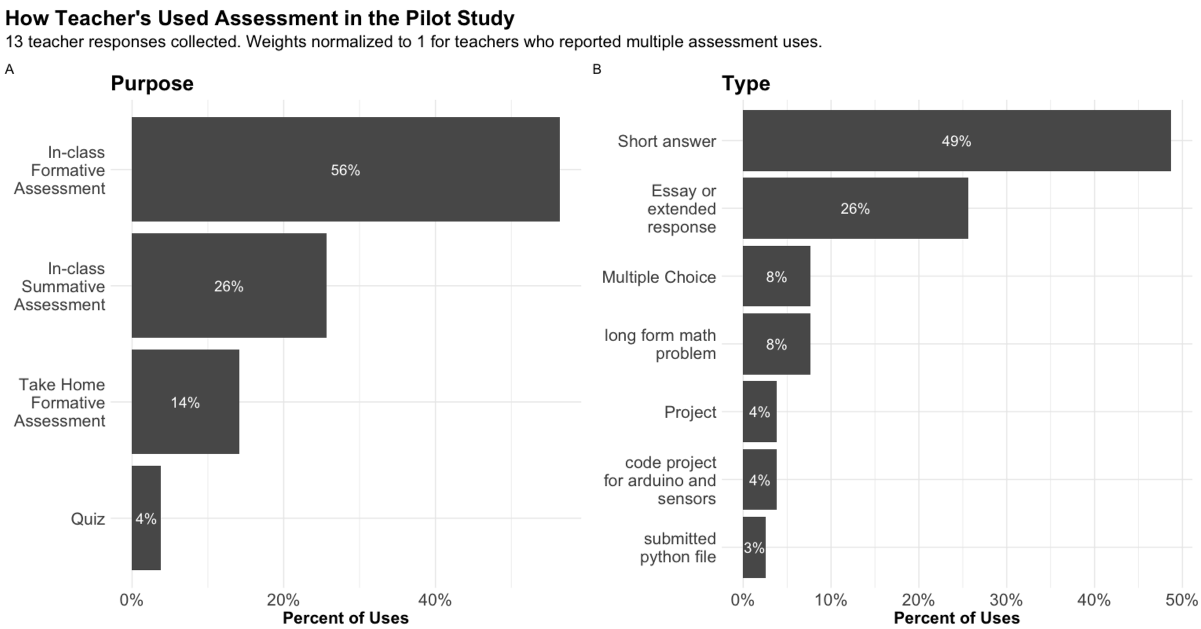


Figure 2

ing automated scores. Yet the variability in student engagement, along with uneven resubmission activity, reinforces a central finding from our qualitative analysis: teacher mediation remains essential to interpreting and contextualizing AI output. Teachers did not simply deploy automation, instead they integrated it into their classroom practices to balance speed with pedagogical intent.

3.2 Teacher Survey Data

13 out of 19 teachers returned survey forms about their use of the AI Grading platform. In addition to the data on the implementation context, they also evaluated the quality of the AI generated rubrics and the AI generated feedback.

3.2.1 Rubric Quality

Over 60% of the teachers indicated that they were able to use the AI generated rubrics in their classroom assignments. The majority indicated that they made minor changes to the rubric, indicating that they were not willing to fully accept the AI generated content without review and adjustment. Interestingly, no teachers indicated that they made major changes to the AI generated rubric. Only 7% of teachers indicated that the rubrics could not be used in their classroom - either because they needed major revisions or were simply not applicable. Roughly a quarter of teachers reported that they did not attempt to use the AI generated rubrics at all. Note that some teachers submitted multiple response forms for their different classrooms, the overall results are weighted so that each teacher has equal weight.

3.2.2 AI Generated Feedback Quality

57% of teachers indicated that the AI feedback provided clear, actionable feedback for teachers or students, with 41% indicating that the feedback was useful for both teachers and students, 14% indicating that the feedback was only useful for students, and 3% indicating that it was only useful for teachers. 42% of teachers indicated that the feedback was not useful, with 24% indicating that the feedback was vague or unhelpful and 18% indicating that it was incorrect or misleading.

3.3 Discussion Transcript & Interview Qualitative Analysis

To deepen our understanding of how teachers experienced the AI-powered assessment, grading, and feedback features in real K-12 classroom contexts, we conducted a qualitative analysis as well. The qualitative analysis yielded three central themes, reflecting both the promise of AI to enhance feedback workflows and the structural and pedagogical tensions that emerge in educational contexts.

3.3.1 AI Grading: Feedback as Formative Scaffold over Numerical Scores

Across classroom contexts, teachers consistently emphasized the pedagogical value of narrative feedback over numerical grades. While the platform offered a mechanism for scoring, many teachers found the AI's application of point values to be inconsistent or misaligned with their rubrics. One teacher shared, "The tool scored some students out of 20 points and others out of 10, when I had specified the assessment was worth 10 points" (Marine Biology, Grades 10–12). Others noted the AI "took points off for things not in the rubric" or used standards "outside of the students' current skill level" (Engineering, Grades 9–12). By contrast, the system's narrative feedback was frequently praised for its specificity, clarity, and alignment with formative goals. Teachers described it as a useful "first draft" that helped identify student misconceptions and suggest improvement strategies. "While I found the feedback from the AI to be fairly accurate, it seemed inconsistent in terms of how it attached numbers to that feedback," stated by the same grades 10–12 marine biology teacher. This tension between qualitative and quantitative outputs suggests that current LLM-based assessment systems may be best positioned as formative tools, generating scalable, revisable feedback that scaffolds learning, rather than reliable summative graders. Teachers expressed interest in treating AI grading as a fast first-pass diagnostic, followed by human adjustment. "[Students] loved the prospect of getting a grade and feedback with such a quick turnaround, rather than waiting the 2–3 weeks that it usually takes me to grade their writing" (English, Grade 11).

This orientation toward feedback-first design reinforces the importance of transparency and explainability in AI-powered assessment tools. When numerical scores lack clarity or consistency, but written comments hold pedagogical value, the role of the AI should be reimagined: not to replace teacher judgment, but to scaffold learning through accessible, timely, and editable feedback.

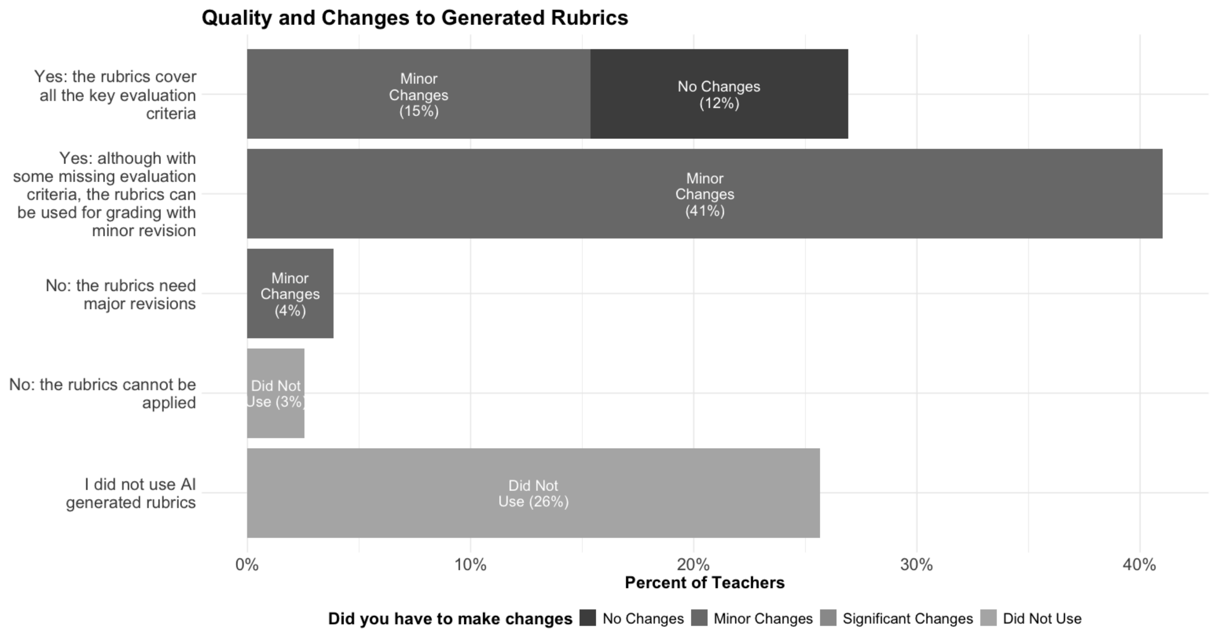


Figure 3: Teacher survey responses to the quality of the AI generated rubric and whether it was necessary to make changes.

3.3.2 Teacher Oversight Enables Trust and Personalization

Teachers reported that AI feedback, while efficient, was not passively accepted by students. Students actively scrutinized AI generated evaluations' fairness, clarity, and alignment with their work. This dynamic created new expectations for teachers to engage in the grading and feedback process, not just as overseers, but as collaborators who could validate, revise, or clarify the AI's output. One educator noted, "My students were confused why some feedback was so positive, yet the score was low. They came to me asking if the grade was accurate and what it really meant" (English, Grade 11). Far from seeing this as a burden, many teachers described this supervising and collaborative role as essential and empowering. It allowed them to reinforce instructional goals, personalize communication with students, and restore fairness to the grading process. "I took the feedback and put it in the AI Chat… told it, 'give me one paragraph in teacher voice,'" one teacher explained, reflecting the effort to mediate AI output in a way that aligned with their teaching persona and classroom discourse (ELA, Grade 10). Another explained, "I asked if [students] would be ok if the AI graded all their work and they all said no! They want to know I'm reading their work. They want me to see their jokes and emotions. They feared that AI would just be like a checklist. I thought her [the AI assistants] feedback was better than mine though. (But they thought it'd be great as a pre-submission self check grade)" (Science, Grade 11). These moments highlight that personalization is not merely the product of generative automation, it is co-produced through educator framing and student trust. Even among those critical of the AI's limitations, teachers valued the system's ability to streamline initial feedback, reduce turnaround time, and make space for higher-order instructional moves. "I see AI grading tools as a kind of new TA: it gives fast, helpful first-pass feedback that enables students to make improvements right away, but I still review and make final grading decisions." (Engineering, Grade 9-12). Even teachers who were critical of the AI's limitations noted its utility for surfacing initial insights that they could refine or expand. In this human-AI collaborative process, automation enhances efficiency, but teacher oversight ensures that outputs align with pedagogical goals and relational norms. Teachers stressed that speed alone was insufficient: the AI's utility depended on whether its feedback meaningfully reflected classroom expectations. "The time-saving is great. But only if the comments represent how I would actually respond to student work. Otherwise I have to re-do it anyway." (Math, Grades 7–8). This convergence of student demand and teacher professional judgment highlights a collaborative model of assessment: one where AI tools extend

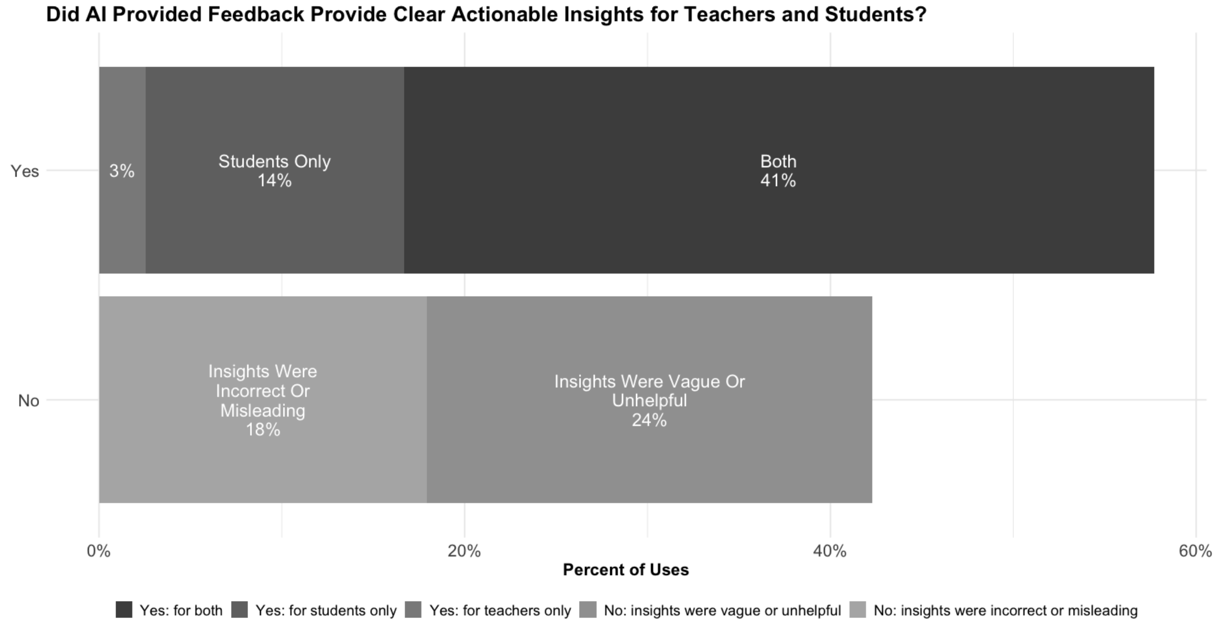


Figure 4: Teacher survey responses to the quality of the AI generated assessment feedback

instructional reach, but teachers retain interpretive authority. Personalization, in this view, is not the result of automation alone, it is made meaningful when filtered through pedagogical expertise and enacted in response to learner needs.

3.3.3 Student Engagement is Mediated by Interface Design and Accessibility Considerations

Teacher noted that student responses to AI evaluations varied significantly, shaped not only by content quality but also by interface design and prior technology exposure. On one hand, Several teachers reported strong engagement among struggling or anxious learners, who appreciated the opportunity to receive feedback before submitting to peers. For example, a grade 11 IB English teacher shares, "some of my struggling students… liked having someone to give feedback before sharing. It made them more confident." On the other hand, some students were overwhelmed by the volume or complexity of the comments. One teacher noted, "They thought it was a lot of feedback... It might have been better to let me limit it to just a few things" (World Language, Grades 9–11). Technical usability also posed barriers: some students had trouble uploading assignments, locating relevant sections of the AI-generated feedback, or were put off by first impressions of the interface, which "looked old," all of which may have led them to abandon the tool after initial attempts. The mixed reception reinforces the importance of usability design and accessibility. Without scaffolds for clarity and navigation, AI systems may inadvertently heighten disparities in experience and learning outcomes among students with different levels of digital fluency. To avoid these pitfalls, developers must prioritize transparency, explanation, and accessibility in system design. Features such as adjustable feedback volume, simpler and fashionable interfaces, and teacher-led onboarding may be critical to ensuring that AI systems support meaningful engagement across all learners.

4 Discussion & Conclusions

This study offers early empirical insight into how K–12 educators engage with AI-powered grading systems in real classroom contexts. Through a co-design pilot with 20 teachers, we observed that while automated scoring tools are increasingly capable of streamlining feedback and assessment workflows, their successful classroom implementation hinges on how well they align with formative goals, support teacher expertise, and align with student expectations for fairness. Teachers used the platform to generate rubrics, assign assessments, deliver formative feedback, and manage revision cycles, but they did not treat AI output as final. Instead, they exercised discretion, editing feedback, clarifying grades, and recontextualizing comments to maintain pedagogical coherence. This model where automation accelerates routine processes but teachers retain interpretive control emerged as a key condition for productive use.

Throughout the study, three themes emerged: (1) teachers emphasized narrative feedback over numeric scores, valuing elaborated comments generated by AI that revealed misconceptions and next steps for learning; (2) teacher mediation was essential to address discrepancies between comments and grades, underscoring that AI should augment rather than replace educator judgment; and (3) student responses varied—some benefited from low-stakes feedback with a quick turnaround, while others experienced cognitive overload or usability challenges, revealing heterogeneity considerations tied to digital literacy and AI competency in modern classrooms.

While this pilot study provides valuable insight into teacher experiences with AI-powered grading tools, several limitations warrant consideration. First, the sample was geographically limited to the Puget Sound region and comprised volunteers who may be more open to AI technology use than the broader teaching population, potentially introducing selection bias. Second, not all participating teachers completed post-implementation surveys, which may skew those findings toward those with stronger opinions or more successful experiences. Third, the study relied on teacher self-reported data and platform logs rather than direct observation of classroom implementation, limiting our ability to assess actual student interaction with the AI system. Finally, this study is of a single generative AI based platform, and findings may not fully generalize to other AI grading and feedback systems.

Despite its exploratory scope, this study yields several insights that are likely to generalize beyond the immediate implementation context. Most notably, teachers consistently valued AI-generated narrative feedback as a formative tool, even when they questioned the reliability of automated scoring. This suggests that LLM-based grading systems may be best positioned not as replacements for teacher judgment, but as scaffolds for feedback-rich instruction. Additionally, the finding that students desired teacher involvement—even when AI feedback was accurate—underscores the importance of maintaining human connection and interpretive authority in automated systems. Finally, the study highlights design considerations for future AI tools: systems should allow for teacher oversight, offer clear interfaces for student understanding, and support workflows that enable iterative revision. These features are likely to be essential across a wide range of school settings and instructional models.

Acknowledgments

This work is supported by the Institute of Education Sciences of the U.S. Department of Education, through Grant R305C240012 and by several awards from the National Science Foundation (NSF #2043613, 2300291, 2405110) to the University of Washington, and a NSF SBIR/STTR award to Hensun Innovation LLC (#2423365). The opinions expressed are those of the authors and do not represent views of the funders.

References

Oke James Ajogbeje. 2023. Enhancing Classroom Learning Outcomes: The Power of Immediate Feedback Strategy. *International Journal of Disabilities Sports and Health Sciences*, 6(3):453–465. Number: 3 Publisher: NDP Academic Publishing.

Heidi Goodrich Andrade. 2005. Teaching with rubrics: The good, the bad, and the ugly. *College teaching*, 53(1):27–31. ISBN: 8756-7555 Publisher: Taylor & Francis.

Frank Benham. 1962. A Method for Processing Test Scores with Minimal Punched Card Equipment. *Florida Journal of Educational Research*, 4(1):65–70. ISBN: 2575-6109.

Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74. ISBN: 0969-594X Publisher: Taylor & Francis.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101. ISBN: 1478-0887 Publisher: Taylor & Francis.

Susan M. Brookhart. 2018. Appropriate Criteria: Key to Effective Rubrics. *Frontiers in Education*, 3. Publisher: Frontiers.

Susan M. Brookhart and Alice Oakley. 2022. Gathering Feedback from Student Work.

Ruth Butler and Mordecai Nisan. 1986. Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of educational psychology*, 78(3):210. ISBN: 1939-2176 Publisher: American Psychological Association.

Vania Castro, Ana Karina de Oliveira Nascimento, Raigul Zheldibayeva, Duane Searsmith, Akash Saini, Bill Cope, and Mary Kalantzis. 2025. Implementation of a Generative AI Assistant in K-12 Education: The CGScholar AI Helper Initiative. *arXiv e-prints*, page arXiv: 2502.19422.

Faieza Chowdhury. 2018. Application of Rubrics in the Classroom: A Vital Tool for Improvement in Assessment, Feedback and Learning. *International Education Studies*, 12(1):61.

Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. A LLM-Powered Automatic Grading Framework with Human-Level Guidelines Optimization. *arXiv preprint*. ArXiv:2410.02165 [cs].

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2024. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3):483–505. ISBN: 0162-3737 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Roberta E. Dihoff, Gary M. Brosvic, Michael L. Epstein, and Michael J. Cook. 2004. Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *The Psychological Record*, 54:207–231. ISBN: 0033-2933 Publisher: Springer.

Sy Doan, Joshua Eagan, David Grant, and Julia H Kaufman. 2024. American instructional resources surveys: 2024 technical documentation and survey results. american educator panels. research report. rr-a134-24. *RAND Corporation*.

Thomas R. Guskey and Jane M. Bailey. 2001. *Developing Grading and Reporting Systems for Student Learning*. Corwin Press. Google-Books-ID: O37oL0PL8wUC.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112. ISBN: 0034-6543 Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Therese N. Hopfenbeck, Zhonghua Zhang, Sundance Zhihong Sun, Pam Robertson, and Joshua A. McGrane. 2023. Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. *Frontiers in Education*, 8. Publisher: Frontiers.

Flora Ji-Yoon Jin, Wei Dai, Bhagya Maheshi, Roberto Martinez-Maldonado, Dragan Gašević, and Yi-Shan Tsai. 2025. Feedback in K-12 and higher education: Educators' perspectives. *Teaching and Teacher Education*, 156:104933. ISBN: 0742-051X Publisher: Elsevier.

Vladimir Jocovic, Milan Marinkovic, Sasa Stojanovic, and Bosko Nikolic. 2024. Automated assessment of pen and paper tests using computer vision. *Multimedia Tools and Applications*, 83(1):2031–2052.

Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144.

Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. LLM-based Automated Grading with Human-in-the-Loop. *arXiv preprint*. ArXiv:2504.05239 [cs].

Pei Yee Liew and Ian K. T. Tan. 2025. On Automated Essay Grading using Large Language Models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, CSAI '24, pages 204–211, New York, NY, USA. Association for Computing Machinery.

Phoebe Lin and Jessica Van Brummelen. 2021. Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–12.

Euan D. Lindsay, Mike Zhang, Aditya Johri, and Johannes Bjerva. 2023. The Responsible Development of Automated Student Feedback with Generative AI. *arXiv preprint arXiv:2308.15334*.

Irving Lorge. 1942. Tabulating and Test-Scoring Machines: Applications of International Business Machines to Educational Research. *Review of Educational Research*, 12(5):550–557. Publisher: [Sage Publications, Inc., American Educational Research Association].

Luke Mandouit and John Hattie. 2023. Revisiting "The Power of Feedback" from the perspective of the learner. *Learning and Instruction*, 84:101718. ISBN: 0959-4752 Publisher: Elsevier.

Jay McTighe and Grant Wiggins. 2013. *Essential questions: Opening doors to student understanding*. Ascd.

Marco A. Muñoz and Thomas R. Guskey. 2015. Standards-based grading and reporting will improve education. *Phi Delta Kappan*, 96(7):64–68. ISBN: 0031-7217 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Ken O'Connor. 2007. A repair kit for grading. *Portland, OR: Educational Testing Service*.

W. James Popham. 2011. *Transformative assessment in action: An inside look at applying the process*. ASCD.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Maria Araceli Ruiz-Primo and Min Li. 2013. Examining formative feedback in the classroom context: New research perspectives. In *SAGE handbook of research on classroom assessment*, pages 215–232. SAGE Publications, Inc.

Zewei (Victor) Tian, Lief Esbenshade, Alex Liu, Shawon Sarkar, Zachary Zhang, Kevin He, and Min Sun. 2025. Rubric Generation in Colleague AI: Transforming Assessment in Education. *Social Innovations Journal*, 30.

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252:124167. ISBN: 0957-4174 Publisher: Elsevier.

Dylan Wiliam. 2011. *Embedded formative assessment*. Solution tree press.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112. ISBN: 0007-1013 Publisher: Wiley Online Library.

A Codebook of Educator Feedback on AI-Powered Assessment

Table 2: Qualitative Coding Scheme with Frequencies

Code	Parent Code	Child Code	Frequency	Child Code Description
WF1	Workflow and Implementation	Feature Setup Challenges	29	Describes obstacles in setting up assessments, rubrics, or assignments using the AI tools.
WF2	Workflow and Implementation	Alternative Use Cases	8	When teachers adapted or repurposed features for different pedagogical intents.
WF3	Workflow and Implementation	Time-Saving Potential	9	Mentions of AI helping reduce grading load or turnaround time.
FB1	Feedback Quality and Utility	Feedback Customization	5	Teachers modifying AI-generated feedback to suit student needs or tone.
FB2	Feedback Quality and Utility	Feedback Usefulness	12	Teachers' perceptions of whether the feedback is pedagogically meaningful or accurate.
FB3	Feedback Quality and Utility	Student Perception of Feedback	2	How students perceive or react to AI feedback.
ST1	Student Impact	Increased Engagement	5	Positive changes in student engagement or willingness to revise based on AI feedback.
ST2	Student Impact	Student Confusion or Frustration	5	Instances of student difficulty with interface, grading accuracy, or expectations.
ST3	Student Impact	Equity of Support	2	Reflections on how AI tools affected different learner groups.
TR1	Trust and Accuracy	Inconsistency of Grading	2	Reports of AI producing different results for the same submission or not aligning with rubric.
TR2	Trust and Accuracy	Human Oversight	10	Emphasis on teacher's role in verifying or revising AI grading before finalizing.
UI1	Usability and Interface	Clunky Interface or Poor UX	2	Descriptions of confusion or dissatisfaction with platform usability.
UI2	Usability and Interface	Preferred Interaction Pathways	2	Teacher workarounds or preferences for using other tools.
PR1	Professional Use and Reflection	Teacher Accountability and Editing	1	Teachers feeling responsible for editing and verifying AI output.
PR2	Professional Use and Reflection	Planning for Growth	4	Teachers thinking about scaling or adjusting practice using AI.
SD1	Suggestions for Development	Workflow Simplification	2	Recommendations to reduce clicks or streamline setup.
SD2	Suggestions for Development	Granular Feedback Requests	2	Suggestions for item-level feedback or clearer linkage to rubrics.
SD3	Suggestions for Development	Feature Expansion	2	Ideas like nudging systems, PDF exports, or data summaries by student.

Compare Several Supervised Machine Learning Methods in Detecting Aberrant Response Pattern

Yi Lu, Yu Zhang, Lorin Mueller

Federation of State Boards of Physical Therapy

Abstract

An aberrant response pattern, e.g., a test taker is able to answer difficult questions correctly, but is unable to answer easy questions correctly, are first identified lz and lz*. We then compared the performance of five supervised machine learning methods in detecting aberrant response pattern identified by lz or lz*.

1 Introduction

Investigating fraudulent testing behavior, especially for high-stakes assessments, has been a common practice for maintaining test score validity. In practical assessment, one of the important problems is to ensure that the test taker's response pattern is consistent with the expected item score pattern. When the difference between the observed and the expected pattern is large, it is classified as an aberrant response pattern (Magis, Raiche, & Beland, 2012; Meijer & Tendeiro, 2014). One example is test taker is able to answer difficult questions correctly, but is unable to answer easy questions correctly. Lz and its modification Lz*, two well-known person-fit statistics are applied in the study to detect aberrant response pattern specified above.

The rapid advancement of machine learning (ML) techniques has led to their widespread application across various domains. In recent years, several studies have conducted comprehensive comparisons of machine learning models to understand their relative strengths and limitations across diverse tasks (e.g., Caruana and Niculescu-Mizil, 2006; Neagu et al., 2007; Raschka, 2018). Collectively, these studies provide a foundational basis for applying and evaluating machine learning algorithms in the present study, which focuses on detecting aberrant response patterns using indices such as the lz and lz* statistics. In the field of educational science, several studies explored machine learning to detect exam cheating (e.g., Man et al., 2019; Pan et al., 2022; Zopluoglu, 2019). There are relatively few studies implementing machine learning methods to investigate aberrant response pattern as specified in the current study.

2 Data

Data used for this study was selected from a licensure exam that is administered multiple times each year. We selected one test form that was administered twice in one year for this study. We used item responses from 2561 examinees who took this form in April as training data. We used item responses from 492 examinees who took the same form in October as test data. There were 200 scored items in this form. Item response for these 200 items was taken as input features. The target variable for each examinee is either flagged as an aberrant response pattern or not based on lz or lz* person fit statistics. In literature, the cutoff value of -4 is used to flag examinees of aberrant response patterns (Tendeiro, Meijer, & Niessen, 2016). In our operational analysis, we used the criteria listed in Table 1 on page 7 to flag aberrant response pattern. Using flagging criteria in Table 1, "flagged #" column in Table 2 on page 7 lists the number of flagged cases in training and test data based on lz and lz* indices, respectively. For our data, the examinees with aberrant response pattern are the minority. A much smaller number of positive cases (aberrant response pattern examinees) can lead to bias in model prediction. To handle the issue of data imbalance, we then conducted data simulation. That is, based on the response pattern of the flagged cases, we simulated one time and two times of examinees that have very similar responses as the flagged aberrant response pattern. The last two columns in Table 2 present the simulated number of aberrant response pattern. Those simulated cases were then randomly inserted and replaced normal response

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 21–24
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

pattern in the original data. In this way, the total number of examinees in training and test data remain the same.

3 Methods

3.1 Lz and Lz* Person-fit Statistics

Drasgow, Levine, & Williams (1985) proposed a standardized version of lz

$$lz = \frac{l_0 - E(l_0)}{V(l_0)} \quad (1)$$

Where l_0 is the log likelihood function of any response pattern, $E(l_0)$ and $V(l_0)$ are the mean and variance of l_0

Snijder (2001), proposed lz*, in which true ability estimates were replaced by sample ability estimates. Magis et al (2012) illustrated lz* as

$$lz^* = \frac{Wn(\hat{\theta}) - c_n(\hat{\theta}) * r_0(\hat{\theta})}{\tilde{V}[l_0(\hat{\theta})]^{1/2}} \quad (2)$$

where $Wn(\theta)$ is a statistic, $r_0(\theta)$ is an estimator, $c_n(\theta)$ is a function modifying $r_0(\theta)$, $\tilde{V}[l_0(\theta)]$ is the modified variance. Magis et al. (2012) has detailed illustrations of those statistics. From equations 1 and 2, we can say that lz* index is a rescaled version of lz by adjusting both its mean and its variance. Lz and lz* are implemented in the current study to identify aberrant response patterns, as illustrated in the data section.

.

3.2 Supervised Machine Learning Methods

Machine learning is broadly categorized into four main types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. As stated below, five supervised learning methods are implemented in the current study to flag aberrant response pattern identified by lz or lz*.

K-Nearest Neighbor (KNN): KNN is a learning algorithm that attempts to classify new samples by allocating them to the class of the most similar labeled cases. In this study, the KNN algorithm was employed to classify examinee response vectors flagged by the lz or lz* indices as either aberrant or normal. The algorithm does not make assumptions about the underlying data distribution, making it particularly suitable for exploratory and diagnostic contexts. The simplicity and interpretability of KNN provide a valuable benchmark against which more complex models—such as neural networks or Support Vector Machines—can be compared.

Naïve Bayes: The Naïve Bayes classifier is a probabilistic machine learning model based on Bayes' Theorem. Bayes' Theorem is formally expressed as:

$$P(C_k|x) = \frac{P(x|C_k)*P(C_k)}{P(x)} \quad (3)$$

Under the naïve conditional independence assumption, the joint likelihood simplifies to a product of individual feature likelihoods:

$$P(C_k|x_1, x_2, \ldots x_n) \propto P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (4)$$

The classification rule then becomes:

$$\hat{y} = \underset{k\in\{1,k\}}{arg\ max}\, P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (5)$$

Based on equation above, Naïve Baye classification algorithm can be used for categorizing new observation into predefined classes for the initiated data. In this study, the Gaussian Naïve Bayes variant was applied to detect aberrant response pattern identified by the lz or lz* indices. The model was implemented using the GaussianNB class from the sklearn.naive_bayes module in Python.

Logistic regression: Logistic regression models the probability that a given input belongs to a specific class. It does this by applying the sigmoid (logistic) function to a linear combination of the input features (Hosmer, Lemeshow, & Sturdivant, 2013).

The sigmoid function is defined as:

$$S(y) = \frac{1}{1+e^{-y}} \quad (6)$$

In the context of logistic regression, the input to the sigmoid function is a linear combination of the predictor variables:

$$p = \frac{1}{1+e^{-(mx+b)}} \quad (7)$$

Where *p* is the estimated probability that the instance belongs to class 1 (e.g., exhibiting aberrant response pattern), *m* represents the weight coefficients (slopes), *X* is the feature vector (e.g., item responses), and *b* is the intercept (bias).

Logistic regression learns these parameters during model training by maximizing the likelihood of the observed data. In binary classification, a threshold (typically 0.5) is applied to the predicted probability to assign class labels. The model was implementedusing the LogisticRegression class from the sklearn.linear model module in Python.

Support Vector Machine (SVM): The central idea behind SVM is to find the optimal hyperplane that best separates data points from different classes in a high-dimensional space. For binary classification, as in the current study, the goal is to maximize the margin between the two classes—the distance between the hyperplane and the nearest data points from each class, known as support vectors.

In this study, a Support Vector Machine (SVM) classifier was employed to detect examinees with aberrant response patterns, as flagged by the lz or lz* indices. The SVM model was implemented using the SVC class from the scikit-learn library in Python. The default SVM configuration with a radial basis function (RBF) kernel was used, which allows the model to capture non-linear relationships in the data.

Neural networks (NNs): NNs are a class of machine learning models inspired by the structure and function of the human brain. They consist of layers of interconnected processing nodes (neurons), where each neuron applies a transformation to the input and passes the result to subsequent layers. Each connection between neurons is associated with a weight that is learned during training through optimization algorithms such as stochastic gradient descent and backpropagation. To classify examinees based on aberrant response pattern identified by the lz or lz* indices, a feedforward neural network was implemented using TensorFlow and Keras.

In the current study, the architecture of the neural network included the following items:

- An input layer with 200 features (corresponding to the number of items),
- Two hidden layers with ReLU activation functions,
- Dropout layers for regularization to mitigate overfitting, and
- A final output layer with a sigmoid activation function for binary classification.

4. Software for Estimation

In this experimental stage, we used Google Colab for estimation. Oversample method was applied in Colab to make sure all aberrant response patterns have been sampled when training the model.

5. Results

One essential tool to evaluate the performance of machine learning models is confusion matrix. A confusion matrix is a simple table that shows how well a classification model is performed by comparing its predictions to the actual results. A confusion matrix adapted to the context of the current study is presented in Table 3 on page 7. Below is a brief explanation on evaluation metrics that applied in the study to evaluate the performance of these supervised machine learning mothods.

Precision$=\frac{TP}{TP+FP}$

Precision focuses on the accuracy of the model's positive predictions. It tells us how many of the instances predicted as positive are actually positive.

Recall/Sensitivity $= \frac{TP}{TP+FN}$

Recall measures the proportion of correctly predicted positive instances among all actual positive instances.

F1score=2 $* \frac{Precison * Sensitivity}{Precision + Sensitivity}$

F1 score combines precision and recall into a single metric to balance their trade-off. It provides a better sense of a model's overall performance, particularly for imbalanced datasets. F1 score ranges from 0 to 1, with 1 indicating the best possible performance.

Accuracy$=\frac{TP+TN}{TP+TN+FN+FP}$

Accuracy measures how often the model's predictions are correct overall. It gives a general idea of how well the model is performing.

In the current study, under different conditions on the number of aberrant response pattern, the resulting classification performance was compared among five supervised machine learning models. Tables 4 and 5 on pages 8 and 9 summarize the classification performance of five machine learning models in detecting aberrant response patterns as identified by the Lz and lz* index, respectively.

Results in these two tables show that, under the condition of the real number of flagged cases, most models—particularly KNN and SVM—struggled to detect aberrant responses, often yielding near-zero F1-scores. Logistic regression consistently achieved high precision but suffered from low recall, while Naïve Bayes and neural networks offered more balanced but modest performance. These results underscore the effectiveness of simulation-based data augmentation for enhancing model sensitivity and suggest that sample size and class balance are critical factors in building reliable aberrant response detectors.

6. Conclusion

In this study, we implemented five supervised machine learning models in detecting aberrant response pattern identified by lz and lz* indices. Across both the Lz and Lz* indices, machine learning models demonstrated consistently high accuracy in identifying normal response patterns. However, performance in detecting aberrant response patterns varied considerably and was highly sensitive to class imbalance. As the number of aberrant responses increased through simulation (1x and 2x the original cases), all models showed marked improvement in identifying aberrant patterns, with F1-scores for class 1 increasing by 2–3 times or more.

In our research, the primary goal of this study has been to compare and choose the best machine learning models. Based on the evaluation metrics—including precision, recall, and F1 score—logistic regression and neural network models demonstrated the strongest performance in detecting aberrant response patterns (in the condition of a real number of aberrant response pattern). However, it is important to note that training the neural network required substantially longer computation time compared to logistic regression. While both models show promise, their effectiveness should be further validated using independent datasets to ensure generalizability. Future research may also explore the potential of alternative machine learning models to enhance detection accuracy and efficiency in various operational contexts.

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161–168.

Drasgow, F., Levine, M. V., & Williams, M. E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37* (1), 57–81.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of Data Mining Methods to Detect Test Fraud. *Journal of Educational Measurement*, 56(2), 251–279.

Meijer, R.R. & Tendeiro, J. N. (2014). The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us. Law School Admission Council Research Report 14-03, March 2014.

Neagu, D. C., Guo, G., Trundle, P. R., & Cronin, M. T. D. (2007). A Comparative Study of Machine Learning Algorithms Applied to Predictive Toxicology Data Mining. *Journal of Chemical Information and Modeling*, 47(2), 716–729.

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A Machine Learning Approach for Detecting Item Compromise and Preknowledge in Computerized Adaptive Testing. *Psychological Test and Assessment Modeling*, 64(4), 385–424.

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv preprint arXiv:1811.12808.

Snijders, T. A. B. (2001). Asymptotic Null Distribution of Person Fit Statistics with Estimated Person Parameter. *Psychometrika, 66*(3), 331–342.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R Package for Person Fit Analysis in IRT. *Journal of Statistical Software, 74*(5). doi: 10.18637/jss.v074.i05

Zopluoglu, Z. (2019). Detecting Examinees with Item Preknowledge in Large-Scale Testing Using Extreme Gradient Boosting (XGBoost). *Educational and Psychological.*

Leveraging multi-AI agents for a teacher co-design

Hongwen Guo *[1], Matthew S. Johnson†[1], Luis Saldivia‡[1], Michelle Worthington§[1], and Kadrye Ercikan¶[1]

[1]ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541

***Paper prepared for 2025 AIME-con at Pittsburgh**

Abstract

This study uses multi-AI agents to accelerate teacher co-design efforts. It innovatively links student profiles obtained from numerical assessment data to AI agents in natural languages. The AI agents simulate human inquiry, enrich feedback and ground it in teachers' knowledge and practice, showing significant potential for transforming assessment practice and research.

Keywords: Human-centered AI, AI agents, large-scale assessment, response and process data, feedback

1 Background

1.1 Literature review

The existing work in learning analytics and educational data mining has provided a strong foundation for understanding student learning through data. Researchers have been seminal in leveraging fine-grained log data from digital learning environments to offer deep insights into complex learning processes (e.g., Baker and Yacef, 2009; Baker, 2021; Thomas et al., 2025; Darvishi et al., 2024). Work in these learning areas also demonstrates the practical application of learning process data in refining intelligent tutoring systems and prediction of student learning outcomes (e.g., Khan Academy, 2025; Ritter et al., 2013; Zheng et al., 2019).

While learning analytics leverages diverse student interaction data (e.g., from learning management systems) to provide feedback and improve instructional design, assessment analytics applies similar data mining and statistical techniques to interpret student performance on tests, evaluate item quality, ensure assessment validity and fairness, and develop measurement innovation (Ercikan et al., 2023; Ercikan and Pellegrino, 2017). Recently, process data collected from log data in large-scale assessments (LSAs) has been gaining momentum in educational measurement, largely due to data availability from NAEP, PISA, TIMSS, etc. (National Assessment Governing Board, 2020; Organisation for Economic Co-operation and Development, 2020; International Association for the Evaluation of Educational Achievement, 2020). Studies using process data in LSAs and other assessments can be found in areas such as test-taking strategies, score validity on the assessments, its relationship with performance (Ercikan et al., 2020; Guo and Ercikan, 2021; Pools and Monseur, 2021), and problem-solving patterns (Greiff et al., 2016; Zoanetti and Griffin, 2017).

Process/log data, as exhibited in the above studies, contain nuanced information about how students engaged with tasks and assessments. Such large and complex data from LSAs may pose challenges to traditional psychometric analysis but offer opportunities for using AI to discover data insights. Recent studies (e.g., Guo et al., 2024a,b) attempted to use NAEP multi-source data (i.e., response data and process data) and human-centered AI (HAI) frameworks to generate preliminary student profiles, which show promises in contextualizing a performance score and providing meaningful and actionable feedback to classroom teachers. These preliminary student profiles were created based on multi-source data when students interacted with LSAs digital platforms. The HAI approach helped to identify near a dozen preliminary profiles, many associated with low-performing students. For teaching, such profiles are intended to provide educators with rich, meaningful feedback, helping them understand how students engaged with the assessment beyond a performance score, which can shed light on students' learning skills to inform classroom teaching practices. Similar

*hguo@ets.org; Corresponding author
†msjohnson@ets.org
‡lsaldivia@ets.org
§mworthington@ets.org
¶kercikan@ets.org

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 25–34
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

to AI applications in learning systems, the AI applications on LSAs can help to drive significant innovation in LSA practice and research, informing both teaching and learning practices.

However, for such preliminary student profiles generated from LSA research to make a real impact on teaching, they need to be refined and improved and grounded in classroom practice with a teacher co-design.

1.2 Aims

As a stepping stone toward transforming LSA research to teaching practice, the primary goal of the current study is to leverage multiple AI agents and their reasoning capabilities to facilitate an effective teacher co-design for transforming assessment research into teaching professional development.

More specifically, in the current study, we propose to use AI multi-agents to find a common ground before we collaborate with real teachers. AI agents will act as experienced educators to understand the multi-source data, refine the preliminary student profiles, generate highlights of student strengths and needs, and suggest possible intervention. These AI-educator agents also communicate with an AI-researcher agent, so that these jointly-created feedback/narratives about a student will be better grounded in both teachers' classroom practice and assessment data for the later teacher co-design. This study addresses the following research questions:

RQ-1: How to extract explainable features that can be mapped into natural languages, so that AI agents can understand?

RQ-2: How to create a coherent crew of AI agents that produce feedback based on empirical data?

RQ-3: How to evaluate whether AI agents' outputs are consistent with research findings?

The project intersects with current advancements in AI and education technologies to give back more data insights to educators to bridge assessment outcomes and learning needs. Deep data insights from LSAs provide indicators of broader student attributes (time management, test navigation regulation, engagement, learning needs) beyond a performance score, which offers rich information for teachers to prepare for personalized intervention. The current study exemplifies an innovative AI application in measurement research. The use of distinct AI agent personas - representing teachers in varied contexts, a coach, and a researcher - demonstrates an attempt to model diverse expert reasoning and tackle the complexity of student data interpretation. The AI-crew-generated feedback will help accelerate and enrich the teacher co-design, so that AI-agents' results can be better communicated to and understood by real teachers, allowing them to endorse, reject, or revise the results to support their students.

In the following method section, we briefly introduce the data and insights produced from previous research, describe explainable feature creation to address RQ-1; describe a crew of multi-AI agents for our exploration to address RQ-2; and additional experiment with AI to address RQ-3. In the result section, we display examples of outputs from the AI crew, highlighting the diverse perspectives from the AI agents, the AI-refined student profiles, and other useful feedback to educators. In the result section, we also show the evaluation of outputs from AI crew. In the last section, we discuss the contributions of this study, its limitations, and the future directions.

2 Methods

2.1 Data

In this study, we used a subset of data that contained manually labeled preliminary student profiles produced from Guo et al.'s (2024b) using the National Assessment of Educational Progress (NAEP) Grade 8 Mathematics assessment. The NAEP multi-source data contain a student's item responses, item response times, number of item visits, digital tools uses, as well as the sequences of item navigation (i.e., how much time was spent on an item and in what order). For details, please refer to National Assessment Governing Board's (2020) for NAEP process data released for secondary analysis.

A human-centered AI (HAI) architecture was proposed for human experts and AI collaboration to produce preliminary profiles for over ten thousand students who took one NAEP math block. The proposed HAI framework (refer to Figure 1) is built on a three-step architecture. This structure underscores: firstly, the critical input of human knowledge in the data preprocessing; secondly, the application of AI algorithms (including machine learning, deep learning) to improve data analysis and identify patterns; and thirdly, the integration of AI's computational power (e.g. active learning) with human expert judgment to finalize the profiles. Researchers and content experts investigated the extracted features and visualization of the multi-source student data and created students' preliminary profiles with AI for all students.

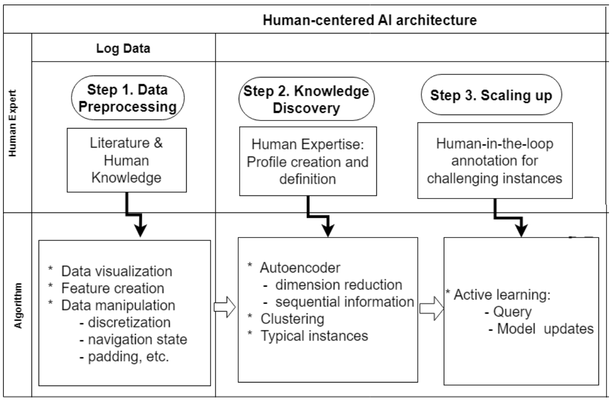


Figure 1: The human-centered AI architecture with three key steps from data preprocessing to scaling up to produce student profiles (Guo et al., 2024a,b).

2.2 Feature Creation and Mapping

In previous multi-source data studies (Guo et al., 2024a,b), deep learning models (i.e., autoencoders) were used to compress sequential data to produce latent features for profile prediction. Because of well-known challenges in latent feature interpretation, in the current study, we extracted features that were explainable from the multi-source data (refer to Table 1 for main features) to address RQ-1. These explainable features enabled mapping numerical values into natural languages for exploration of multi-AI agents. Features, not presented, also include mean and standard deviation of item visits and item scan, locations of longest bursts and longest jump, etc.

Among the explainable features, the new features, including navigation regularity, scan, scan burst, and jump, were created to address the challenge in describing a student's sequential navigation behaviors. Definitions of some of the new features are straightforward, as described in Table 1. Below we focus on the definition of the navigation regularity (Reg) feature which uses the concept of entropy to quantify and measure the unpredictability or randomness of a student's navigation behaviors when interacting with the test as a whole.

More specifically, let $Y = \{y_1, y_2, \cdots, y_{m+1}\}$ be the sequence of item numbers a student visited from the beginning of the test session to the end, where $m + 1$ is the total number of item visits; let $X = \{x_1, x_2, \cdots, x_m\}$ be the lag difference (i.e., $x_t = y_{t+1} - y_t$. Let $X^* = \{x_1^*, x_2^*, \cdots, x_m^*\}$ be the absolute value of X , where $x_t^* = |x_t|$.[1]

The entropy H of the sequence $X^* = \{x_1^*, x_2^*, \cdots, x_m^*\}$ is

$$H(X) = -\sum_{1}^{m}[p(x_i) * \log(p(x_i))],$$

where $p(x_i)$ is the probability of x_i. The navigation regularity (or simply, Regularity) is defined as:

$$\mathrm{Reg}(X) = \frac{1}{1 + H(X)}, \tag{1}$$

so that the upper bound of $Reg(X)$ is 1. That is, for students to have the value of $Reg(X) = 1$ on the test, they have to navigate the test very orderly (i.e., moving between adjacent items only). A value of $Reg(X)$ close to zero indicates unregulated navigation behaviors.

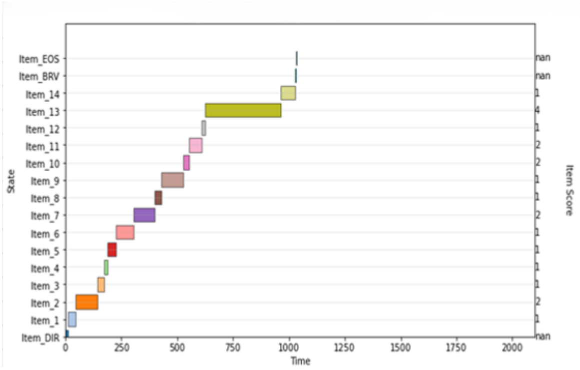


Figure 2: Navigation plots of student A. In the plot, the x-axis stands for the testing time, the y-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the y-axis on the right stands for the item score the student obtained. Each colored rectangle shows the time spent on an individual navigation item state (Guo et al., 2024b). The plot shows Student A with $Reg(X) = 1$.

A navigation plot is the visualization of three sequences (navigation item state, time on the state, and score received (Guo et al., 2024b). Refer to two examples in Figures 2 and 3, respectively, which shows two students' navigation patterns. One student (Student A; $Reg(X) = 1$) worked linearly one item at a time following the item presentation order on the test; the other (Student B;

[1]Taking absolute value is to conveniently define the jump event. A jump event occurs when a $X_t^* \geq 2$. Readers can modify these definitions based on their circumstances.

$Reg(X) = 0.29$) exhibited an irregular navigation pattern, showing behaviors such as quick item scans, skipping items, and jumping among items.

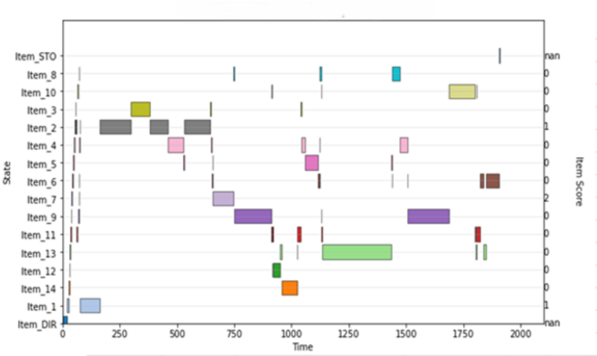


Figure 3: Navigation plots of Student B. In the plot, the x-axis stands for the testing time, the y-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the y-axis on the right stands for the item score the student obtained. Each colored rectangle shows the time spent on an individual navigation item state (Guo et al., 2024b). The plot shows Student B with $Reg(X) = 0.29$.

2.3 AI Agents

In this exploration, to address RQ-2, we created a crew of five AI agents embodying "teacher", "professional coach", and "researcher" personas (refer to Figure 4). Their roles logically build upon each other sequentially.

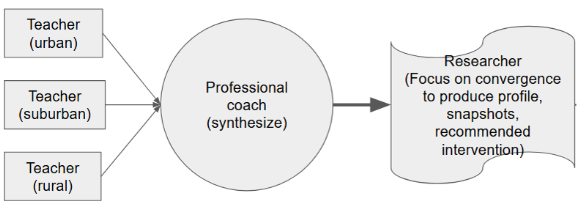


Figure 4: A crew of AI agents

The three math teacher AI-agents represent experienced teachers working in the urban, suburban, and rural settings, respectively, who help to pretest concepts, provide diverse contextual lenses, and react to an assessment idea from research results, flagging potential misunderstandings or concerns early on. Each teacher AI-agent is required to read the student data (i.e., features and preliminary profiles), reflect on their knowledge, and provide factual feedback to improve the preliminary profile, highlight students' strength and growth areas, and recommend potential intervention in a whole person approach.

The math coach AI-agent reads and synthesizes the three teacher AI-agents' results, including convergence and divergence in teachers' reports. The math coach's report may help to identify hypothetical points of friction or alignment in how each persona views data insights and practicalities from the assessment.

The scientist AI-agent reads the math coach's report, checks against the preliminary student profile, to ensure alignment and conciseness of the refined profiles, student's strength and needs, and recommended intervention.

This workflow (from multiple initial teacher analyses to coach synthesis, and then to research-informed summary), as shown in Figure 4, mimics a rigorous, collaborative human inquiry process, moving from divergent thinking to convergent, to provide meaningful collaboration in the teacher co-design.

2.4 Evaluation

To compare the AI-crew-refined profiles with the original manually-labeled preliminary profiles to address RQ-3, we conducted experiments using sentence embedding approaches and the GenAI agent approach. In the first approach, we cluster the sentence embedding results into ten clusters, and then evaluate the consistency between the embedding-generated clusters and preliminary profiles. In the second approach, we created an independent editing agent (with the persona of a meticulous editor and data analyst specializing in educational data). This AI-editor read and analyzed the refined profiles generated by the AI-crew, and then put them into ten clusters. Evaluation is carried out again on the consistency between the AI-Editor-generated clusters and preliminary profiles.

3 Results

Given the computation load of GenAI, in this study, we selected 50 students with manually-labeled preliminary profiles (five students in each profile) to explore the multi-AI agent application. Refer to Table 2 for the descriptions of the preliminary profiles, modified from those in (Guo et al., 2024b) based on explainable features introduced in this study.

To explore the multi-AI agent approach and prepare for the next-step teacher co-design, we used

Name	Description	Interpretation
Total score	Sum of item scores	Value range (discrete): [0, 21]. High value: good performance; low value: low performance. Most important feature, affecting interpretation of others.
Total time	Sum of item response times	Value range (continuous): (0, 1800]. Low value: less engaged; high value: issues in time management. Important feature, affecting interpretation of others
Total visit	Sum of item visits	Value range (discrete): [1, 113]. A student visiting all items just once has a value of 14. Low value: few item visits/less engaged; high value: issues in behavior regulation. Important feature, affecting interpretation of others
Not-reached (NR)	Number of not-reached items (no response time)	Value range (discrete): [0, 13]. High value: worked on few items; low value: worked on many items. Speeded: if non-zero NR & high time.
Rapid response (RR)	Sum of rapid-responded items	Value range (discrete): [0, 114]. Low value: less engaged; high value: issues in time management. RR: likely not spent adequate time to understand/work the item and associated with low effort.
Prolonged time(PL)	Sum of items with prolonged times (over 95 percentile)	Value range (discrete): [0, 4]. High value: likely struggling on high number of items. Non-zero value may indicate struggling, mostly due to lack of knowledge and skills to solve the problem(s), and subsequently likely led to NR items (i.e., test is speeded).
Navigation Regularity (Reg)	A measure to show whether a student mostly followed the order of item presentation on the test	Value range (continuous): [0.29, 1]. High value: orderly navigation through items (value of 1 indicates always moving forward or backward one item a time; no skipping around); low value: irregularly navigated through items. Also refer to Burst, scan, and jump related features for context.
Scan	Number of quick item scan behaviors in the entire session. Five seconds or less spent on an item is flagged as a scan behavior.	Value range (discrete): [0, 72]. High value: unregulated scan behaviors; low value: engaged with items (i.e, slow and steady win the race).
Longest scan burst	Number of longest scan behaviors in a burst	Value range (discrete): [0, 20]. High value may indicate global review, especially when its location is high; low value may indicate local review.

Table 1: Main features created and their interpretation.

Label	Description & Preliminary Profile
1	Attempted little to no items. Unengaged group
2	Very Low score, low/regular time, and regular visit behavior. Low engagement with very low performance, navigated through most items with low time
3	Low score, low/regular time, and regular visit behavior. Low engagement with low performance, navigated through most items with low time.
4	Low score, full/regular mixed time, and regular visit behavior. Engaged with low performance, navigated through most items, used mixed strategies
5	Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness
6	Low score, full/regular time with some prolonged item response times. Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness
7	Medium score, regular time and visit behavior, Medium performing group in all dimensions
8	Medium score, full/regular time with some prolonged item response times, and regular visit behavior. Medium performing, show strategic engagement behaviors (such as strategical response times)
9	High score, regular time and visit behavior. High performing group, expected navigation patterns.
10	Very high score, regular time and visit behavior. Highest performing group, expected navigation patterns

Table 2: The ten preliminary profiles, modified from (Guo et al., 2024b).

Azure OpenAI API (model: GPT-4o-mini; OpenAI, 2024) for its cost efficiency and CrewAI (Moura and contributors, 2024) for its easy implementation.

3.1 Example reports

To illustrate the work by the crew of the mulit-AI agents, we show the final outputs for two students (Student A and Student B).

Student A obtained a perfect score of 21 points, spent a total of 1029 seconds on the test, and visited all the 14 items linearly without any revisits, scan behaviors, or prolonged response times. The navigation regularity index is 1 (refer to the navigation plot in Figure 2). Student A's preliminary profile is 10 (i.e., *Very high score, regular time and visit behavior. Highest performing group, expected navigation patterns*).

Based on these facts of Student A, the AI crew provided the following feedback.

```
Key Strengths:
a. Strong understanding of mathematical concepts
and procedures.
b. Excellent problem-solving skills, reflected in high
scores and structured navigation.

Primary Growth Areas:
a. Continued exploration of advanced mathematical
concepts to prepare for high school.
b. Addressing potential gaps in resources or technology
access that may limit enrichment opportunities.

Critical Recommended Interventions:
a. Cognitive Support:
1. Implement differentiated instruction featuring
advanced math challenges to sustain engagement and
deepen understanding.
2. Introduce project-based learning opportunities that
integrate real-world applications of math.

b. Holistic Support (engagement, learning skills):
1. Establish mentorship programs connecting students
with advanced peers or professionals in STEM fields.
2. Organize community workshops focused on resilience,
stress management, and healthy coping strategies.
```

For Student A, the refined profile from the AI crew is

```
Very high cognitive performance with a perfect NAEP
score. Engages effectively with assessments and manages
time well. Exhibits strong self-regulation in navigation
behaviors. Resilient, supported by community influences.
Prepared for advanced mathematical challenges and
enrichment opportunities.
```

Student B obtained a total score of 2 out of the maximum 21 points, spent maximum total time allowed on the test; visited all 14 items but with 74 total visits. This student had one prolonged response time without any rapid responses. The longest scan burst is 39, and the largest jump is 10, with a navigation regularity of 0.29 (refer to the navigation plot in Figure 3). Student B's preliminary profile is 5 (i.e., *Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with/without speededness*).

Based on these facts of Student B, the crew provided the following feedback .

```
Key Strengths:
a. Demonstrated persistence and commitment to
engaging with all test items.
b. Ability to revisit questions, indicating a desire
for clarity and understanding.

Primary Growth Areas:
a. Need for strengthening foundational math skills
and conceptual understanding.
b. Development of effective time management
and test-taking strategies.

Critical Recommended Interventions:
a. Cognitive Support:
1. Implement targeted small group instruction focusing
on foundational math skills through real-world
applications.
2. Introduce structured practice sessions with timed
quizzes to improve pacing and time management skills.

b. Holistic Support (engagement, learning skills):
1. Foster a growth mindset by framing mistakes as
learning opportunities and encouraging reflective
discussions.
2. Create mentorship or peer tutoring programs to
provide emotional support and academic guidance.
```

For this student, the refined profile from the AI crew is

```
Very low cognitive performance in math,
high engagement with all test items, challenges in
time management and self-regulation, potential
struggles with anxiety, demonstrated resilience in
facing academic tasks, requires targeted support for
foundational skill development and emotional
resilience.
```

As shown in these examples, the outputs from the AI crew greatly enriched the interpretation of the student preliminary profile with depth, nuance, and the whole person learning perspective. These outputs will serve as a starting point for us to communicate with real teachers to collaborate on creating meaningful and actionable data insights for professional training.

3.2 Evaluation

We experimented several sentence embedding techniques, but results were unsatisfactory, mainly due to the fact that the available sentence embedding models in NLTK, without additional manipulations, did not differentiate the degree of importance for different features (e.g., Total Score has the upmost importance in profiling). Because of the space limit, results from the embedding approach are not presented.

Results from the AI-editor are presented in Table 3, that compares the clusters from the AI agent's analysis and the preliminary profile labels.

From Table 3, we observed that only one student's cluster was not consistent with the preliminary profile. That is, this student's preliminary

Table 3: Contingency Table for Preliminary Profile (Label) and AI-editor's Cluster (Cluster ID)

	Cluster ID										
Label	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**All**
1	0	0	0	0	0	0	0	0	0	5	5
2	0	0	0	0	0	0	0	0	5	0	5
3	0	0	0	0	0	0	0	5	0	0	5
4	0	0	0	0	0	0	5	0	0	0	5
5	0	0	0	0	0	5	0	0	0	0	5
6	0	0	0	0	4	1	0	0	0	0	5
7	0	0	0	5	0	0	0	0	0	0	5
8	0	0	5	0	0	0	0	0	0	0	5
9	0	5	0	0	0	0	0	0	0	0	5
10	5	0	0	0	0	0	0	0	0	0	5
All	5	5	5	5	4	6	5	5	5	5	50

profile (*6: Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness*) was classified into the adjacent preliminary profile by AI-Editor based on the AI Crew description (i.e. *5: Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness*). The major difference between these two preliminary profiles resides in navigation regularity, while students in Preliminary Profile 6 showed slightly better navigation behaviors (e.g., a higher value of Navigation Regularity Index). This discrepancy of one student's profile in Table 3 may indicate that it is challenging for AI agents to differentiate these two preliminary profiles.

4 Discussion and Conclusion

As AI continues to transform education and assessment practices, the current study explores the opportunity of using multi-AI agents to enhance, accelerate, and innovate measurement research to support education.

This multi-AI agents approach allows for rapid, low-cost exploration of diverse viewpoints, facilitating the identification of areas for deeper, evidence-based discussion with classroom teachers, as well as potential shortcomings in the research design. The outputs from the AI crew helps to develop a better teacher co-design study that aims at providing meaningful and actionable feedback to teachers from the big and rich LSAs' multi-source data.

In this multi-AI agent exploration, we found that AI crew (agents of teachers, math coach, and researcher) were able to enrich feedback, and their narratives were likely to be more grounded in teachers' knowledge and classroom practice than those preliminary profiles from research findings. This would help us to move one step closer to the teacher co-design to bridge the gap between assessment research and teacher practice.

There are a few observations worth mentioning in this exploration. First, even though we asked teacher agents in the crew to consider student data (features and preliminary profiles), we observed that final outputs still contain speculation, without data evidence, on why certain behaviors occurred on the assessment. For AI agents to generate factual profiles, we ended up with requiring every AI agent to refer back to student data, to ensure the final narratives were anchored in empirical data. That is, we used the empirical student data as a guardrail for AI agents to generate outputs. Another observation is the discrepancy between preliminary profiles and AI editor's analysis, which indicates that features, feature mapping, as well as AI agents' persona instructions in this study need to be improved. If AI agents have difficulties to differentiate some student profiles, they are likely to be challenging to real classroom teachers. This multi-AI agent exploration offers opportunities for us to improve our study design.

Overall, this study explored the use of multi-AI agents to prepare and accelerate the process of a teacher co-design for transforming research findings from LSAs to teaching practice. Based on data-driven student profiles obtained from NAEP multi-source data, we assembled a crew of AI agents that mimicked a rigorous human inquiry process to prepare for the teacher co-design. Built on previous studies, we proposed a few innovative approaches to link assessment-data-driven research that uses numerical features to AI agents that use natural languages. Among these innovations explored in this study, explainable feature creation is one of the key steps, which enables the mapping of numerical features into natural languages, and providing empirical data bases for AI agents to reason and produce factual feedback. Features associated with the visual navigation plot, particularly the navigation regularity index, will find wider applications in capturing a behavior process in assessment analytics, learning analytics, and other areas. Most importantly, this set of features will enable generalization of AI methodologies proposed in this study to other item blocks and even other tests in the future work.

Our exploration showed that the AI crew could enrich feedback and ground it in teachers' knowledge and practice, better preparing researchers for the real teacher co-design. Note that these AI-generated profiles are exploratory in the study. Given the increasing capabilities of GenAI, AI agent uses, evaluation, and validation need further research to empower researchers and educators. In addition, AI outputs in the current study need to be improved further by human experts and teachers. Meaningful understanding and valid insight still require direct engagement with actual teachers and researchers to capture genuine experiences and build trust for impactful assessment research and practice. Such AI systems, built on rich LSA data, research, and teacher co-designs, will be able to promote a more consistent and thorough initial analysis for all students' data, ensuring that feedback includes key factors (cognitive, engagement, learning skills) meaningful to guide teaching professional development with the evolving educational technologies. Collaboration between AI and human experts provides deeper analytical support at a larger scale than might be possible with human expertise alone for education innovation.

Acknowledgments

This work was supported, in part, by the Gates Foundation [INV-086180] and National Center for Education Statistics (NCES). Any opinions expressed in this work are those of the authors alone and shall not be attributed to the Gates Foundation, NCES, or ETS.

References

R. S. Baker and K. Yacef. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

Ryan Baker. 2021. *Artificial intelligence in education: Bringing it all together*, pages 43–54. OECD.

Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. Impact of ai assistance on student agency. *Computers & Education*, 210:104967.

Kadriye Ercikan, Hongwen Guo, and Qiwei He. 2020. Use of response process data to inform group comparisons and fairness research. *Educational assessment*, 25(3):179–197.

Kadriye Ercikan, Hongwen Guo, and Han-Hui Por. 2023. Uses of process data in advancing the practice and science of technology-rich assessments. In Natalie Foster and Mario Piacentini, editors, *Innovating Assessments to measure and support complex skills*, pages 211 – 228. OECD Publishing.

Kadriye Ercikan and James Pellegrino. 2017. *Validation of score meaning in the next generation of assessments: The use of response processes*. Routledge, New York, NY.

Samuel Greiff, Christoph Niepel, Ronny Scherer, and Romain Martin. 2016. Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61:36–46.

H. Guo, M. Johnson, K. Ercikan, L. Saldivia, and M. Worthington. 2024a. Large-scale assessments for learning: A human-centered AI approach to contextualize test performance. *Journal of Learning Analytics*, 11(2):229–245.

H. Guo, M. Johnson, L. Saldivia, M. Worthington, and K. Ercikan. 2024b. Human-centered ai for discovering student engagement profiles on large-scale educational assessments. *Journal of Measurement and Evaluation in Education and Psychology*, 30(12):282–301.

Hongwen Guo and Kadriye Ercikan. 2021. Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, 26(5-6):302–327.

International Association for the Evaluation of Educational Achievement. 2020. TIMSS 2023 international databasee.

Khan Academy. 2025. Keeping your streak alive: insights + tips from the last 6 months.

João Moura and contributors. 2024. CrewAI. Software library. Accessed: June 3, 2025. Please update year and version based on your usage.

National Assessment Governing Board. 2020. Response process data from the 2017 NAEP grade 8 mathematics assessment. Technical report, National Assessment Governing Board. Last accessed on June 2, 2025.

OpenAI. 2024. Generative pre-trained transformer 4 omni (gpt-4o). Model variant: GPT-4o mini. Accessed via Microsoft Azure.

Organisation for Economic Co-operation and Development. 2020. PISA 2018 database.

E. Pools and C. Monseur. 2021. Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assess Educ*, 9(10).

Steve Ritter, Ambarish Joshi, Stephen E. Fancsali, and Tristan Nixon. 2013. Predicting standardized test scores from cognitive tutor interactions. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, pages 169–176, Memphis, Tennessee, USA. International Educational Data Mining Society.

Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. 2025. Does multiple choice have a future in the age of generative ai? a posttest-only rct. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 494–504, New York, NY, USA. Association for Computing Machinery.

G. Zheng, S. E. Fancsali, S. Ritter, and S. Berman. 2019. Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, 6(2):153–174.

Nathan Zoanetti and Patrick Griffin. 2017. Log-file data as indicators for problem-solving processes. In Beno Csapo and Joachim Funke, editors, *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, chapter 11. OECD Publishing, Paris.

Long context Automated Essay Scoring with Language Models

Christopher Ormerod
Cambium Assessment Inc.
christopher.ormerod@cambiumassessment.com

Gitit Kehat
Cambium Assessment Inc.
gitit.kehat@cambiumassessment.com

Abstract

Transformer-based language models are architecturally constrained to process text of a fixed maximum length. Essays written by higher-grade students frequently exceed the maximum allowed length for many popular open-source models. A common approach to addressing this issue when using these models for Automated Essay Scoring is to truncate the input text. This raises serious validity concerns as it undermines the model's ability to fully capture and evaluate organizational elements of the scoring rubric, which requires long contexts to assess. In this study, we evaluate several models that incorporate architectural modifications of the standard transformer architecture to overcome these length limitations using the Kaggle ASAP 2.0 dataset. The models considered in this study include fine-tuned versions of XLNet, Longformer, ModernBERT, Mamba, and Llama models.

1 Introduction

Automated Essay Scoring (AES) is the application of statistical models to approximate the grading of essays by a human using a rubric. The initial models employed for AES were based on word frequencies and hand-crafted features (Page, 2003). The methods and models applied to AES have closely followed those used in more general Natural Language Processing (NLP) applications. The models employed in AES include recurrent and convolutional neural networks (Taghipour and Ng, 2016), models with attention mechanisms (Dong et al., 2017), and transformer-based large language models (LLM) (Rodriguez et al., 2019). Currently, LLMs are readily used to perform AES in research and large-scale assessment (Lottridge et al., 2023).

The first transformer-based LLM to be applied to AES was the Bidirectional Encoder-based Representations by Transformers (BERT) (Devlin et al., 2018). Since BERT arrived on the scene, the BERT model and its derivatives have readily provided state-of-the-art results in a wide range of downstream NLP tasks (Wang et al., 2019). The key to the success of these LLMs has been due to the transformer architecture (Vaswani et al., 2017) and to the ability to pretrain the model weights on a large corpus of unlabeled data on a semisupervised task such as next-token prediction (Radford et al., 2018) or masked-word prediction (Devlin et al., 2018). While we often say that the pretraining provides the model with some limited "understanding", the model weights are simply encoding enough information to encode the necessary word-probability functions.

Transformer-based models are deep feed-forward networks utilizing residual connections between layers that help stabilize training and prevent vanishing gradients (Vaswani et al., 2017). Each layer uses a multiheaded attention mechanism, similar to those used in recurrent networks (Graves et al., 2013). The input is defined by the addition of a positional embedding and a word embedding, which also defines the fixed length of the feedforward network. Since the computing power required by the attention mechanism scales quadratically with length, the length chosen for BERT was 512 (Devlin et al., 2018). This length became something of a standard for the most popular transformer-based LLMs.

The need for models that could overcome the limitations imposed by the transformer architecture became an active area of research shortly after BERT's release. We selected five different models that employ distinct approaches to addressing this challenge. These include versions of XLNet (Yang et al., 2019), Longformer (Beltagy et al., 2020), ModernBERT (Warner et al., 2024), Mamba (Gu and Dao, 2024), and a generative Llama model (AI@Meta, 2024) fine tuned for scoring using parameter-efficient methods (Xu et al., 2023). We give a brief explanation as to how each of these models addresses this limitation in

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 35–42
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

§2. The most novel of these approaches is applied in the Mamba model, which is the only pretrained language model in this study that uses the state-space model (SSM) (Gu et al., 2021). For SSMs, the computing power required scales linearly with the length of the input.

To understand the limitations of AES, researchers introduced the Automated Student Assessment Prize (ASAP) Dataset using the Kaggle platform (Shermis and Hamner, 2013). This Dataset consists of essay responses to eight prompts, some of which were assessed using trait scoring and some of which were assessed using a simpler holistic rubric. While this dataset became the definitive benchmark for AES methods, most essay responses possessed fewer than 512 tokens. This meant that, while LLMs showed superior performance with respect to traditional AES criteria (Williamson et al., 2012), the dataset did not adequately test the length issues that are often critical in the application of LLMs in large-scale assessment (Lottridge et al., 2023).

A second dataset, known as the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus (Crossley et al., 2022), which was originally designed to evaluate the performance of models that annotate the argumentative components of essays, was later extended to the Automated Student Assessment Prize v2 (ASAP 2.0) (Crossley et al., 2025). We will describe the dataset in more detail below, but many responses in the ASAP 2.0 dataset are too long for most language models.

This article is organized as follows: We use §2 to highlight the characteristically different approaches of the models chosen for this study. This is followed by §3 in which we describe the data used and the training methods. We have two different training regimes: one regime for classification models, such as those obtained by appending a classification, and another regime for generative LLMs. This is followed by the results in §4 and a discussion in §5.

2 Models

In this section, we discuss each model used in this study and why we chose to include it. We have attempted to illustrate if and how these models circumvent the architecturally imposed length restrictions of the standard transformer architecture.

2.1 DeBERTa

The DeBERTa model has a context length of 512. It has been chosen for this study to provide a strong benchmark for models typically used for AES. It is widely regarded as one of the best-performing models in a range of tasks. The model was trained as a discriminator, similarly to the ELECTRA models (Clark et al., 2020). The DeBERTa models also deviate from the standard BERT model by disentangling the word-embedding from the positional embedding (He et al., 2021).

2.2 Longformer

The Longformer model attempts to reconcile the need for local attention with a selective form of global attention. The local attention is applied in the form of a sliding window, similar to attention using convolutional units (Wu et al., 2019) coupled with a form of global attention only applied to special tokens (Beltagy et al., 2020), such as the beginning, ending, and mask tokens. This model still possesses a length limitation, however, by only using attention selectively, the computational burden is mitigated, allowing for pretraining over larger context lengths.

2.3 XLNet

The XLNet model uses the recurrent definition of attention introduced by the Transformer-XL model (Dai et al., 2019). These models have recently been discussed for essays, where the long context was useful in accurately annotating the argumentative components of essays (Ormerod et al., 2023). Almost all masked-language models are encoder-only models; however, the XLNet model is also distinguished as one of the few decoder models that was autoregressively pretrained as a masked-language model (Yang et al., 2019).

To demonstrate the recurrence, suppose any input sequence of length L is denoted $s_\tau = [x_{\tau,1}, \dots, x_{\tau,L}]$ while the hidden state for n-th layer associated with s_τ is $h_\tau^n \in \mathbb{R}^{L \times d}$. The recurrence relation defining $h_{\tau+1}^n$ as a function of h_τ^{n-1} and $h_{\tau+1}^{n-1}$ is given as follows:

$$
\begin{aligned}
\tilde{h}_{\tau+1}^{n-1} &= [SG(h_\tau^{n-1}) \circ h_{\tau+1}^{n-1}], && \text{(1a)}\\
q_{\tau+1}^{n} &= h_{\tau+1}^{n-1} W_q, && \text{(1b)}\\
k_{\tau+1}^{n} &= \tilde{h}_{\tau+1}^{n-1} W_k, && \text{(1c)}\\
v_{\tau+1}^{n} &= \tilde{h}_{\tau+1}^{n-1} W_v, && \text{(1d)}\\
h_{\tau+1}^{n} &= \mathrm{MHA}(q_{\tau+1}^{n}, k_{\tau+1}^{n}, v_{\tau+1}^{n}), && \text{(1e)}
\end{aligned}
$$

where SG is the stop gradient, $[x \circ y]$ is the concatenation operation of two sequences, and MHA is an abbreviation for the typical multiheaded attention mechanism for the transformer layer. The recurrence is built into the definition of $\tilde{h}^n_\tau$, affecting the keys and values. Digging deeper into (1) tells us that while the definition allows for infinite input lengths, there is a functional limitation of the architecture in which the output of any token is only a function of at most LD of the previous tokens where D is the depth of the network. The base and large pretrained models released with (Yang et al., 2019) has $L = 512$ and $D = 12$ and $D = 24$ respectively. This effectively caps the practical length to $6,000$ and $12,000$ for these models, respectively.

2.4 ModernBERT

The ModernBERT model is an encoder-based masked language model benefiting from much of the research that has been conducted since BERTs release (Warner et al., 2024). In particular, applications of generative LLMs have pushed the context length limitations in ways that the previous models stated above have not. The key to the context length of 8196 has been the Rotational Position Embedding (RoPE) (Su et al., 2024). There is a pretraining step in which the model is trained at short lengths with a large rotational component, then further trained on a model that interleaves rotational embedding with small and large rotational values to capture contributions from close and distant tokens. This method, developed in (Fu et al., 2024), was key to extending the context length for a range of popular models such as the herd of Llama models (AI@Meta, 2024).

2.5 Llama

The Llama series is a family of open-source generative LLMs from Meta (AI@Meta, 2024). The models have become as ubiquitously associated with open-source generative models as BERT was to masked language models. These generative models use RoPE (Su et al., 2024) in combination with the methods used to extend context lengths to 128k (Fu et al., 2024). In terms of architecture, the Llama models are a variant of the decoder-only transformer-based models, utilizing RMSNorm layers and a particular activated fully connected layer. We present this architecture in Figure 1, paying particular attention to the linear layers normalizing the input into the multi-headed attention (MHA) mechanism.

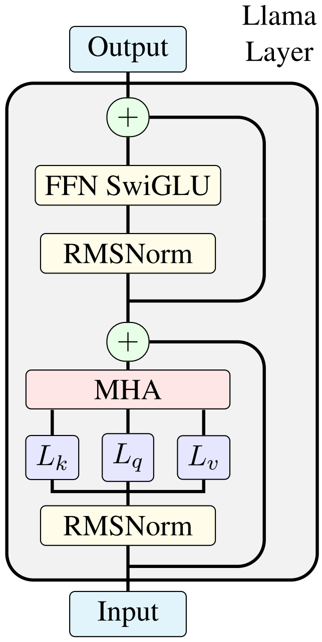


Figure 1: A layer of the Llama decoder-only architecture.

As a generative model, it was trained to predict the next token (Radford et al., 2018), followed by instruction tuning (Chung et al., 2022), followed by a reinforcement learning phase to make the models more useful (Kaufmann et al., 2024). These models come in a variety of sizes. The latest models include multi-modal capabilities; however, the models employed in this article are limited to text.

2.6 State-Space Models

This novel architecture completely replaces the transformer layer and attention with a simpler system based on discretizations of the state-space model (SSM). The SSM is a family of differential equations specified by the matrix equations

$$x'(t) = Ax(t) + Bu(t), \tag{2a}$$
$$y(t) = Cx(t) + Du(t), \tag{2b}$$

where x, u, and y are vectors and A, B, C, and D are matrices. This is a class of models broadly used in control theory. A standard discretization of (2) provides us with the recurrence relation of the form

$$h_t = Ah_{t-1} + Bx_t, \tag{3a}$$
$$x = Ch_t. \tag{3b}$$

A Mamba Layer, in contrast with the Transformer Layer, uses (3) as one component in addition to linear projections, a convolutional layer, and activation functions, as shown in Figure 2.

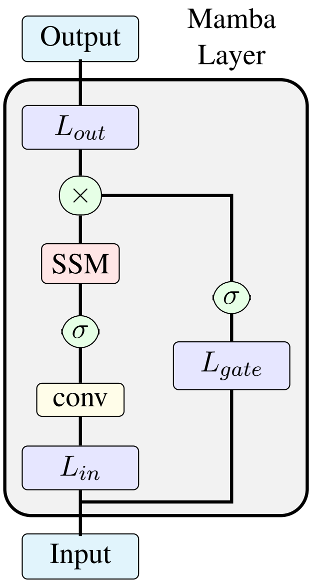


Figure 2: A single layer of the Mamba model.

The Mamba blocks can be computed with linear complexity, making them well-suited for long context tasks (Gu et al., 2021). This claim has been validated empirically by the superior performance of the Jamba models, which is an ensemble of transformer and Mamba layers (Lieber et al., 2024), on RULER benchmarks (Hsieh et al., 2024). As we seek longer and longer context lengths, models with linear complexity may be favorable from an efficiency standpoint.

2.7 Data

The reason we chose the ASAP 2.0 dataset (Crossley et al., 2025) is that this dataset provides a much-needed update of the original ASAP dataset (Shermis and Hamner, 2013), which could be considered to be saturated at this point. This dataset, derived as an extension of the PERSUADE corpus (Crossley et al., 2022), consists of essays written by students from grades 6 to 10 on a wide range of prompts.

Since a key feature of this study is our ability to handle long contexts, it is important to consider the length and grade level characteristics of the data. Because we are using a variety of LLMs, each of which has adopted different subword tokenizations (Kudo and Richardson, 2018), we have no unified notion of what defines a token. In lieu of a uniform tokenization, we will report the word count reported in the dataset. These length characteristics have been presented in Table 1.

	Train		Test	
Grade	Count	Avg. Words	Count	Avg. Words
6	2094	292.2	527	268.3
8	1648	339.9	921	295.9
9	4002	426.1	0	-
10	9563	385.8	5973	356.4
Total	17307	376.1	7421	342.7

Table 1: The size and length characteristics of the ASAP 2.0 dataset.

To evaluate the data, we use the standard metrics specified for AES (Williamson et al., 2012). The main metric used is the agreement statistic known as quadratic weighted kappa (QWK). Generally, the weighted kappa is specified by the equation

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad (4)$$

where $O_{i,j}$ is the observed agreement between the first rater giving a score of i and the second rater a score of j, and $E_{i,j}$ is the expected agreement only assuming the two raters' general distribution. This becomes QWK under the weighting

$$W_{i,j} = \frac{(i-j)^2}{(n-1)^2}$$

where n is the number of scores. It is generally understood that this is a measure of agreement above random chance, where a QWK of 1 is perfect agreement and -1 is perfect disagreement. In practical terms, lower scores represent the level of reliability between raters (McHugh, 2012), and our models should be compared against human-human agreements (Williamson et al., 2012). The QWK between the raters is reported to be 0.745

3 Methods

In order to perform essay scoring using LLMs, we distinguish two different cases. We call the first case traditional LLM-based scoring, where the underlying LLM is a masked-language model, such as BERT (Devlin et al., 2018), or a next word predictor such as the Generative Pretrained Transformer (GPT) (Radford et al., 2018). The second class of models considered was generative, which are distinguished by typically possessing an order of magnitude more parameters, and being trained in three phases: pretaining, instruction tuning, and reinforcement (OpenAI, 2023).

3.1 Traditional LLM based scoring

The typical procedure for traditional scoring is to convert a next word or masked word prediction model into a classifier by removing the linear head that would otherwise predict a token and append, in its place, a classification head with as many targets as there are scores (Rodriguez et al., 2019). The classification head is randomly initialized.

To train each of these models, 10% of the training set was designated as a development set. The models were trained by applying the Adam optimizer with a weight decay mechanism (Loshchilov and Hutter, 2019) to the cross-entropy loss function. An initial learning rate of 10^{-6} and a linear learning rate scheduler that reduces the learning rate to 0 over 10 epochs was used with a batch size of either 4 or 1 due to the length of some essays. The QWK was optimized on the development set using an early stopping mechanism.

To fine-tune our Mamba models for classification, we appended a learnable classification head, however, we were required to effectively freeze the weights associated with the SSM, L_{gate}, and the convolutional layer (See Figure 2). Full model training seemed to readily lead to model collapse, perhaps due to the requirement that certain weights take a particular form (Gu et al., 2021). Hence, we fine-tuned the embedding layer and the associated L_{in} and L_{out} weights of every layer. This is a memory-efficient way to fine-tune that provides excellent results. We used the Adam optimizer above with a learning rate of 10^{-5} and a batch size of 8.

3.2 Generative LLM based scoring

Many attempts in the literature seek to optimize the prompting of closed-source generative models to yield higher agreement rates (Xiao et al., 2024). While this is an interesting approach, we believe fine-tuning is necessary to obtain reasonable success. Due to the large size of the models, in order to do this with reasonable computational resources, we need to employ parameter-efficient methods (Xu et al., 2023). These methods can be applied without reference to an API and, hence, can be effectively employed securely, and privately, generating a fraction of the carbon emissions (Bulut et al., 2024).

In the case of fine-tuning generative models, the dataset used mimics an instruction set the model has been trained on. This means that any element of the training set appears to be a user prompting the model to score an essay to a rubric (Ormerod and Kwako, 2024). To do this, we used the following prompt template:

User

```
Assign a **Score** to the
**Essay** using the **Rubric**
provided.

**Rubric**: {rubric}

**Essay**:
```

Assistant

```
**Score**: {score}
```

This template highlights the important aspects by using markdown, due to the formatting of the corpus the model was trained on. Given that variations in prompting can have a significant bearing on the results, we exploit this by allowing the model to summarize and rephrase the rubric in 20 different ways. We optimized the variation of the rubric by evaluating the QWK of the model before fine-tuning on a development set that consisted of 10% of the training set.

We apply the method of low-rank adapters (Hu et al., 2021) and quantization (QLoRA) by (Dettmers et al., 2023). To apply QLoRA to a model, we must specify which linear layers to apply the adapter to, the rank of the adapter, scaling factors, the usual learning rate, and batch size. Concerning Figure 1, we seek to apply low-rank adapters to L_q, L_k, and L_v in the Llama model.

4 Results

The study evaluated various long-context language models on the ASAP 2.0 dataset to assess their effectiveness in automated essay scoring (AES). Models tested included traditional encoder-only architectures like DeBERTa-Base and XLNet-Base, extended-context models such as Longformer and ModernBERT, a state-space model (Mamba-130m), and generative decoder-based models like Llama-3.2-8B.

Model	Reference	L	Model Size	Overall	Grade 6	8	10
Human	(Crossley et al., 2025)	inf		0.745			
DeBERTa-Base	(He et al., 2021)	512	183M	0.790	0.696	0.659	0.800
XLNet-Base	(Yang et al., 2019)	8k*	110M	0.784	0.654	0.640	0.798
Longformer	(Beltagy et al., 2020)	4k	149M	0.798	0.698	0.658	0.811
ModernBERT	(Warner et al., 2024)	8k	149M	0.790	0.639	0.658	0.804
Mamba-130m	(Gu and Dao, 2024)	8k*	130M	0.797	0.674	0.640	0.812
Llama-3.2-8B	(AI@Meta, 2024)	8k	8B	0.792	0.667	0.672	0.803

Table 2: The performance of each model in terms of QWK, given by (4). These context lengths for XLNet models and Mamba models are not specified. The value of 8k was implemented as a mechanism to bound the memory required for training.

Human-human rater agreement stood at 0.745, serving as the baseline for comparison. All models surpassed this baseline, with Longformer achieving the highest overall QWK of 0.798. Notably, Mamba-130m performed competitively despite its smaller parameter size, demonstrating that linear-complexity models can rival attention-based transformers in AES tasks. Key findings revealed that long-context models, particularly those using advanced architectural innovations like RoPE-based positional embeddings and selective state spaces, are well-suited for handling lengthy student essays. Traditional models like DeBERTa and XLNet showed strong performance but lagged slightly behind Longformer and Mamba. Despite their large parameter counts and sophisticated training methods – such as instruction tuning and reinforcement learning —- generative models did not significantly outperform encoder-based models. However, they do offer the promising capability of providing feedback (Ormerod and Kwako, 2024).

5 Discussion

Overall, the results affirm the viability of long-context models in automated scoring systems, especially when dealing with complex, lengthy texts where global coherence and argument structure are crucial. Using long context models should not be about getting higher agreement, but rather addressing a glaring flaw from a modeling perspective; it is difficult to argue that traditional language models are faithfully modeling aspects of the rubric, such as organization, when essays are being truncated at 512 tokens.

Our modeling results indicate that both the selective attention mechanism and Mamba's linear complexity architecture deliver robust AES performance on lengthy texts. The study's most notable finding is Mamba's exceptional performance despite its simplified architecture. These differences between these models also suggest a potential for ensemble approaches. Several factors position Mamba and related architectures like Jamba (Lieber et al., 2024) as compelling alternatives for large-scale assessment applications. The linear scaling relationship between computational complexity and sequence length offers significant advantages over traditional transformer architectures. Additionally, optimized implementations may achieve 2-8x speed improvements compared to transformer-based models. These efficiency gains, combined with demonstrated effectiveness on long-context tasks, make state space models like Mamba practical solutions for automated assessment and similar applications requiring efficient processing of extended sequences.

References

AI@Meta. 2024. Llama 3 Model Card.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint*. ArXiv:2004.05150 [cs].

Okan Bulut, Maggie Beiting-Parrish, Jodi M. Casabianca, Sharon C. Slater, Hong Jiao, Dan Song, Christopher M. Ormerod, Deborah Gbemisola Fabiyi, Rodica Ivan, Cole Walsh, Oscar Rios, Joshua Wilson, Seyma N. Yildirim-Erbasli, Tarid Wongvorachan, Joyce Xinle Liu, Bin Tan, and Polina Morilova. 2024. The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *arXiv preprint*. ArXiv:2406.18900.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*. ArXiv:2210.11416 [cs].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Technical Report arXiv:2003.10555, arXiv. ArXiv:2003.10555 [cs] type: article.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54:100667.

Scott Andrew Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A Large-Scale Corpus for Assessing Source-Based Writing Quality: Asap 2.0.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv preprint*. ArXiv:1901.02860 [cs, stat].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical Report arXiv:1810.04805, arXiv. ArXiv:1810.04805 [cs] type: article.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data Engineering for Scaling Language Models to 128K Context. *arXiv preprint*. ArXiv:2402.10171 [cs].

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with Deep Bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278.

Albert Gu and Tri Dao. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint*. ArXiv:2312.00752 [cs].

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently Modeling Long Sequences with Structured State Spaces.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint*. Number: arXiv:2111.09543 arXiv:2111.09543 [cs].

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv preprint*. ArXiv:2404.06654 [cs].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A Survey of Reinforcement Learning from Human Feedback. *arXiv preprint*. ArXiv:2312.14925.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, and 3 others. 2024. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv preprint*. ArXiv:2403.19887 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv preprint*. ArXiv:1711.05101 [cs, math].

Susan Lottridge, Chris Ormerod, and Amir Jafari. 2023. Psychometric Considerations When Using Deep Learning for Automated Scoring. In *Advancing Natural Language Processing in Educational Assessment*. Routledge. Num Pages: 16.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Christopher Ormerod, Amy Burkhardt, Mackenzie Young, and Sue Lottridge. 2023. Argumentation Element Annotation Modeling using XLNet. *arXiv preprint*. ArXiv:2311.06239 [cs].

Christopher Michael Ormerod and Alexander Kwako. 2024. Automated Text Scoring in the Age of Generative AI for the GPU-poor. *arXiv preprint*. ArXiv:2407.01873 [cs].

Ellis Batten Page. 2003. Project Essay Grade: PEG. In *Automated essay scoring: A cross-disciplinary perspective*, pages 43–54. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. *arXiv preprint*. Number: arXiv:1909.09482 arXiv:1909.09482 [cs, stat].

Mark D. Shermis and Ben Hamner. 2013. Contrasting State-of-the-Art Automated Scoring of Essays. pages 335–368. Publisher: Routledge Handbooks Online.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Technical Report arXiv:1804.07461, arXiv. ArXiv:1804.07461 [cs] type: article.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint*. ArXiv:2412.13663 [cs].

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2011.00223.x.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay Less Attention with Lightweight and Dynamic Convolutions. *arXiv preprint*. ArXiv:1901.10430 [cs].

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint*. ArXiv:2401.06431 [cs] version: 1.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *arXiv preprint*. ArXiv:2312.12148 [cs].

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Optimizing Reliability Scoring for ILSAs

Ji Yoon Jung **Ummugul Bezirhan** **Matthias von Davier**
TIMSS & PIRLS International Study Center at Boston College
{jiyoon.jung, bezirhan, vondavim}@bc.edu

Abstract

This study proposes an innovative method for evaluating cross-country scoring reliability (CCSR) in multilingual assessments, using hyperparameter optimization and a similarity-based weighted majority scoring within a single human scoring framework. Results show that this approach provides a cost-effective and comprehensive assessment of CCSR without the need for additional raters.

1 Introduction

Constructed response (CR) items are valued for their ability to assess students' higher order thinking skills, offering deeper insights into student performance compared to multiple choice items (Livingston, 2009; Scully, 2017). However, their widespread use in large-scale assessments has been constrained by concerns about human scoring reliability. While extensive rater training and structured scoring protocols can enhance inter-rater reliability, rater effects such as leniency, severity, and the halo effect often persist (Myford & Wolfe, 2003; Yamamoto et al., 2017).

These scoring challenges are particularly pronounced in international large-scale assessments (ILSAs). In multilingual contexts, achieving high consistency among human raters from diverse cultural and linguistic backgrounds is difficult, even with centralized scoring guides (Wang & Li, 2020). The substantial time, effort, and resources required for global human rater training, scoring vast numbers of responses, and monitoring scoring procedures across multiple countries further complicate the process.

Cross-country scoring reliability (CCSR), designed to measure international scoring consistency (von Davier et al., 2023) in the Progress in International Reading Literacy Study (PIRLS), exemplifies these challenges. This valuable measure operates as a separate, additional burden alongside the main scoring process and encounters significant logistical hurdles. It evaluates scoring consistency using a common set of 200 English language responses for specific PIRLS reading items, but its scope is critically limited to human raters who are either native English speakers or proficient in English. Consequently, the conventional CCSR approach assesses a narrow subset of responses and relies on an underrepresented rater pool. This restricts its ability to provide a comprehensive assessment of scoring consistency across the full range of CR items and participating countries.

To address these logistical and methodological limitations, we recently proposed a novel reliability scoring framework that combines similarity-based majority voting (Jung et al., under review).

The current study focuses on the systematic optimization of that framework through hyperparameter tuning while also providing a transparent step-by step implementation of the full pipeline. This method aims to offer a more efficient and reliable measure of cross-country scoring consistency, reducing dependency on extensive human rater resources.

2 Background

Human scoring in multilingual assessments presents significant challenges, primarily due to difficulties in maintaining consistency across different human raters, languages, and countries (Jung et al., 2025; Okubo et al., 2023). The inherent linguistic and sociocultural diversity among raters may influence the interpretation of student responses and the application of scoring guides, introducing systematic variance in scoring outcomes (Ercikan & Por, 2020; Wang & Li, 2020).

Double or multiple scoring by independent raters is a foundational practice in educational

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 43–49
October 27-29, 2025

measurement for ensuring scoring consistency. However, this approach is costly and time-intensive, requiring the recruitment and training of multiple raters for every item and response (Fliss et al., 1981; Gwet, 2014; Wiggins, 1990).

Alternative cost-saving strategies have emerged to alleviate these resource constraints. One common approach is to double score only a randomly selected subset of responses, though this strategy may be suboptimal when the precise classification of students into performance levels is critical (Finkelman et al., 2009). Alternatively, targeted double scoring (TDS) focuses on responses falling near the critical score range (e.g., pass/fail cutoff), aiming to improve scoring accuracy and reliability (Finkelman et al., 2009; Miao et al., 2023; Sinharay et al., 2022). However, the effectiveness of TDS depends on the accurate identification of the critical score range. Xu and Wind (2025) also found no notable psychometric advantage for TDS over random double-scoring approaches.

Importantly, double or multiple scoring, whether applied to all responses or a subset, substantially increases costs and time compared to single human scoring, creating a persistent tension between scoring quality and practical feasibility. This study explores a novel strategy to optimize reliability scoring within a single human scoring framework, achieving cost-effective and comprehensive measurement without the need for additional human scoring.

3 Method

3.1 Dataset

The PIRLS assesses fourth-grade students' reading comprehension in more than 50 countries globally on a five-year cycle since 2001. In PIRLS 2021, approximately half of the participating countries (n=27) transitioned to computer-based testing (digital PIRLS). From the 18 items with reported CCSR values in PIRLS 2021, we selected 2 two-point CR items, using data from all countries participating in digital PIRLS (see Table 1). These two-point items were selected as they are the only two-point "trend" items that will be reused for PIRLS 2026, and this study supports PIRLS 2026 scoring preparation. Notably, one item exhibited the most problematic CCSR of 0.768, making it a challenging yet ideal candidate for validating our new reliability scoring approach.

Item	Process	*N*	CCSR
1	Focus on and retrieve	14,875	0.868
2	Straightforward inferences	14,151	0.768

Table 1: PIRLS trend items used in the study

3.2 Multilingual Response Translation

We utilized a standardized prompt template with GPT-4o to translate non-English responses into English and to rectify spelling and grammatical errors in English responses using GPT-4o (i.e., gpt-4o-2024-08-06). The prompt template incorporated four key components, as detailed in Table 2 (Jung et al., under review). This Zero-Shot-Chain-of-Thought (Zero-Shot-CoT) is task-agnostic (Kojima et al., 2022), enabling its application across diverse items to generate contextually appropriate translations.

Component	Content
Instruction	Comprehensive guidance on AS
Reading passage	A written text serving as the stimulus
Question	A question consisting of one or two sentences
Scoring guide	Rubric for scoring an item, including descriptions and examples

Table 2: PIRLS scoring template components

3.3 Response Flagging and Auto-Scoring

Following translation, we implemented a two-stage data flagging process. First, untranslated responses were flagged as 'missing' and excluded from subsequent analysis. Second, semantically meaningless responses were flagged as 'meaningless', assigned a score of 0, and retained as valid responses for analysis (included in the weighted majority scoring). Detailed criteria for each flagging stage are provided below.

Missing Flagging: Responses were classified as 'missing' if they met either of two criteria: (1) GPT-4o explicitly marked them as 'untranslatable' during translation, or (2) their English vocabulary was less than 75% of tokenized words. This missing flag was only applied to responses exceeding 8 characters. Linguistic preprocessing included lower-casing, lemmatization, and tokenization by spaCy's en_core_web_lg model in Python. The English vocabulary percentage was calculated using the PyEnchant dictionary. Proper nouns (e.g., "California" or "Marie"), identified via spaCy's Named Entity Recognition, counted as valid English vocabulary.

Meaningless Flagging: After excluding missing responses, we flagged 'meaningless' responses if they were: (1) extremely short or (2) semantic outliers. These responses were assigned a score of 0 but retained in the dataset. Very short responses were defined as those with a normalized translation length L_i<0.03, representing the bottom 3% of the length distribution. Translation length was normalized using min-median normalization to mitigate the impact of extreme outliers:

$$L_i = \frac{l_i - \min(l)}{median(l) - \min(l)} \quad (1)$$

where l_i is the length of the translated response *i*.

Semantic outliers were identified through a multi-faceted assessment. First, responses with a coherence score (C_i) below 0.20 are flagged. C_i was computed as the average cosine similarity between the embedding of response *i* and the embeddings of all other responses, excluding self-similarity:

$$C_i = \frac{1}{N-1} \sum_{i \neq j}^{N} sim(E_i, E_j) \quad (2)$$

where $sim(E_i, E_j)$ is the cosine similarity between embeddings of response i and *j*. Response embeddings were generated using the Sentence Transformer model (all-MiniLM-L6-v2) in Python.

Second, responses with a meaningfulness score (M_i) below m were also identified as semantic outliers. The meaningfulness threshold m was determined following the hyperparameter optimization. M_i integrates both coherence and normalized length with weights:

$$M_i = 0.80 \times C_i + 0.20 \times L_i \quad (3)$$

M_i was examined when responses were deemed semantic outliers if the average cosine similarity of their top k most similar responses (as determined during the hyperparameter optimization phase) fell below 0.80.

3.4 Reliability Scoring with Optimal Hyperparameters

Our reliability scoring approach scored responses using a weighted majority scoring algorithm based on cosine similarity between response embeddings.

Similarity Measurement: Response embeddings were generated using the all-MiniLM-L6-v2 model, and cosine similarities were calculated between all response pairs. For each response i, we identified the top k most similar responses based on the highest cosine similarities, where k is a hyperparameter optimized through grid search.

Weighted Majority Scoring: For each response *i*, the majority score $s^* \in \{0, 1, 2\}$ was determined as:

$$s^* = \arg max_s \left(W_{is} = \sum_{j \in S_{is}} sim\left(E_i, E_j\right)\right) \quad (4)$$

where S_{is} is the set of the top *k* similar responses (neighbors) to response *i* with human score *s*. The score s^* was assigned only if its proportion of the total weighted score exceeds the weight threshold *WT*, which was optimized via grid search. Otherwise, the response was flagged as 'inconsistent' if the proportion fell below *WT*, indicating that human scores among similar responses varied too widely to assign a reliable majority score.

$$\frac{W_{is^*}}{\sum_s W_{is}} > WT \quad (5)$$

Hyperparameter Tuning via Grid Search: We conducted a systematic grid search over $k \in \{1, 2, 3, 4, 5, 10, 15\}$ (number of similar responses) and $WT \in \{0.60, 0.65, 0.70, 0.75\}$ (weight threshold) to optimize the reliability scoring. All 28 unique hyperparameter combinations were examined using Python's itertools.product.

3.5 Evaluation

The grid search evaluated each hyperparameter combination based on two criteria: (1) minimizing the proportion of responses labeled as 'inconsistent', and (2) maximizing weighted exact agreement (Weighted EA).

Weighted EA quantifies the agreement between human and majority scores, assigning more weight to matches (where human score equals majority score) that exhibit higher cosine similarity. It was calculated as the ratio of the sum of average cosine similarities for responses with matching to the sum of average cosine similarities for all responses. After determining optimal values for *k* and *WT*, several meaningfulness thresholds (*m*) were tested to identify the optimal threshold for detecting semantic outliers. The appropriateness of each threshold was evaluated by analyzing human score distributions, with accurate flagging confirmed by human scores of 0.

Following the hyperparameter optimization, the optimized reliability scoring was analyzed in detail, focusing on the majority score (s^*) distribution and cosine similarity statistics.

4 Results

Hyperparameter Optimization: The grid search results identified the optimal hyperparameter setting as WT=0.60 and k=3, which minimized the inconsistency proportion and maximized the weighted EA, as detailed in the Appendix. Under this configuration, the inconsistency proportions were very low (0.80% for Item 1 and 2.02% for Item 2), and the weighted EAs (0.881 for Item 1 and 0.755 for Item 2) closely aligned with their corresponding CCSR values (0.868 for Item 1 and 0.768 for Item 2).

Using the optimal hyperparameters (WT=0.60 and k=3) along with m = 0.30, we achieved highly accurate detection of semantic outlier responses, as shown in Tables 3 and 4. For Item 1, 99.40% of responses flagged as ‘meaningless’ received a human score of 0, compared to 87.08% for Item 2. The reduced detection accuracy for Item 2 was anticipated, as it showed the most significant CCSR issues in PIRLS 2021 (CCSR = 0.768), suggesting inconsistent cross-country scoring, or a higher prevalence of borderline responses susceptible to scoring variations across countries and languages. Given the more reliable performance of Item 1, we adopted m = 0.30 for our optimized reliability scoring.

Meaningfulness (m)	Human Score (%)		
	0	1	2
0.25	99.02	0.98	0.00
0.26	99.14	0.86	0.00
0.27	99.23	0.77	0.00
0.28	99.30	0.70	0.00
0.29	99.36	0.64	0.00
0.30	99.40	0.60	0.00

Table 3. Human score distribution for ‘meaningless’ responses to Item 1

Meaningfulness (m)	Human Score (%)		
	0	1	2
0.25	93.94	4.94	1.12
0.26	91.86	6.86	1.29
0.27	91.10	7.57	1.33
0.28	90.69	8.10	1.21
0.29	88.13	10.16	1.71
0.30	87.08	10.83	2.09

Table 4. Human score distribution for ‘meaningless’ responses to Item 2

Reliability Scoring Assessment: First, we examined the majority score distribution (s^*), as presented in Table 5. The average proportions of inconsistent and missing responses were 1.41% (n=203) and 1.51% (n=218), respectively. This indicates that our reliability scoring approach effectively assigned scores to most responses (97.69% for Item 1 and 96.48% for Item 2) by leveraging their top three most similar neighbors. As expected, Item 2 exhibited a slightly higher inconsistency proportion of 2.02%, consistent with its problematic CCSR. The proportion of missing responses was also low across both items, suggesting that GPT-4o demonstrated a strong capability in translating non-English language responses, including those from low-resource languages such as Arabic, Lithuanian, and Slovak, into English.

Majority score	Item 1		Item 2	
	n	%	n	%
0	4314	29.00	5049	35.68
1	3356	22.56	6364	44.97
2	6862	46.13	2240	15.83
Inconsistent	119	0.80	286	2.02
Missing	224	1.51	212	1.50

Table 5. Majority score distribution

Next, we analyzed cosine similarity statistics to assess the effectiveness of our reliability scoring in capturing semantically similar responses, both across all responses and within each response’s top three similar neighbors (see Table 6). The mean of average cosine similarities was high, at 0.932 for Item 1 and 0.891 for Item 2, with standard deviations below 0.1, indicating very low variability across responses (see Figures 1 and 2). Additionally, the top three cosine similarities per response tend to be tightly clustered, with very low standard deviation reflecting minimal internal semantic variability among each response's nearest neighbors. These demonstrate the robust performance of our reliability scoring in detecting semantically coherent neighbors.

Item	Mean avg cos sim	SD of avg cos sim	Avg SD of top 3 cos sim
1	0.932	0.098	0.007
2	0.891	0.095	0.012

Table 6. Statistics on average cosine similarity

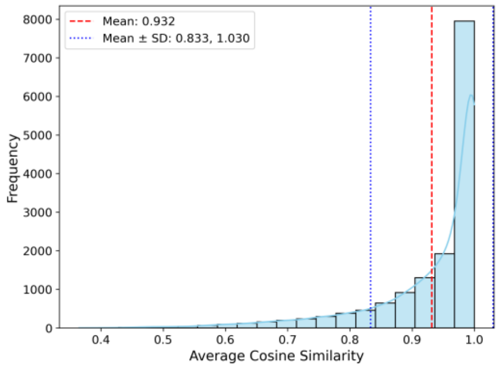


Figure 1. Average cosine similarity for Item 1

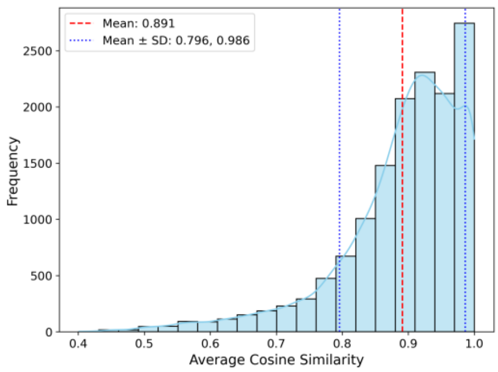


Figure 2. Average cosine similarity for Item 2

5 Discussion

Our findings demonstrate that optimized reliability scoring can effectively evaluate CCSR in multilingual contexts without requiring additional human raters. Although double or multiple scoring has traditionally been the gold standard for achieving consistency (Williamson et al., 2012), prior research (Sinharay et al., 2023; Song & Lee, 2022; Wiggins, 1990) highlights its resource-intensive nature and associated practical and methodological challenges. Our method provides a resource-efficient alternative, utilizing initial human scoring with all responses (over 14,000 responses per item) to achieve results comparable to established CCSR practices. Moreover, this approach enables a comprehensive assessment of individual countries' scoring practices on a global scale using weighted EA or kappa statistics disaggregated by country and language. This facilitates the detection of possible scoring inconsistencies in specific countries or languages and the identification of problematic items (Jung et al., under review).

Despite these promising results, this study has limitations. First, we examined only two two-point "trend" items with available CCSR values, selected for the PIRLS 2026 scoring preparation. Future studies should examine the scalability of this approach across a wider range of item types, including both one- and two-point items. Second, while our approach successfully identified the three most similar neighbors for all responses, responses with low average cosine similarity require further scrutiny. Specifically, responses assigned an initial human score of 2 but exhibiting very low average cosine similarity scores may indicate initial human scoring errors, limitations in our reliability scoring, or both. These cases warrant review by content experts to better understand the sources of scoring discrepancies.

6 Conclusion

This study highlights the effectiveness of optimizing reliability scoring through key hyperparameter optimization and a similarity-aided weighted majority scoring method. This approach robustly measures cross-country consistency by leveraging initial human scoring alongside all responses, offering a more inclusive and cost-effective alternative to existing CCSR. Our novel approach provides a valuable measure for evaluating scoring consistency on a global scale, enabling more accurate and reliable reporting to participating countries.

References

Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural assessment contexts. Comparability of large-scale educational assessments: Issues and recommendations, 205-225.

Finkelman, M., Darby, M., & Nering, M. (2009). A two-stage scoring method to enhance accuracy of performance level classification. Educational and Psychological Measurement, 69(1), 5-17.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement. Statistical methods for rates and proportions, 2(212-236), 22-23.

Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Jung, J. Y., Tyack, L., & von Davier, M. (2025). Towards the Implementation of Automated Scoring in International Large-scale Assessments: Scalability and Quality Control. Computers and Education: Artificial Intelligence, 100375.

Jung, J. Y., Tyack, L., & von Davier, M. (under review). Optimizing Automated Scoring in ILSAs with Prompt Compression Computers and Education: Artificial Intelligence.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Livingston, S. A. (2009). Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11. Educational Testing Service.

Miao, J., Sinharay, S., Kelbaugh, C., Cao, Y., & Wang, W. (2023). Evaluating targeted double scoring for the performance assessment for school leaders using imputation and decision theory. ETS Research Report Series, 2023(1), 1-10.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of applied measurement, 4(4), 386-422.

Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, C. S., & Bastianic, T. (2023). AI scoring for international large-scale assessments using a deep learning model and multilingual data.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. Practical Assessment, Research and Evaluation (PARE), 22(1), 1-13.

Sinharay, S., Johnson, M. S., Wang, W., & Miao, J. (2023). Targeted double scoring of performance tasks using a decision-theoretic approach. Applied Psychological Measurement, 47(2), 155-163.

Song, Y. A., & Lee, W. C. (2022). Effects of Using Double Ratings as Item Scores on IRT Proficiency Estimation. Applied Measurement in Education, 35(2), 95-115.

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Eds.). (2023). Methods and Procedures: PIRLS 2021 Technical Report. Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/methods

Wang, Y., & Li, S. (2020). Issues, challenges, and future directions for multilingual assessment. Journal of language teaching and research, 11(6), 914-919.

Wiggins, G. (1990). The case for authentic assessment. Practical assessment, research, and evaluation, 2(1).

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated

Xu, Y., & Wind, S. A. (2025). Examining the Psychometric Impact of Targeted and Random Double-Scoring in Mixed-Format Assessments. Educational Measurement: Issues and Practice, 44(1), 18-30.

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2017). Developing a machine-supported coding system for constructed-response items in PISA. ETS Research Report Series, 2017(1), 1-15.

Appendices

A. Grid Search Results

Weight threshold	k	Inconsistency (%)	Weighted EA
	1	13.18	0.867
	2	13.14	0.800
0.60	3	0.80	0.881
0.65	3	6.70	0.880
0.70	3	5.27	0.753
0.75	3	5.55	0.753
0.60	4	6.06	0.851
0.65	4	0.97	0.851
0.70	4	6.70	0.851
0.75	4	10.66	0.805
0.60	5	8.70	0.861
0.65	5	8.75	0.825
0.70	5	19.70	0.825
0.75	5	6.70	0.825
0.60	10	10.66	0.861
0.65	10	11.33	0.837
0.70	10	12.76	0.820
0.75	10	19.70	0.788
0.60	15	13.23	0.853
0.65	15	10.66	0.833
0.70	15	15.43	0.806
0.75	15	17.64	0.772

Table 1. Grid search results on Item 1

Weight threshold	k	Inconsistency (%)	Weighted EA
	1	25.43	0.738
	2	26.92	0.604
0.60	3	2.02	0.755
0.65	3	2.54	0.753
0.70	3	40.85	0.509
0.75	3	40.88	0.509
0.60	4	18.18	0.670
0.65	4	18.18	0.670
0.70	4	18.19	0.670
0.75	4	30.73	0.584
0.60	5	15.10	0.690
0.65	5	28.49	0.605
0.70	5	28.50	0.605
0.75	5	28.50	0.605
0.60	10	20.01	0.661
0.65	10	28.29	0.604
0.70	10	33.91	0.569
0.75	10	42.51	0.502
0.60	15	22.61	0.643
0.65	15	29.35	0.595
0.70	15	39.31	0.524
0.75	15	48.88	0.451

Table 2. Grid search results on Item 2

Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment

Jill Burstein*
Ramsey Cardwell
Ping-Lin Chuang
Allison Michalowski
Steven Nydick
Duolingo
{jill, ramsey, pinglin, allison.michalowski, steven.nydick}@duolingo.com

Abstract

Practice tests for high-stakes assessment are intended to build test familiarity, and reduce construct-irrelevant variance which can interfere with valid score interpretation. Generative AI-driven, automated item generation (AIG) scales the creation of large item banks and multiple practice tests, enabling repeated practice opportunities. We conducted a large-scale observational study (N = 25,969) using the Duolingo English Test (DET)—a digital, high-stakes, computer-adaptive English language proficiency test to examine how increased access to repeated test practice relates to official DETscores, test-taker affect (e.g., confidence), and score-sharing for university admissions. To our knowledge, this is the first large-scale study exploring the use of AIG-enabled practice tests in high-stakes language assessment. Results showed that taking 1-3 practice tests was associated with better performance (scores), positive affect (e.g., confidence) toward the official DET, and increased likelihood of sharing scores for university admissions for those who also expressed positive affect. Taking more than 3 practice tests was related to lower performance, potentially reflecting *washback* – i.e., using the practice test for purposes other than test familiarity, such as language learning or developing test-taking strategies. Findings can inform best practices regarding AI-supported test readiness. Study findings also raise new questions about test-taker preparation behaviors and relationships to test-taker performance, affect, and behaviorial outcomes.

1 Introduction

For millions of international test takers, scores on high-stakes English language proficiency (ELP) assessments can profoundly impact their educational and professional goals. As a result, they engage in various test preparation strategies. For example, practice tests aim to build familiarity for a specific test; reading books and articles can improve English language reading skills; and, deliberate engagement in conversations with peers and instructors can strengthen English language speaking and listening skills.

This paper focuses on *practice tests*. Practice tests aim to build test familiarity to reduce test-design-related *construct-irrelevant variance* (CIV). CIV is associated with the introduction of factors unrelated to the skills a test is intended to measure (the *target construct*) (Messick, 1982; Powers, 1985). For instance, CIV can stem from unfamiliar technical features (e.g., *drag-and-drop*), lack of familiarity with the device required for taking a test (e.g., test requirements to use a laptop for test takers who have limited laptop experience (Koné et al. (2024)), or anxiety triggered by an unfamiliar format (Winke and Lim, 2017).

Conventional practice tests, often developed by testing organizations, aim to reduce CIV. However, they typically contain a limited number of fixed forms, restricting opportunities for repeated test practice. Modern generative AI-powered automated item generation (henceforth, AIG) alleviates this constraint by enabling the creation of large item pools for digital practice tests. As a result, practice test generation can be scaled to support repeated practice test opportunities for test takers.

The Duolingo English Test (DET)is a digital, AI-driven, high-stakes, computer-adaptive ELP assessment used for international student university admissions. The DET is taken by hundreds of thousands of test takers each year.

To help test takers become familiar with the test, the DET offers a *free* practice test that simulates the official DET. As such, the practice test provides exposure to the DET task types, mirroring the official test in both appearance and administration order. It also provides an estimated score range, giving test takers a sense of how they are likely to perform on

*Authors are listed alphabetically to reflect equal contributions.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 50–57

the official test. Like the official DET, the practice test is also computer-adaptive, but drawing from a separate item pool than the official test. The large practice-test item pool, enabled by AIG, is used to dynamically generate versions of the practice test with different item sets, offering test takers repeated opportunities for practice (Naismith et al., 2025).[1,2]

The study presented in this paper examines how access to repeated test practice (i.e., the number of tests taken)—enabled by AIG— relates to test-takers' official DET scores, test-taker affect (e.g., confidence), and test-takers' decision to share their official DET scores for university admissions.

2 Background

Language assessment research has examined various aspects of test preparation, including test-taker preparation preferences (O'Sullivan et al., 2021), the relationship between preparation and affect (such as anxiety) (Chang and Read, 2008; Powers and Alderman, 1983; Winke and Lim, 2017), and the link between preparation and test performance (Green, 2007; Knoch et al., 2020; Liu, 2014; Powers, 1985; Xie, 2013). These studies suggest that test preparation can reduce anxiety (Chang and Read, 2008; Powers and Alderman, 1983), increase confidence (Powers and Alderman, 1983), and improve test scores (Green, 2007; Knoch et al., 2020; Xie, 2013). Knoch et al. (2020) investigated repeat test takers, showing how they changed their test preparation strategies over time to try to improve their test score. Xie (2013) demonstrated how test takers use test preparation to develop strategies for score improvement. Green (2007) examined the comparative impact of test preparation courses for a high-stakes language assessment. These three studies highlight *washback effect* with regard to test preparation, whereby a test influences language teaching and learning (Messick, 1996).

Automated item generation research related to assessment and instruction is extensive, but much predates modern generative AI. For example, Mitkov et al. (2006) showed that NLP-assisted item generation with human review can be more time-efficient than manual creation. Heilman and Smith (2010) proposed a framework for automatically generating and evaluating questions from text, demonstrating the feasibility of transforming declarative sentences into fact-based questions. Similarly, Madnani et al. (2016) discussed the Language Muse system, which used NLP to generate reading comprehension exercises for U.S. middle school texts for English learners. More recent research has shifted toward evaluating item quality and comparing system performance using large language models. For instance, Laverghetta Jr and Licato (2023) investigated GPT-4 for test item generation, demonstrating its potential to create psychometrically valid items.[3]

AIG is now integrated into the development of digital, high-stakes language assessments. Specific to this paper, the official DET and its practice test are dynamically assembled using AIG-created item banks with human review (Attali et al., 2022). After generating items with prompts used to fine-tune the AIG, human experts conduct a review. To ensure item quality and appropriateness, a multistage process for human review is implemented. This process begins with automated checks for linguistic accuracy and social appropriateness, followed by human expert review focused on copyediting, fact-checking, and identifying potential fairness and bias issues that could disadvantage certain test-taker groups (Church et al., 2025).

An internally-developed review platform is used to coordinate item reviews, track reviewer performance, and ensure inter-rater consistency. The final items are used to automatically create the DET practice and *official* DET tests.

As mentioned earlier, prior research about test preparation for high-stakes assessment has studied test-taker preferences, and established links between test preparation, test-taker affect, and performance outcomes. However, we are unaware of research examining how test takers' access to repeated practice tests—now enabled by AIG—relates to these factors. This likely stems from the limited scalability of conventional practice tests, which rely on human test developers who cannot generate test items at the same scale as AIG. He et al. (2024) conducted an extensive literature review, including 66 studies about research for second language test preparation. No themes emerged demonstrating research that examined technology

[1]The practice test items are created using the same AIG methods as the official DET.

[2]Successive versions of OpenAI's GPT models were used to develop the practice test, reflecting generative AI advances.

[3]Also see Flor (2025) for a comprehension discussion of automated item generation.

or AI to enhance test preparation.

3 The Study

This *observational study* examined how access to repeated practice test opportunities—enabled by AIG—related to test takers' official DET performance, test-taker affect, and test-taker decisions to share their official DET scores for university admissions. The study addressed the research question: What are the *observed* relationships between the number of practice tests taken and test-takers' official DET performance, test-taker affect, and test-taker score sharing decisions?

3.1 Methods

3.1.1 Survey instrument

To measure test takers' affect, we developed a brief survey instrument (henceforth, *survey*) that elicited perceptions of *achievement, confidence, motivation, preparedness*, and *anxiety* in relation to the official DET. The survey items reflect affective factors commonly used in prior research on assessment (e.g., Winke and Lim, 2017) and instructional contexts (e.g., Ling et al., 2021). We acknowledge that typical affective surveys include more items per construct. However, because the DET is an operational, high-stakes assessment, there are required constraints: we had to limit the number of post-test, *offboarding*[4] questions to avoid overburdening test takers. Consequently, the survey consisted of five items, each rated on a six-point Likert-style scale. The survey was presented to all test takers as shown in **Figure 1**.

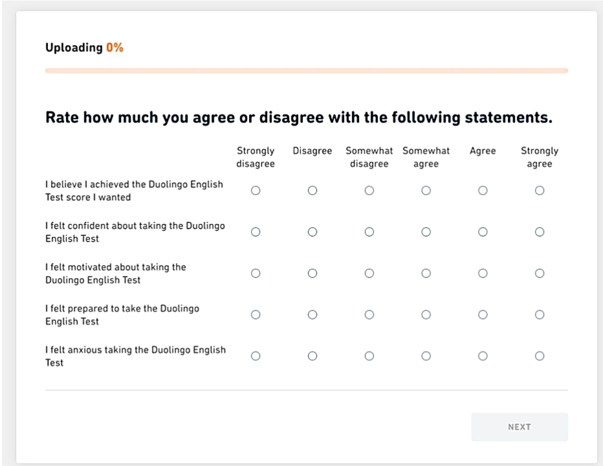


Figure 1: Post–DET Affective Perceptions Survey

	ACH	CON	MOT	PREP	ANX
ACH	1.00	0.74	0.59	0.67	0.04
CON	0.74	1.00	0.68	0.72	-0.03
MOT	0.59	0.68	1.00	0.64	0.07
PREP	0.67	0.72	0.64	1.00	0.06
ANX	0.04	-0.03	0.07	0.06	1.00

Table 1: Spearman Correlations Between Responses to Survey Items; ACH=Achieved; CON=Confident; MOT=Motivated; PREP=Prepared; ANX=Anxious

3.1.2 Data Collection

The survey was administered during September 2023. Upon completion of the DET, test takers were presented with the survey during the DET offboarding process.

Of the original 32,599 test-taker participants (henceforth, test takers) who took the survey, responses were retained from 25,969 test-takers for the analysis. Responses were retained only for participants who: (1) responded to all survey items; (2) were taking the official DET for the first time[5]; (3) received an official DET score that was validated by human proctors; and, (4) had taken the practice tests within 60 days prior to taking the official DET.

Table 1 shows the Spearman rank-order correlations between the survey items. The pairwise correlations between *I believed I achieved the DET score I wanted* (Achieved), *I felt confident about taking the DET* (Confident), *I felt motivated about taking the DET* (Motivated), and *I felt prepared to take the DET* (Prepared) are moderately high. This suggests that these positive affective statements may be related to a similar construct. By contrast, *I felt anxious taking the DET* (Anxious) is effectively uncorrelated with the other items.

3.1.3 Participant Demographics

Test taker demographic information is collected from test takers during the official DET's offboarding process. Offboarding items ask test takers about their *gender*, *age*, *testing intent* (i.e., obtaining an undergraduate or graduate degree), and *first language*.[6] **Table 2** shows the self-reported, test-taker demographics, also comparing the participant

[4]Offboarding takes place once the test is completed. Test takers are asked questions related to, e.g., demographics and their target score.

[5]Prior testing may have provided additional practice, complicating the analysis.

[6]One hundred unique languages were reported by at least five participants.

Demographic	TTs (%)	DET(%)
Gender		
Female	44.0	47.6
Male	55.9	52.3
Age Group		
16–20 years	19.0	32.7
21–25 years	36.6	34.1
26–30 years	18.8	14.8
Testing Intent		
Undergraduate	43.0	47.1
Graduate	43.7	37.0
First Language		
English	13.7	9.5
Mandarin	10.8	17.8
Telugu	10.3	5.8
Spanish	8.8	10.0
Arabic	5.9	5.1

Table 2: Test-Taker Demographics; TTs=Test takers from this study; DET=DET population

sample to the DET test-taker population (Naismith et al., 2025). The sample includes all demographic subgroups from the DET population, though with some variation in proportions. This may be because the study included only first-time test takers, while the DET test-taker population includes both first-time and repeat test takers.

3.2 Analyses

This section discusses relationships that emerged between test takers' DET practice test engagement (i.e, *number of practice tests taken*), and their official DET scores, their affect (as self-reported in the survey), and their score-sharing decisions.[7]

Table 3 shows official DET scores by number of practice tests taken. Test takers were grouped into six bins (*count groups)* by number of practice tests completed (0, 1, 2–3, 4–6, 7+). We chose these categories to distinguish between 0, 1, and multiple practice test-taking sessions. Multiple practice test counts were grouped to balance the bin sample sizes.

Table 3 suggests a relationship between practice tests taken and official DET scores. For each practice test count group, we included 95% confidence intervals of the mean test score. The highest average scores were observed among those who took 1–3 practice tests (in **bold rows**). Confidence intervals of the mean test score for these rows did not overlap with those for 0, or 4 or more practice tests, showing significant differences. Those who took 0, or 4 or more practice tests scored slightly lower.[8]

# of PT	N	%	M	CI 95%
0	4,742	18.3	108.5	[107.8, 109.2]
1	**6,128**	**23.6**	**112.4**	**[111.8, 113.0]**
2–3	**6,469**	**24.9**	**112.3**	**[111.8, 112.8]**
4–6	4,142	16.0	111.1	[110.5, 111.7]
7+	4,488	17.3	108.6	[108.1, 109.1]
Total	25,969			

Table 3: Mean (M) Overall DET Score by Number of Practice Tests Taken (# of PT)

The finding that scores do not continue to increase with 4 or more practice tests aligns with expectations: practice tests are intended to build test familiarity, which on its own, should not facilitate large jumps in language proficiency.

Table 4 illustrates the relationship between number of practice tests taken, test-taker affect, and test takers' official DET score. As no clear differences emerged across the original Likert-scale categories (Figure 1), the six Likert-scale categories were collapsed into two. *Agree* contained: Strongly Agree, Agree, and Somewhat Agree. and *Disagree* contained: Strongly Disagree, Disagree, and Somewhat Disagree.

We included 95% confidence intervals of the difference between the mean scores for those who Agree and Disagree.[9] Rows in **bold** indicate that the confidence interval did not include 0, showing significant differences. Table 4 consistently shows that among test takers who took 0–3 practice tests, those who Agreed with positively-oriented items (Achieved, Confident, Motivated, Prepared) performed significantly better on the official DET than those who Disagreed. For those who Agreed they were Motivated and Prepared, better performance was also observed for 7+, and 4-6 and 7+ groupings, respectively.

Test takers who took 0 or 1 practice test showed a significant score difference between those who

[7]We used test takers' unique, official DET IDs to link to their practice test activity and score report sharing.

[8]Average scores across all groups hovered around the B2 CEFR level—a benchmark for independent language users and a common minimum for admission to English-medium universities (Council of Europe, 2020). However, it is important to note that where the test taker sits in the B2 CEFR range (lower vs. higher in the range) can impact their acceptance to a university.

[9]The Disagree mean score was subtracted from the Agree mean score.

Agreed and Disagreed across all positive statements. As well, test takers who practiced 2-3 times also showed significant differences between those who Agreed and Disagreed with the positive statements. This finding suggests that for some test takers, access to repeated test practice was related to positive affect and higher test scores.

Across the large proportion of test takers who indicated they felt Anxious (70.8%-75.3%), there was no signficant relationship found based on the number of practice tests taken. A possible explanation is the high-stakes nature of the DET. In recent work in classroom settings, Deho et al. (2025) found relationships between test anxiety and demographic factors. This is something that could be explored in future research.

Table 5 indicates a relationship between number of practice tests taken, likelihood of score sharing for university admissions, and test-taker affect.

We used 95% confidence intervals for the share rates (proportions) of those who Agreed or Disagreed with each of the statements. Rows in **bold** indicate that the corresponding Agree and Disagree confidence intervals did not overlap, which showed significant differences. Test takers who took 0, 1, or 2-3 practice tests and Agreed with the Achieved, Confident, and Prepared statements had non-overlapping confidence intervals with test takers who took 0, 1, or 2-3 practice tests and Disagreed with those statements. For those who Agreed with the Motivated statement, only those who took 2-3 practice tests had share rate confidence intervals that did not overlap with the corresponding confidence intervals with those who Disagreed. Note that test takers were always more likely to share their scores if they Agreed with positive statements.

As expected, further analysis showed that test takers who shared their scores tended to have higher mean scores. Scores typically aligned with a mid- to high B2 CEFR level. This is an expected outcome, as test takers are more likely to share scores that meet university requirements. Scores were highest among those who took 0–3 practice tests and Agreed with positive sentiment statements. For example, those who Agreed with the Achieved category had mean scores of 119.1, 120.9, and 119.0 for 0, 1, and 2–3 tests taken, respectively. This trend held across all positive sentiment categories. Scores declined slightly for those who took 4–6 tests (about 1 point lower) and more noticeably for those with 7+ tests (about 3 points lower). A similar pattern emerged for the Anxious category.

4 Discussion

Integrated into the DET pipeline, AIG generates large item pools. This scales the creation of DET practice tests, which increases test takers' access to repeated practice opportunities. To our knowledge, this is the first study to examine how AIG can contribute to increased practice opportunities and how, in turn, access to more practice is related to test-taker affect and outcomes. The study explored relationships between (1) practice test engagement and test score. (Table 3), (2) test-taker affect and official DET scores (Table 4), and (3) affect and score-sharing decisions for university admissions (Table 5).

Three key findings emerged from the analysis to address our research question: What are the *observed* relationships between the number of practice tests taken, and official DET performance, test-taker affect, and score-report sharing decisions?

First, repeated test practice was related to higher test scores to an extent. (Table 3). Those who took 1, or 2-3 practice tests had comparatively higher scores than those who took 0, or more than 3. As taking 2-3 practice tests was related to higher test scores, this suggests a potential benefit of access to repeated practice for some test takers. These test takers may have come to the practice test with higher proficiency and were using the practice test for its intended purpose—i.e., test familiarity.

By contrast, taking more than 2-3 practice tests was associated with lower performance. This may be related to washback effect (mentioned earlier). Specifically, test takers may have used the practice test for reasons beyond test familiarity, such as building English language skills (i.e., positive washback that supports language learning), or test-taking strategies, such as trying to *game* the test (i.e., negative washback that does not support language learning) (Knoch et al., 2020; Xie, 2013). In this scenario, test takers' repeated practice testing may be an example of *wheel spinning*, where learners repeated attempts to master a skill are unsuccessful (Beck and Gong, 2013; Mu et al., 2020).

Second, test takers who took more practice tests reported feeling more positively (Table 4). Based on the number of practice tests taken, higher proportions of test takers reported positive affect toward the official DET regarding their beliefs that

#	Agree		Disagree		CI 95%
	%	M	%	M	
Achieved					
0	**85.7**	**109.5**	**14.3**	**102.2**	**[5.1, 9.5]**
1	**82.8**	**113.8**	**17.2**	**105.6**	**[6.6, 9.9]**
2-3	**84.3**	**112.8**	**15.7**	**109.1**	**[2.3, 5.3]**
4-6	87.4	111.3	12.6	109.7	[-0.3, 3.5]
7+	91.0	108.6	9.0	108.3	[-1.6, 2.2]
Confident					
0	**85.4**	**109.9**	**14.6**	**100.5**	**[7.2, 11.6]**
1	**82.8**	**114.0**	**17.2**	**104.6**	**[7.8, 11.1]**
2-3	**84.4**	**113.2**	**15.6**	**107.4**	**[4.2, 7.2]**
4-6	86.6	111.3	13.4	109.6	[-0.2, 3.6]
7+	91.4	108.7	8.6	108.3	[-1.6, 2.4]
Motivated					
0	**90.9**	**109.1**	**9.1**	**102.1**	**[4.1, 9.9]**
1	**89.8**	**113.0**	**10.2**	**107.5**	**[3.2, 7.7]**
2-3	**91.7**	**112.6**	**8.3**	**108.8**	**[1.7, 5.8]**
4-6	93.0	111.2	7.0	110.1	[-1.6, 3.7]
7+	**95.2**	**108.8**	**4.8**	**105.8**	**[0.4, 5.5]**
Prepared					
0	**85.3**	**110.0**	**14.7**	**99.9**	**[7.8, 12.2]**
1	**82.5**	**114.3**	**17.5**	**103.6**	**[9.1, 12.3]**
2-3	**85.4**	**113.3**	**14.6**	**106.2**	**[5.5, 8.6]**
4-6	**88.3**	**111.6**	**11.7**	**107.3**	**[2.3, 6.2]**
7+	**92.7**	**108.9**	**7.3**	**105.4**	**[1.4, 5.5]**
Anxious					
0	70.8	107.6	29.2	110.6	[-4.6, -1.6]
1	72.9	112.2	27.1	113.1	[-2.2, 0.4]
2-3	73.9	112.3	26.1	112.1	[-0.9, 1.4]
4-6	75.3	111.2	24.7	110.6	[-0.7, 1.9]
7+	75.0	108.5	25.0	108.9	[-1.5, 0.8]

Table 4: Mean (M) Overall DET Score by Practice Tests Taken (#) and Affective Perceptions

#	Agree		Disagree	
	%	CI 95%	%	CI 95%
Achieved				
0	**41.9**	**[40.3, 43.4]**	**32.2**	**[28.6, 35.7]**
1	**43.5**	**[42.1, 44.8]**	**31.9**	**[29.1, 34.7]**
2-3	**43.4**	**[42.1, 44.7]**	**33.6**	**[30.7, 36.5]**
4-6	42.2	[40.6, 43.8]	38.0	[33.9, 42.2]
7+	44.3	[42.8, 45.8]	40.2	[35.5, 45.0]
Confident				
0	**42.2**	**[40.7, 43.7]**	**30.3**	**[26.8, 33.7]**
1	**43.6**	**[42.2, 44.9]**	**31.4**	**[28.6, 34.2]**
2-3	**43.5**	**[42.2, 44.8]**	**32.8**	**[29.9, 35.7]**
4-6	42.1	[40.5, 43.7]	38.8	[34.7, 42.8]
7+	44.0	[42.5, 45.5]	43.3	[38.4, 48.2]
Motivated				
0	41.0	[39.5, 42.4]	35.5	[31.0, 40.0]
1	41.9	[40.6, 43.2]	37.9	[34.1, 41.8]
2-3	**42.5**	**[41.2, 43.8]**	**34.7**	**[30.7, 38.7]**
4-6	41.7	[40.1, 43.2]	41.2	[35.6, 46.9]
7+	44.2	[42.7, 45.6]	39.6	[33.1, 46.1]
Prepared				
0	**42.1**	**[40.6, 43.6]**	**31.1**	**[27.7, 34.5]**
1	**43.5**	**[42.1, 44.9]**	**32.0**	**[29.2, 34.8]**
2-3	**43.5**	**[42.2, 44.8]**	**32.3**	**[29.3, 35.3]**
4-6	42.1	[40.5, 43.7]	38.0	[33.7, 42.3]
7+	44.3	[42.8, 45.8]	38.9	[33.6, 44.2]
Anxious				
0	39.7	[38.0, 41.3]	42.4	[39.8, 45.0]
1	41.1	[39.6, 42.5]	42.6	[40.3, 45.0]
2-3	42.0	[40.6, 43.4]	41.4	[39.1, 43.8]
4-6	41.4	[39.6, 43.1]	42.5	[39.5, 45.6]
7+	43.2	[41.5, 44.8]	46.3	[43.4, 49.2]

Table 5: Proportion of Test Takers who Shared Their DET Score by Number of Practice Tests Taken and Affective Perceptions

they achieved the score they wanted, and their confidence, motivation, and preparedness. As such, the 7+ group consistently had the highest proportion of test takers reporting positive affect. Reported feelings of anxiety were similar across the number of practice tests taken (Table 4). While not surprising in a high-stakes context, the finding is novel compared to prior work suggesting that test preparation could reduce anxiety (Chang and Read, 2008; Powers and Alderman, 1983; Winke and Lim, 2017). However, previous work was conducted in no- or low-stakes experimental settings.

Regarding DET performance, test takers who agreed with the positive statements had higher official DET scores, on average, than those who disagreed; this finding was significant (Table 4). Those who took 1-3 practice tests had the highest scores, on average. Test scores trended lower after taking more than 3 practice tests.

Third, test takers who reported positive perceptions were more likely to share their official DET score report for university admissions (Table 5). This finding was consistent across the number of practice tests taken with comparatively higher proportions for those who Agreed than Disagreed with the positive survey items. Share rates were significantly higher for those who took 0-3 practice tests and Agreed with the Achieved, Confident and Prepared statements, and for those who took 2-3 practice tests and Agreed with the Motivation statement, as compared to those who Disagreed. Like other outcomes we investigated, Anxiety did not show significant differences in share rates by agreement status.

5 Limitations

This section notes *two* study limitations.

First, as an observational study, our findings are **not causal**. Independent of practice test use, higher English proficiency may underlie positive perceptions, higher scores, and share rates.

Second, the number of survey items was necessarily limited to reduce the burden test takers after taking a high-stakes test. Given this real-world constraint, we prioritized items related to test-taker affect, and did not include an item eliciting information about alternative test strategies. As a result, we lacked data on test takers' use of alternative preparation methods. Related, we do not have information about what motivated test takers' repeated practice. As we continue with this research, we are exploring ways to address this limitation.

6 Conclusions

The DET's practice test simulates the official DET. As a computer-adaptive test, the practice test aims to familiarize test takers with the official DET's item types, its adaptive administration, and the official DET score scale (by providing an estimated test score range). Integrating AIG into the test development pipeline enables scalable production of DET practice tests. This facilitates the creation of multiple practice test versions, offering test takers repeated opportunities to build test familiarity.

The study analysis showed that test takers who took 1-3 practice tests tended to have higher official DET scores. Higher test scores were also related to positive affect (i.e., agreeing with the positive survey items). Higher share rates were also linked to positive affect. This *may* be related to those test takers having higher underlying English proficiency. Therefore, test takers may have used the practice test for its intended purpose—test familiarization, whereby 1-3 practice test repetitions may have been sufficient. This also suggests that for some test takers—those who took 2-3 practice tests— that *limited* repeated practice may have provided extra needed support to sufficiently build their test familiarity.

By contrast, test takers who took more than 3 practice tests had lower performance, on average. It is possible that these test takers may have come to the test with lower proficiency. Their additional test practice may be related to washback, whereby test takers used the practice test for reasons besides building test familiarity (e.g., English language learning or building test-taking strategies). However, we lack data about test takers' preparation strategies, beyond the DET practice test, as well as test-taker goals for taking the practice test. Therefore, this limits interpretation. At the same time, it raises interesting questions with regard to appropriate guidance about test preparation, especially with regard to mitigating negative washback effects, such as using test practice to develop test gaming strategies.

AIG for high-stakes assessment is still in its early stages. The study examines how repeated practice—enabled by AIG—may relate to test-taker performance, affect, and behavioral outcomes (i.e., score sharing). It also raises important questions about test preparation practices when test-takers have access to repeated test practice. Our findings—and future research—could be useful in helping to inform best practices for AI-enhanced test readiness in high-stakes contexts.

Acknowledgements

The authors would like to thank our Duolingo colleagues for helpful reviews of earlier versions of this paper: Audrey Kittredge, Nitin Madnani, Ben Naismith, and Alina von Davier. Thanks to the anonymous reviewers for their feedback.

References

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:1–13.

J. E. Beck and Y. Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013*, pages 431–440. Springer.

A. C. S. Chang and J. Read. 2008. Reducing listening test anxiety through various forms of listening support. *TESL-EJ*, 12(1). N1.

Jacqueline Church, Yena Park, and Jill Burstein. 2025. Guidelines for fair test content: The Duolingo English Test example. Duolingo Research Report DRR-25-02, Duolingo. 19 pages.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing.

Oscar Blessed Deho, Srecko Joksimovic, Maria Vieira, and Ryan Baker. 2025. Beyond predictive accuracy: Fairness and bias in predicting test anxiety. In *Proceedings of the International Conference on Artificial Intelligence and Education*.

Michael Flor. 2025. Question generation with large language models and generative ai. In *Automatic Question Generation*, pages 137–147. Springer.

Anthony Green. 2007. Washback to learning outcomes: A comparative study of ielts preparation and university pre-sessional language courses. *Assessment in Education*, 14(1):75–97.

Shanshan He, Anne-Marie Sénécal, Laura Stansfield, and Ruslan Suvorov. 2024. A scoping review of research on second language test preparation. *Language Testing*, 42(1):11–47.

Michael Heilman and Noah A. Smith. 2010. Question generation via overgenerating transformations and ranking. Technical Report CMU-LTI-10-008, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA.

Ute Knoch, Annemiek Huisman, Cathie Elder, Xiaoxiao Kong, and Angela McKenna. 2020. Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing*, 37(4):550–572.

Kadidja Koné, Paula Winke, and Matthew Gordon. 2024. “we would like to see ourselves in the test:” the experiences of francophone african english learners in high-stakes english proficiency testing.

Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 414–428.

Guangming Ling, Norbert Elliot, Jill C Burstein, Daniel F McCaffrey, Charles A MacArthur, and Steven Holtzman. 2021. Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, 48:100509.

O. L. Liu. 2014. Investigating the relationship between test preparation and toefl ibt performance. *ETS Research Report Series*, (2):1–13.

Nitin Madnani, Jill Burstein, John Sabatini, Kristy Biggers, and Slava Andreyev. 2016. Language muse™: Automated linguistic activity generation for english language learners. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

S. Messick. 1982. Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, 17:67–91.

Samuel Messick. 1996. Validity and washback in language testing. *Language testing*, 13(3):241–256.

Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194.

Tong Mu, Andrea Jetten, and Emma Brunskill. 2020. Towards suggesting actionable interventions for wheel-spinning students. *International Educational Data Mining Society*.

B. Naismith, R. Cardwell, G. LaFlair, S. Nydick, and M. Kostromitina. 2025. Duolingo English Test: Technical manual. Duolingo research report, Duolingo.

B. O'Sullivan, K. Dunn, and V. Berry. 2021. Test preparation: An international comparison of test takers' preferences. *Assessment in Education: Principles, Policy & Practice*, 28(1):13–36.

D. E. Powers. 1985. Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement*, 9(2):179–190.

D. E. Powers and D. L. Alderman. 1983. Effects of test familiarization on sat performance. *Journal of Educational Measurement*, 20(1):71–79.

P. Winke and H. Lim. 2017. The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, 14(4):380–397.

Q. Xie. 2013. Does test preparation work? implications for score validity. *Language Assessment Quarterly*, 10(2):196–218.

Develop a Generic Essay Scorer for Practice Writing Tests of Statewide Assessments

Yi Gui
University of Iowa
yi-gui@uiowa.edu

Abstract

This study examines whether NLP transfer learning techniques, specifically BERT, can be used to develop prompt-generic AES models for practice writing tests. Findings reveal that fine-tuned DistilBERT, without further pre-training, achieves high agreement (QWK ≈ 0.89), enabling scalable, robust AES models in statewide K-12 assessments without costly supplementary pre-training.

1 Introduction

Currently, Automated Essay Scoring (AES) is widely utilized in large-scale standardized tests with writing assessments in the US. However, there are some notable limitations in the current major AES engines that are used for many high-stakes writing assessments, such as the annual statewide assessments in K-12 education. These limitations prevent the provision of instantaneous online essay scoring services in writing practice tests of those statewide assessments for students' daily exercise.

One major limitation of AES algorithms trained with traditional machine learning (ML) approaches is the substantial sample size required for training sets with essays scored by human raters. The random assignment of prompts in practice tests results in some prompts having too few essay samples to effectively train a scoring model using traditional ML methods. For instance, Intelligent Essay Assessor (IEA), a major AES engine developed by Pearson which is used in many operational tests, including several statewide assessments, requires a sample of approximately 500 student responses evaluated by human raters to score essays on a specific prompt in high-stakes assessments (Foltz et al., 2013). While it also scores essays in MyLab Writing online services instantly with immediate overall evaluations, it still needs hundreds of submissions scored by human raters to build scoring models for each prompt (Pearson Inc., 2010).

A precursor area with this frequent lack of "labelled" data quandary in ML is the image classification problem through computer vision. The traditional ML model needs to be trained for a specific task of image classification with the target data from scratch, making no use of the knowledge previously learned from similar tasks. To deal with this predicament, transfer learning is applied because it is able to build accurate models even without enough labeled data from the target domain (Rawat & Wang, 2017). With transfer learning, the model-building process starts from the "knowledge" that has been learned previously instead of zero, when solving relevant problems in the past.

Thus, the purpose of the study is to develop a generic essay scorer generalizable to essays on any prompts in the target domain with Google's BERT (Bidirectional Encoder Representations from Transformers), one of state-of-the-art NLP transfer learning techniques, for low-stakes online writing practice tests of those statewide student assessments, even if there is not enough essay sample to train scoring algorithms with traditional ML approaches. With such a generic essay scorer, students' routine practice essays can be scored similarly to those assessment essays even outside the annual test windows, providing students with timely and meaningful feedback during their preparation.

Transfer learning using Google's BERT revolutionizes traditional ML approaches by leveraging pre-trained models on extensive datasets to improve performance on specific downstream tasks. BERT is pre-trained on a large

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 58–81
October 27-29, 2025

corpus of human language text materials, including the entirety of Wikipedia (comprising roughly 2.5 billion words) and the BookCorpus dataset (comprising approximately 800 million words). This pre-training method is particularly advantageous as it allows BERT to generate deep contextualized word embeddings that capture nuanced relationships within the text and be fine-tuned with minimal labeled target data to develop high-performing models in target domains. Thus, this study seeks to investigate how BERT can be utilized to help develop generic AES models and examine how different treatments of BERT's pre-training affect the models' scoring performances in an AES research experiment designed to answer these research questions. Moreover, an analytic essay scoring method focusing on specific writing traits has been selected in this research. The four traits to be scored are development, organization, language use, and prompt task, based on the ELA Common Score Standards of writing, and the scoring rubrics of the SWAS essays used as the target data in the study.

In this research, the following research questions are expected to be answered:

1) How many hyperparameter settings of the original BERT model, when fine-tuned on target data, achieve a Quadratic Weighted Kappa (QWK) value greater than 0.7 for each writing trait (development, organization, prompt task, language use) without additional pre-training?

2) How many hyperparameter settings result in QWK values greater than 0.7 when a pre-trained BERT model undergoes further pre-training on either "within-task" or "in-domain" materials, followed by fine-tuning? Additionally, do these settings outperform the original BERT model in terms of performance?

3) What is the performance rank orders of fine-tuned scoring models for various writing traits when using the same hyperparameter settings, and what are the implications?

The target domain consists of essays written by high school students, while the scoring results produced by the AES engine, IEA, for the available SWAS essays in the study serve as the reference against which the study's scoring results are compared. The flowchart in Figure 1 illustrates the research design and the experimental procedures of the study.

2 Related Work

Automated Essay Scoring (AES) systems have historically depended on handcrafted linguistic features coupled with traditional machine-learning methods. Early influential systems like Project Essay Grade (PEG) used simple textual proxies—such as sentence length or vocabulary—to approximate human grades (Page, 1966). Later, more sophisticated AES engines, notably IntelliMetric and E-rater, employed extensive feature engineering, including grammar accuracy, lexical diversity, and structural coherence (Attali & Burstein, 2006; Shermis & Burstein, 2013). These approaches established AES as a viable alternative for essay scoring, yet their accuracy and adaptability heavily depended on the quality and quantity of manually crafted features and extensive prompt-specific training data.

The release of the Automated Student Assessment Prize (ASAP) dataset (Shermis & Burstein, 2013) significantly advanced AES research by offering a standardized evaluation benchmark. With this dataset, neural network methods emerged, notably recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which automatically learned textual representations rather than relying solely on manual features. Taghipour and Ng (2016) demonstrated that simple CNN-RNN hybrids could surpass traditional AES baselines by directly learning meaningful text patterns from essays. Still, these early neural models struggled to effectively represent complex, long-range discourse structures characteristic of persuasive and argumentative essays.

The advent of pretrained transformer-based language models, particularly BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), dramatically shifted the AES paradigm. These models, pretrained on massive textual corpora, offered deep contextualized embeddings capable of capturing semantic and syntactic nuances beyond the reach of simpler neural architectures (Devlin et al., 2019). Mayfield and Black (2020) provided an influential early evaluation of fine-tuning BERT for AES, showing that transformer models could achieve accuracy comparable to highly-engineered feature-based systems, although computational demands were notably higher. Their work demonstrated transformers' potential for AES, while also highlighting practical trade-offs in model deployment.

To better exploit transformers' strengths, researchers developed specialized fine-tuning methods. Yang et al. (2019) proposed combining a traditional regression loss with a ranking loss, guiding transformer models toward learning not only accurate score predictions but also correct relative ordering of essay quality. This dual-objective approach improved Quadratic Weighted Kappa (QWK)—a standard AES performance metric—by approximately 2–3 percentage points over standard fine-tuning, demonstrating that carefully crafted training objectives can significantly enhance transformer-based AES.

AES research has also addressed the perennial challenge of data scarcity through domain adaptation and multi-task learning. Typically, each essay prompt has limited training data, posing significant risks of overfitting. Cao et al. (2020) presented a domain-adaptive framework combining adversarial training and auxiliary self-supervised tasks (e.g., sentence-order prediction) to learn prompt-invariant essay representations. Their approach not only improved performance on previously unseen prompts but also established a practical methodology for mitigating prompt-specific data shortages through domain transfer. Similarly, Muangkammuen and Fukumoto (2020) employed multi-task learning by integrating an auxiliary sentence-level sentiment analysis task alongside AES. This hierarchical joint training improved QWK scores, illustrating that complementary learning tasks could enrich the representation learned by AES models, enhancing their generalizability.

Holistic essay scoring, while common, limits the detailed feedback educators desire. Thus, recent AES research emphasizes analytic scoring, separately evaluating distinct writing traits (e.g., organization, content, grammar). Historically, separate models were developed independently for each trait, ignoring the natural correlations among writing dimensions. For example, early analytic scoring models, like those by Persing and Ng (2015, 2016), modeled traits like argument strength or organization independently with trait-specific features and classifiers. More recently, Do et al. (2024) proposed Autoregressive Score Generation for Multi-trait Scoring (ArTS), using a transformer-based T5 model to sequentially generate scores for multiple traits. This innovative framework explicitly modeled trait dependencies, significantly improving trait-level AES performance and marking a notable advancement in providing nuanced formative feedback to students.

Evaluation methods have also become standardized with AES advancements. Quadratic Weighted Kappa (QWK) remains a widely adopted metric, penalizing larger scoring errors more heavily and thus closely aligning automated evaluations with human judgments. Current transformer-based AES models routinely achieve QWK scores around 0.75 to 0.80 on standard benchmarks like ASAP, nearing human inter-rater agreement levels (~0.80–0.85; Mayfield & Black, 2020; Yang et al., 2019). This demonstrates substantial progress in AES technology toward human-level reliability.

Overall, AES research has evolved significantly—from feature-engineered regressors to sophisticated transformer-based methods—driven by transformer architectures, specialized training strategies, multi-task learning, and domain adaptation. These advances collectively address critical challenges such as data scarcity and trait-specific feedback, facilitating robust, reliable, and informative automated scoring systems. This literature provides a robust foundation for the current study's exploration of developing prompt-generic AES models for statewide educational assessments, emphasizing transformer-based methods' potential to improve scoring quality, reduce data requirements, and enhance educational feedback.

3 Method

A distilled version of BERT (DistilBERT) was employed to develop prompt-generic essay scoring models. Three variants were compared:

Group 1 (Baseline): DistilBERT fine-tuned directly on SWAS essays.

Group 2 (ASAP-pretrained): DistilBERT further pre-trained on the ASAP corpus, then fine-tuned on SWAS.

Group 3 (SWAS-pretrained): DistilBERT further pre-trained on a 500-essay "within-task" SWAS subset, then fine-tuned on SWAS.

3.1 Data Preparation

Two corpora were used. The SWAS corpus originally contained 4,500 essays (1,500 per grade for grades 9–11). Handwritten submissions (n = 1,203) were excluded, leaving 3,297 typed essays (Figure 2). A random sample of 500 typed essays was reserved for within-task pre-training. The ASAP corpus, comprising 12,970 essays across eight prompts and two genres (Table 1), was used for in-domain pre-training.

To mitigate score-level imbalance from handwritten-essay removal, RandomOverSampler was applied separately to each analytic trait. The balance improvements were confirmed via stacked-bar plots and annotated tables (Figures 3 and 4), though downstream benefits were minimal. Oversampled sets were used only for diagnostics.

3.2 Model Pre-training and Fine-tuning

DistilBERT weights (66 M parameters) were loaded from the Hugging Face "distilbert-base-uncased" checkpoint. In Groups 2 and 3, intermediate pre-training was performed using a learning rate of 5×10^{-4} and batch sizes of 16 and 32. All pre-training ran for a uniform number of epochs, ensuring each variant saw equal exposure to its respective corpora.

Subsequently, each variant was fine-tuned on SWAS essays using an Ordinal Logistic Regression (OLR) classifier built on DistilBERT embeddings. Hyper-parameters for fine-tuning were selected via grid search over three regularization strengths ($\alpha \in \{0.01, 0.10, 1.00\}$) and three maximum-iteration ceilings ({100, 500, 1000}), yielding nine distinct configurations.

3.3 Evaluation Protocol

Model evaluation employed a leave-one-grade-out design: in three rounds, essays from two grades were used for training and the remaining grade served as the test set (Tables 2–4). Within each round, five-fold cross-validation was executed, and the entire process was repeated with three random seeds to assess stability. Aggregate statistics across folds and seeds were computed for: **Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Exact Accuracy, Adjacent Accuracy (predictions within ±1 score point), Precision, recall, and F1** were calculated per score point (1–5) (see Figures 8–10 for accuracy, Figures 11–12 for precision, recall, and F1).

By systematically comparing baseline and pre-trained variants under consistent optimization settings and a robust leave-one-grade-out protocol, this method section demonstrates how prompt-generic essay scoring can be realized with minimal reliance on prompt-specific labeled data. The design ensures fairness across groups, repeatability via multiple seeds, and comprehensive trait-level analysis through detailed metric computation and visualization.

4 Results

4.1 Agreement and Accuracy Across Splits

Table 2–4 report mean Quadratic Weighted Kappa (QWK) results for each leave-one-grade-out split. When trained on grades 9 & 10 and tested on grade 11 (Table 2), mean QWK ranged from 0.889 to 0.893 across the best hyper-parameter settings. Similar stability was observed for the other splits: training on grades 9 & 11 (Table 3) yielded QWK near 0.892, and training on grades 10 & 11 (Table 4) yielded QWK near 0.893. Exact accuracy, summarized in Tables 5–7, consistently hovered around 0.68–0.69 for all splits. Aggregating across splits (Table 8) confirms mean QWK ≈ 0.89 and mean accuracy ≈ 0.68, demonstrating that two-grade training provides robust linguistic coverage for scoring the held-out grade.

4.2 Impact of Pre-training

Supplementary pre-training did not yield a uniform advantage; effects depended on split, α, and trait. At $\alpha = 1.0$, with the strongest regularization, the no-pretraining baseline (Group I) achieved the highest QWK across all traits in the train 9&11 → test 10 design (Table 3). In other splits, leadership shifted: for train 9&10 → test 11 (Table 2), Group II (ASAP-pretrained) led Organization, Prompt Task, and Development, while Group III (SWAS-pretrained) led Language Use; for train 10&11 → test 9 (Table 4), Group II dominated most traits, with all groups performing similarly on Prompt Task. At lower α, leadership occasionally changed by trait but without clear consistency. Overall, even at $\alpha = 1.0$, where performance was most stable, relative rankings fluctuated across splits, showing that train–test design substantially shaped outcomes and prevented conclusive judgments of model performance.

4.3 Trait-Level Performance

Figures 5–7 plot QWK trajectories across max_iter for each trait in the three splits. Organization consistently scored highest (peak QWK ≈ 0.93), followed by Language Use and Development (≈ 0.90), with Prompt Task trailing (≈ 0.86). Even the most challenging trait, Prompt Task, exceeded the operational QWK threshold of 0.70 in every configuration (Figures 5–7). These rankings held irrespective of pre-training group, confirming a stable hierarchy of trait difficulty. Macro-average F1 scores per trait across splits are summarized in Table 9.

4.4 Precision, Recall, and F1 by Score Point

Figures 11–12 show per-score precision, recall, and F1 for the 9+10→11 split (and supplementary figures for the other splits). All groups peak at the extreme scores (1 & 5) and dip in the mid-range (2–4) for precision and recall, reflecting both data imbalance and inherent scoring difficulty. No group gains a systematic edge from extra pre-training.

4.5 Hyper-parameter Fine-tuning

Hyper-parameter sweeps confirm that regularization strength $\alpha = 1.0$ combined with at least 500 training iterations produces the most stable and highest-performing models. Early stopping at 100 iterations dropped QWK by roughly 0.005–0.006 (see Tables 2–4), and increasing beyond 1,000 iterations yielded diminishing returns. Lower α values (0.01, 0.10) led to mild over-fitting, indicated by higher training QWK but lower test QWK and increased variance across seeds.

4.6 Oversampling Correction

Although RandomOverSampler successfully equalized class frequencies (supplemental bar plots), oversampling did not materially improve modeling outcomes. Precision and recall at rare score points improved slightly in some configurations, but aggregate QWK and accuracy remained unchanged or marginally worse when oversampled sets were used for training.

4.7 Macro-Average F_1 Summary

To condense all per-score results, Table 9 reports the macro-averaged F_1 (mean over score points 1–5) for each trait, group, and leave-one-grade-out split. Together, Tables 2–9 and Figures 5–12 show that a baseline DistilBERT (G1) fine-tuned on two-grade SWAS essays yields high agreement (QWK ≈ 0.89), accuracy (≈ 0.68), and F_1 across traits, without the need for extra pre-training or oversampling.

Overall, these results demonstrate that a baseline DistilBERT model—fine-tuned exclusively on two-grade SWAS data—achieves high agreement (QWK ≈ 0.89) and accuracy (≈ 0.68) across grade splits and analytic traits without requiring additional pre-training or extensive oversampling (Tables 2–9, Figures 5–12).

5 Discussion

The stability of model performance across all three leave-one-grade-out splits suggests that DistilBERT's pre-trained language representations are highly adaptable to essay scoring—even without extensive prompt-specific data. Training on any two adjacent grades yielded nearly identical agreement (QWK ≈ 0.89), exact accuracy (~ 0.68), and Adjacent Accuracy (> 98 %), confirming that essays from two grades supply sufficient linguistic and rhetorical variety to generalize to a held-out grade.

Perhaps most surprisingly, neither large-scale in-domain pre-training on ASAP nor "within-task" pre-training on a SWAS subset produced consistent gains. As Table 9's macro-average F_1 summary shows, the baseline model (Group 1) ties or outperforms both ASAP-pretrained (Group 2) and SWAS-pretrained (Group 3) variants in every trait and split. For instance, Prompt Task F_1 on 9+10→11 is 0.714 for Group 1 versus 0.706 (Group 2) and 0.698 (Group 3). This counter-intuitive result implies that when the BERT's original pretraining corpus is already massive and representative enough, further pre-training can introduce stylistic noise or domain drift instead of strengthening task alignment.

Hyper-parameter analysis reinforces the need for careful regularization and adequate training steps. Models with $\alpha = 1.0$ and at least 500 (ideally 1,000) iterations consistently achieve the highest and most reproducible QWK. Lower α values permit mild over-fitting—evident in higher training QWK but lower test QWK—while very short runs (100 iterations) leave a nontrivial 0.005–0.006 QWK gap compared to longer runs.

Trait-level performance reveals a stable hierarchy of difficulty. Organization is most easily predicted (peak QWK ≈ 0.93), followed by Development and Language Use (≈ 0.90), with Prompt Task trailing (≈ 0.86). Crucially, even the most challenging trait exceeds the operational QWK threshold of 0.70, indicating that all four analytic dimensions can be scored with confidence.

Finally, oversampling to correct class imbalance offered minimal benefit. Although frequency distributions were equalized, aggregate QWK, accuracy, and micro-F_1 remained flat or dipped slightly, suggesting that model capacity and the breadth of cross-grade coverage outweigh precise score-level balance when fine-tuning transformer embeddings.

Taken together, these findings validate a lightweight, prompt-agnostic AES pipeline: fine-tune a standard DistilBERT checkpoint on a representative two-grade corpus with α = 1.0 and 500–1,000 iterations, and skip costly intermediate pre-training or complex oversampling. This approach simplifies system development, reduces computational overhead, and still delivers robust, reproducible scoring across multiple writing traits and grade levels.

6 Conclusion

This study has demonstrated that prompt-generic automated essay scoring (AES) can be achieved efficiently by fine-tuning DistilBERT on representative two-grade essay sets, without the need for extensive prompt-specific pre-training or elaborate data balancing. Across three leave-one-grade-out splits and nine hyper-parameter configurations, baseline DistilBERT models consistently achieved strong agreement (QWK ≈ 0.89), exact accuracy (~0.68), and adjacent accuracy (> 98%). These results challenge conventional assumptions, showing that DistilBERT's general-domain representations suffice for robust scoring when paired with straightforward fine-tuning.

A particularly striking finding was the observation of "knowledge collapse": applying supplementary pre-training settings to overwrite existing parameters paradoxically diminished downstream scoring performance. This counter-intuitive effect—where newly acquired "knowledge" impaired rather than enhanced task ability—underscores the critical need to avoid equating machine learning processes with human learning, and suggests that care must be taken to preserve previously learned representations during transfer learning.

From a practical standpoint, clear hyper-parameter guidelines have emerged: a regularization strength of α = 1.0 and a training horizon of 500–1 000 iterations reliably maximize performance and model stability. This simple recipe offers a low-overhead path to deploying AES in educational contexts, minimizing both computational cost and engineering complexity.

Nonetheless, certain limitations temper the generalizability of these conclusions. The within-task pre-training set was limited to 500 essays covering a single prompt per grade, which may have constrained the potential benefits of task-specific pre-training. Exclusion of 1203 handwritten essays—due to transcription challenges—introduced moderate score-level imbalance and restricted the training corpus's representativeness. Finally, employing a single scoring rubric across all prompts may have simplified the generalization challenge.

To address these gaps, future work should explore larger, more diverse essay collections spanning multiple prompts, genres, and rubrics to assess how prompt variety and score distribution affect adaptability. Alternative machine-learning frameworks beyond ordinal logistic regression—such as ensemble methods or neural classifiers—should be evaluated for further performance gains. It will also be important to develop transfer-learning strategies that explicitly guard against "knowledge collapse," preserving core representations while incorporating new domain information. Integrating advanced handwriting recognition technologies remains essential for inclusive AES that covers all response formats.

In closing, this research provides compelling evidence that a lightly fine-tuned DistilBERT model can serve as a scalable, reliable AES engine for formative writing practice, dramatically reducing the data and computational burdens. By recommending concrete hyper-parameter settings and highlighting the nuanced effects of further pre-training, this work lays a pragmatic foundation for the next generation of accessible, robust AES tools in K-12 education.

A Appendices

ASAP Dataset	Topics
Prompt 1	The effects computers have on people
Prompt 2	Censorship in the libraries
Prompt 3	Respond to an extract about how the features of a setting affected a cyclist
Prompt 4	Explain why an extract from *Winter Hibiscus* by Minfong Ho was concluded in the way the author did
Prompt 5	Describe the mood created by the author in an extract from *Narciso Rodriguez* by Narciso Rodriguez
Prompt 6	The difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there
Prompt 7	Write a story about patience
Prompt 8	The benefits of laughter

Table 1: Topics of Eight Prompts in ASAP Dataset

	Fine-tuning Parameter: Alpha is set to be the same across three groups No. of essays for training=2251 No. of essays for test=1046						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1.0	Language Use	0.940	0.887	0.940	0.876	0.943	0.922
	Organization	0.938	0.851	0.944	0.920	0.946	0.908
	Prompt Task	0.939	0.914	0.940	0.922	0.942	0.917
	Development	0.935	0.897	0.938	0.925	0.940	0.903
0.1	Language Use	0.944	0.873	0.941	0.897	0.938	0.902
	Organization	0.945	0.930	0.942	0.928	0.944	0.893
	Prompt Task	0.937	0.895	0.930	0.896	0.940	0.847
	Development	0.938	0.884	0.939	0.923	0.938	0.876
0.01	Language Use	0.938	0.851	0.934	0.895	0.931	0.886
	Organization	0.940	0.915	0.935	0.912	0.936	0.884
	Prompt Task	0.931	0.863	0.930	0.896	0.925	0.881
	Development	0.933	0.867	0.933	0.904	0.931	0.839

Table 2: Mean QWK Results vs. Alpha for Train on Grade 9&10 and Test on Grade 11

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2170 No. of essays for test=1127						
	maxiter=500	Group I (No Further Pre-training)		Group II (Further Pre-training with ASAP Essays)		Group III (Further Pre-training with SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1	Language Use	0.935	0.922	0.938	0.918	0.941	0.871
	Organization	0.939	0.935	0.944	0.879	0.941	0.853
	Prompt Task	0.941	0.921	0.946	0.911	0.948	0.871
	Development	0.934	0.926	0.940	0.903	0.937	0.871
0.1	Language Use	0.939	0.914	0.936	0.909	0.937	0.882
	Organization	0.941	0.932	0.944	0.827	0.941	0.865
	Prompt Task	0.944	0.919	0.945	0.899	0.945	0.879
	Development	0.936	0.923	0.941	0.895	0.936	0.893
0.01	Language Use	0.933	0.890	0.931	0.886	0.928	0.867
	Organization	0.938	0.924	0.938	0.789	0.935	0.855
	Prompt Task	0.940	0.907	0.939	0.887	0.938	0.858
	Development	0.935	0.902	0.936	0.869	0.929	0.877

Table 3: Mean QWK Results vs. Alpha for Train on Grade 9&11 and Test on Grade 10

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2173 No. of essays for test=1124						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1.0	Language Use	0.946	0.911	0.948	0.913	0.949	0.908
	Organization	0.949	0.922	0.948	0.927	0.949	0.925
	Prompt Task	0.936	0.890	0.940	0.891	0.939	0.892
	Development	0.938	0.903	0.941	0.917	0.940	0.888
0.1	Language Use	0.949	0.909	0.947	0.901	0.946	0.912
	Organization	0.949	0.920	0.945	0.914	0.946	0.900
	Prompt Task	0.936	0.824	0.938	0.858	0.933	0.854
	Development	0.941	0.900	0.940	0.902	0.938	0.870
0.01	Language Use	0.945	0.894	0.941	0.888	0.939	0.874
	Organization	0.944	0.910	0.938	0.881	0.936	0.872
	Prompt Task	0.926	0.815	0.928	0.852	0.924	0.840
	Development	0.936	0.872	0.937	0.881	0.927	0.834

Table 4: Mean QWK Results vs. Alpha for Train on Grade 10 & 11 and Test on Grade 9

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2251 No. of essays for test=1046						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1	Language Use	0.793	0.722	0.792	0.575	0.799	0.760
	Organization	0.791	0.760	0.795	0.697	0.795	0.722
	Prompt Task	0.766	0.718	0.769	0.699	0.776	0.718
	Development	0.762	0.722	0.773	0.738	0.775	0.722
0.1	Language Use	0.805	0.680	0.798	0.667	0.783	0.702
	Organization	0.793	0.702	0.789	0.741	0.789	0.680
	Prompt Task	0.761	0.682	0.761	0.701	0.764	0.682
	Development	0.774	0.675	0.776	0.746	0.771	0.675
0.01	Language Use	0.786	0.658	0.773	0.681	0.755	0.663
	Organization	0.779	0.663	0.763	0.707	0.765	0.658
	Prompt Task	0.744	0.627	0.731	0.659	0.725	0.627
	Development	0.762	0.633	0.765	0.705	0.758	0.633

Table 5: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 9&10 and Test on Grade 11

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2170 No. of essays for test=1127						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1.0	Language Use	0.781	0.762	0.789	0.752	0.799	0.633
	Organization	0.793	0.760	0.805	0.627	0.795	0.595
	Prompt Task	0.781	0.704	0.796	0.684	0.803	0.603
	Development	0.769	0.736	0.788	0.697	0.778	0.643
0.1	Language Use	0.792	0.738	0.782	0.725	0.784	0.672
	Organization	0.794	0.755	0.802	0.547	0.791	0.621
	Prompt Task	0.790	0.702	0.794	0.659	0.791	0.630
	Development	0.774	0.732	0.794	0.683	0.781	0.684
0.01	Language Use	0.773	0.662	0.767	0.670	0.754	0.650
	Organization	0.785	0.732	0.786	0.517	0.775	0.606
	Prompt Task	0.776	0.674	0.771	0.637	0.791	0.630
	Development	0.783	0.675	0.788	0.640	0.778	0.659

Table 6: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 9&11 and Test on Grade 10

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2173 No. of essays for test=1124						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1.0	Language Use	0.806	0.734	0.808	0.740	0.813	0.728
	Organization	0.800	0.739	0.794	0.759	0.801	0.750
	Prompt Task	0.760	0.698	0.768	0.673	0.768	0.691
	Development	0.767	0.678	0.776	0.740	0.773	0.631
0.1	Language Use	0.814	0.726	0.804	0.722	0.801	0.681
	Organization	0.797	0.736	0.783	0.735	0.791	0.699
	Prompt Task	0.758	0.576	0.760	0.612	0.751	0.627
	Development	0.776	0.687	0.770	0.725	0.765	0.609
0.01	Language Use	0.797	0.677	0.785	0.686	0.775	0.611
	Organization	0.797	0.736	0.756	0.669	0.760	0.588
	Prompt Task	0.729	0.570	0.730	0.618	0.718	0.594
	Development	0.767	0.625	0.761	0.701	0.739	0.563

Table 7: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 10&11 and Test on Grade 9

Train → Test	Mean QWK	Mean Accuracy
G9 & G10 → G11	0.893	0.687
G10 & G11 → G9	0.889	0.679
G9 & G11 → G10	0.892	0.673

Table 8: Average Performance by Leave-One-Grade-Out Split

Trait	Model Group	9+10 → 11	9+11 → 10	10+11 → 9
Prompt Task	Baseline (G1)	0.714	0.714	0.645
	ASAP-pretrained (G2)	0.706	0.702	0.630
	SWAS-pretrained (G3)	0.698	0.710	0.626
Organization	Baseline (G1)	0.753	0.741	0.648
	ASAP-pretrained (G2)	0.742	0.725	0.622
	SWAS-pretrained (G3)	0.741	0.730	0.642
Development	Baseline (G1)	0.714	0.722	0.705
	ASAP-pretrained (G2)	0.704	0.710	0.695
	SWAS-pretrained (G3)	0.698	0.716	0.686
Language Use	Baseline (G1)	0.767	0.773	0.762
	ASAP-pretrained (G2)	0.758	0.764	0.752
	SWAS-pretrained (G3)	0.753	0.760	0.740

Table 9: Macro-Average F_1 by Trait, Model Group, and Leave-One-Grade-Out Split

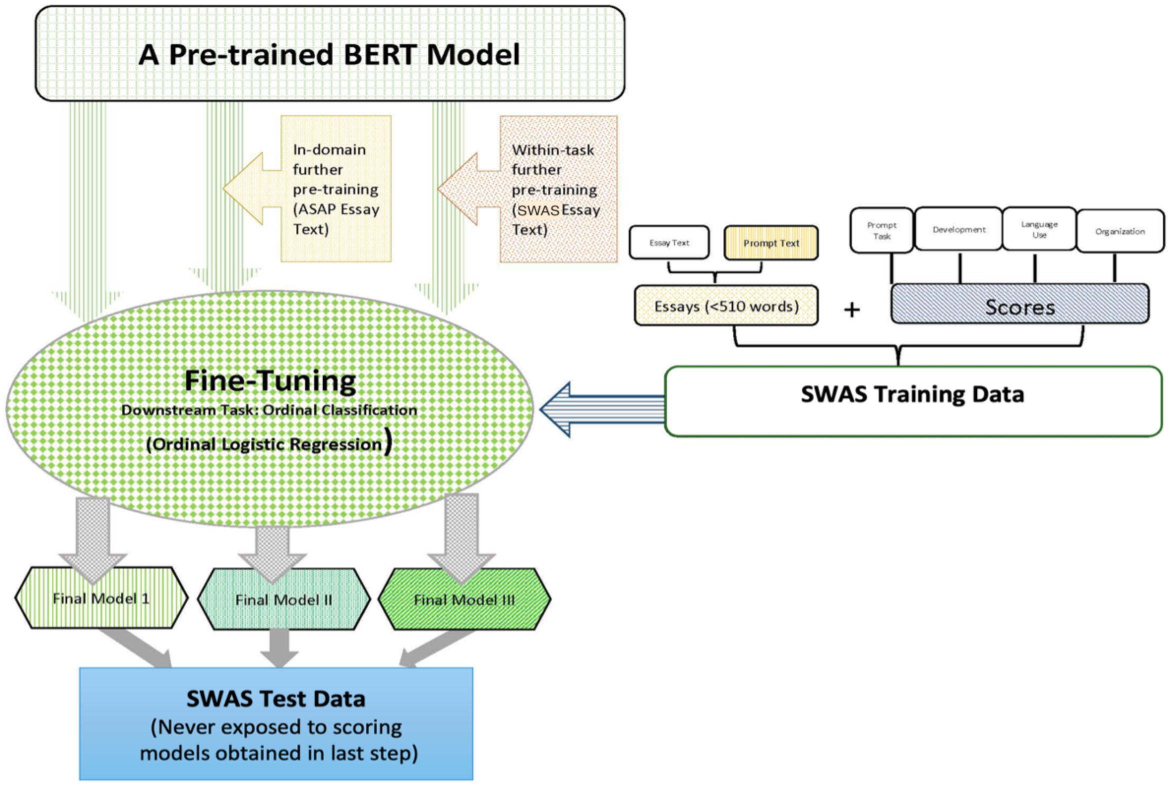


Figure 1: Research Design of the Study

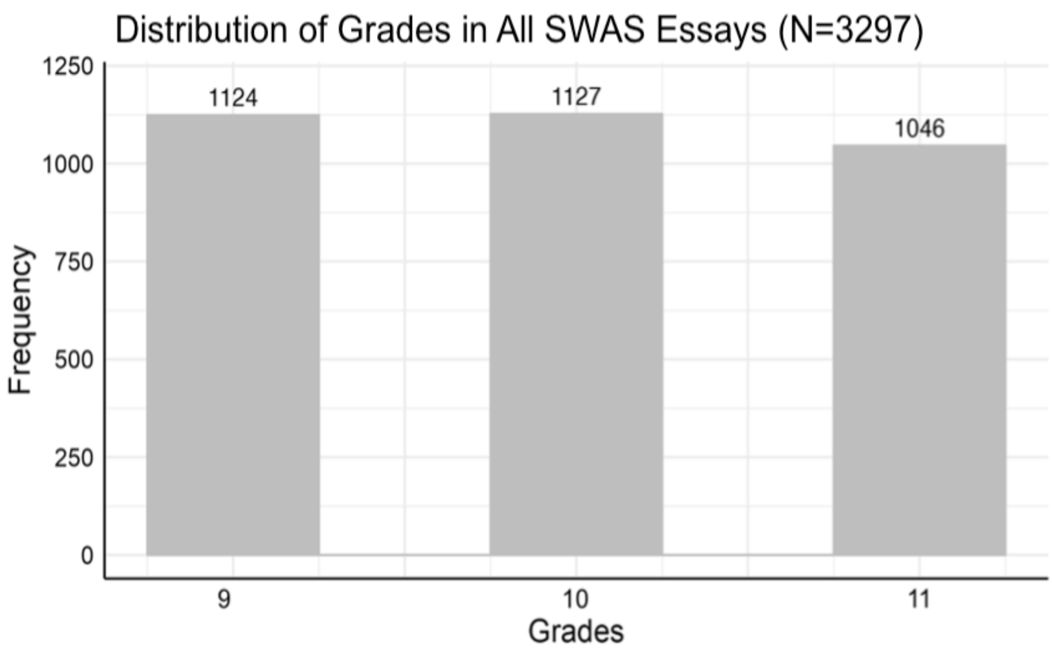


Figure 2: Grade Level Distribution of Available SWAS Essays

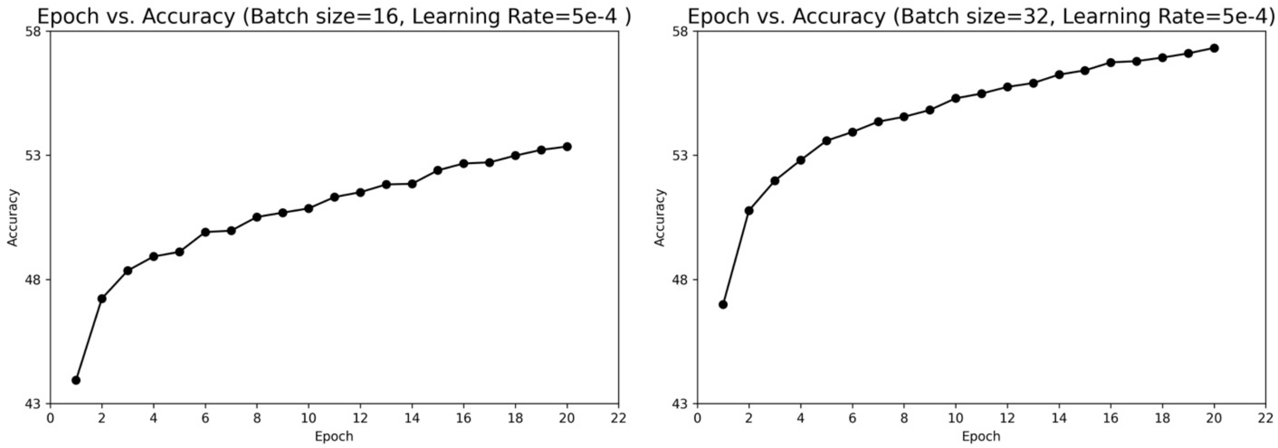


Figure 3: Accuracy in Further Pre-training with ASAP Essays

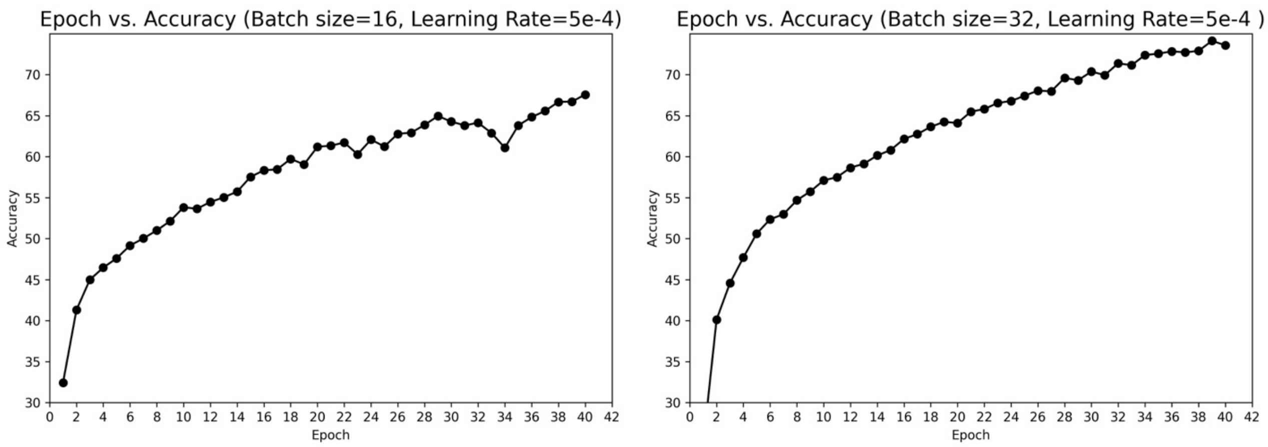


Figure 4: Accuracy vs. Epoch in Further Pre-training with 500 SWAS Essays

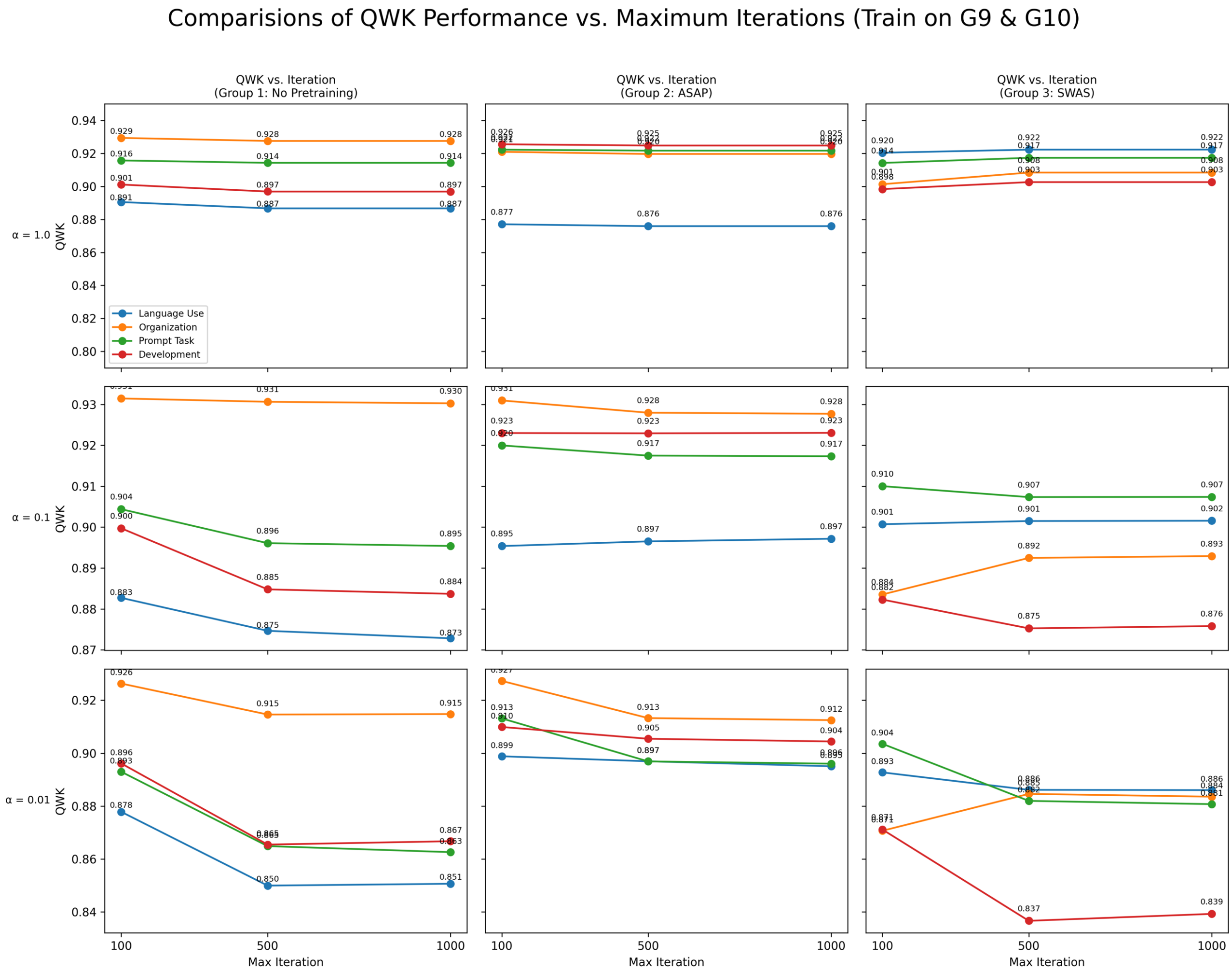


Figure 5: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 9&10 and Test on Grade 11

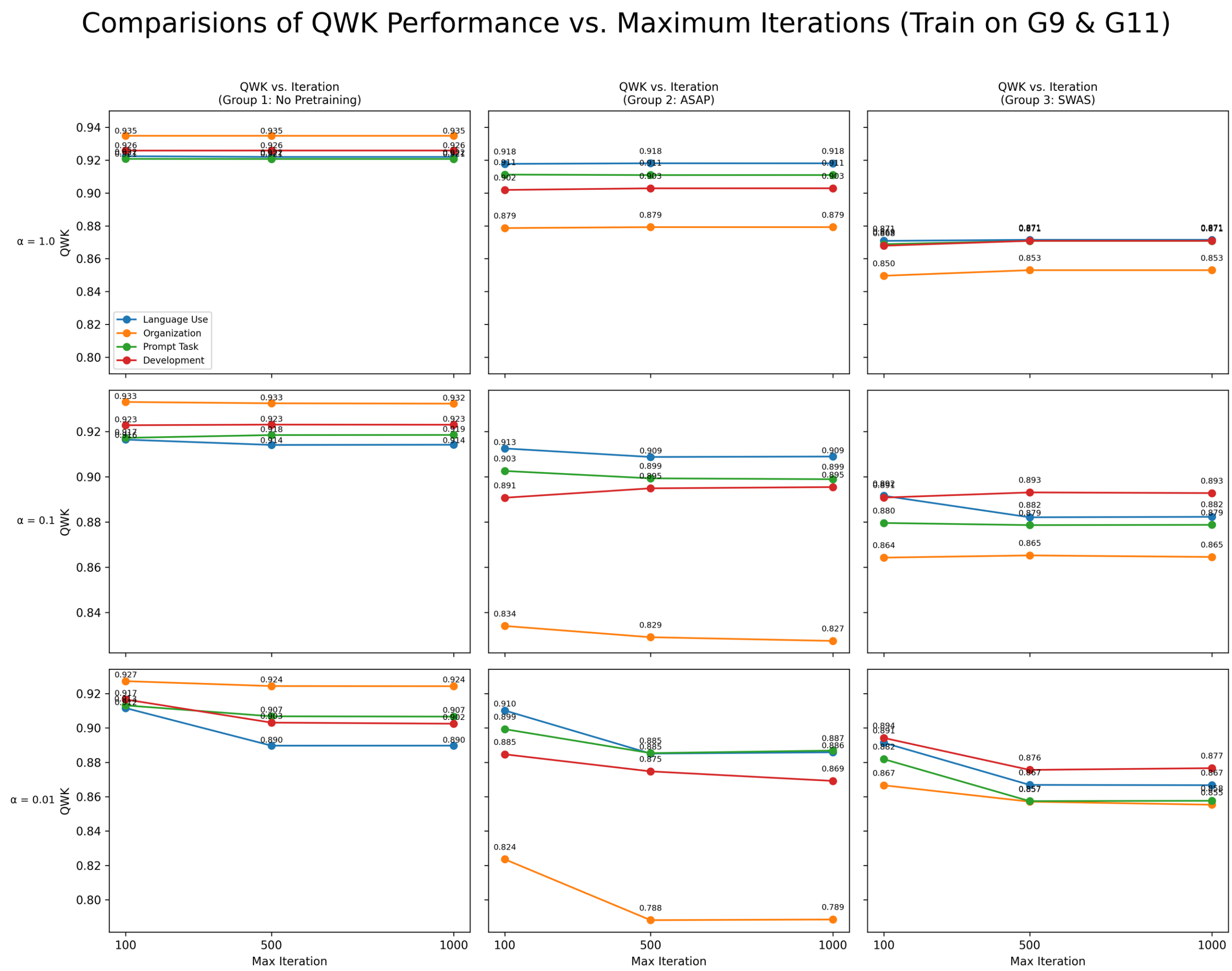


Figure 6: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 9&11 and Test on Grade 10

Figure 7: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 10&11 and Test on Grade 9

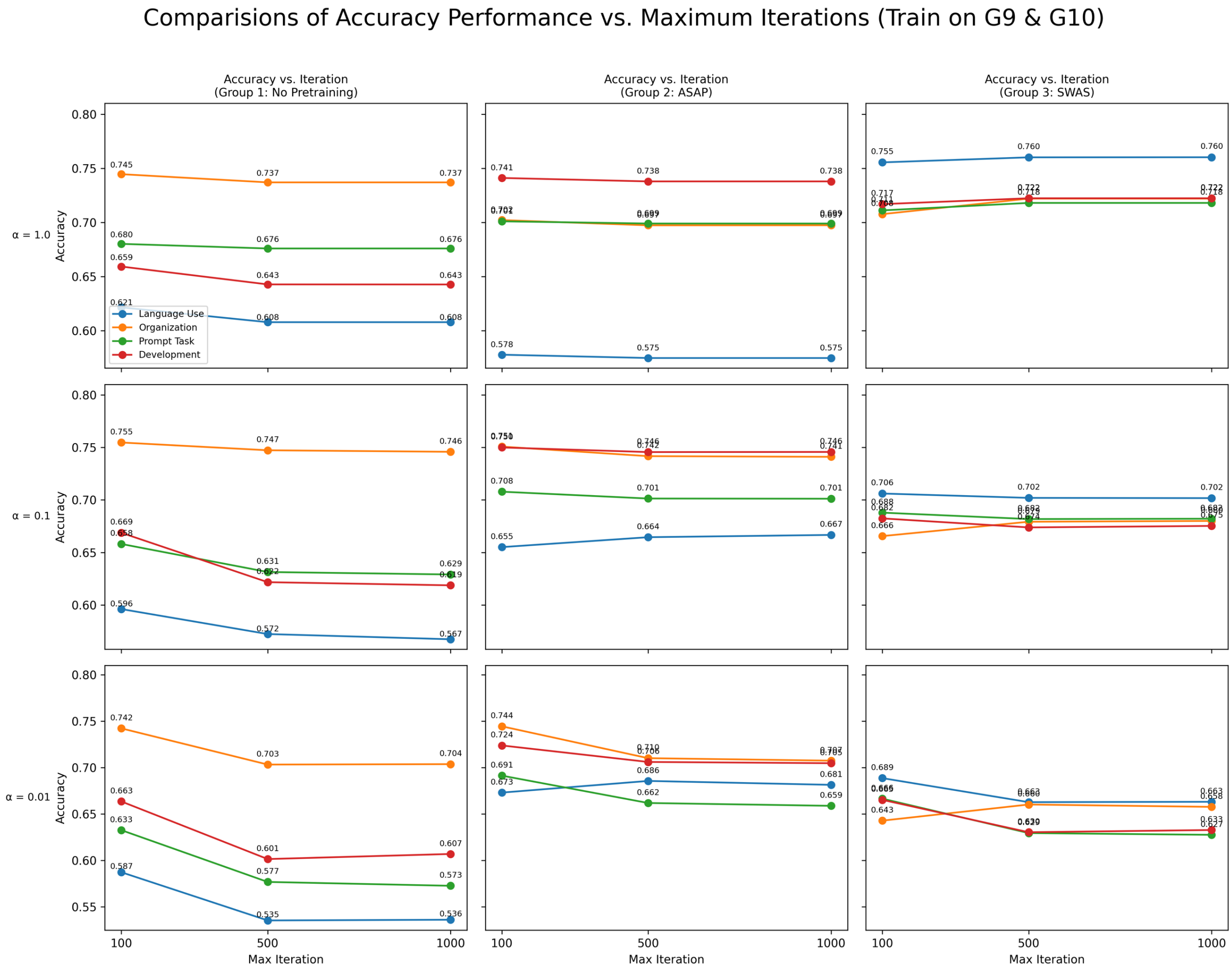


Figure 8: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 9&10 and Test on Grade 11

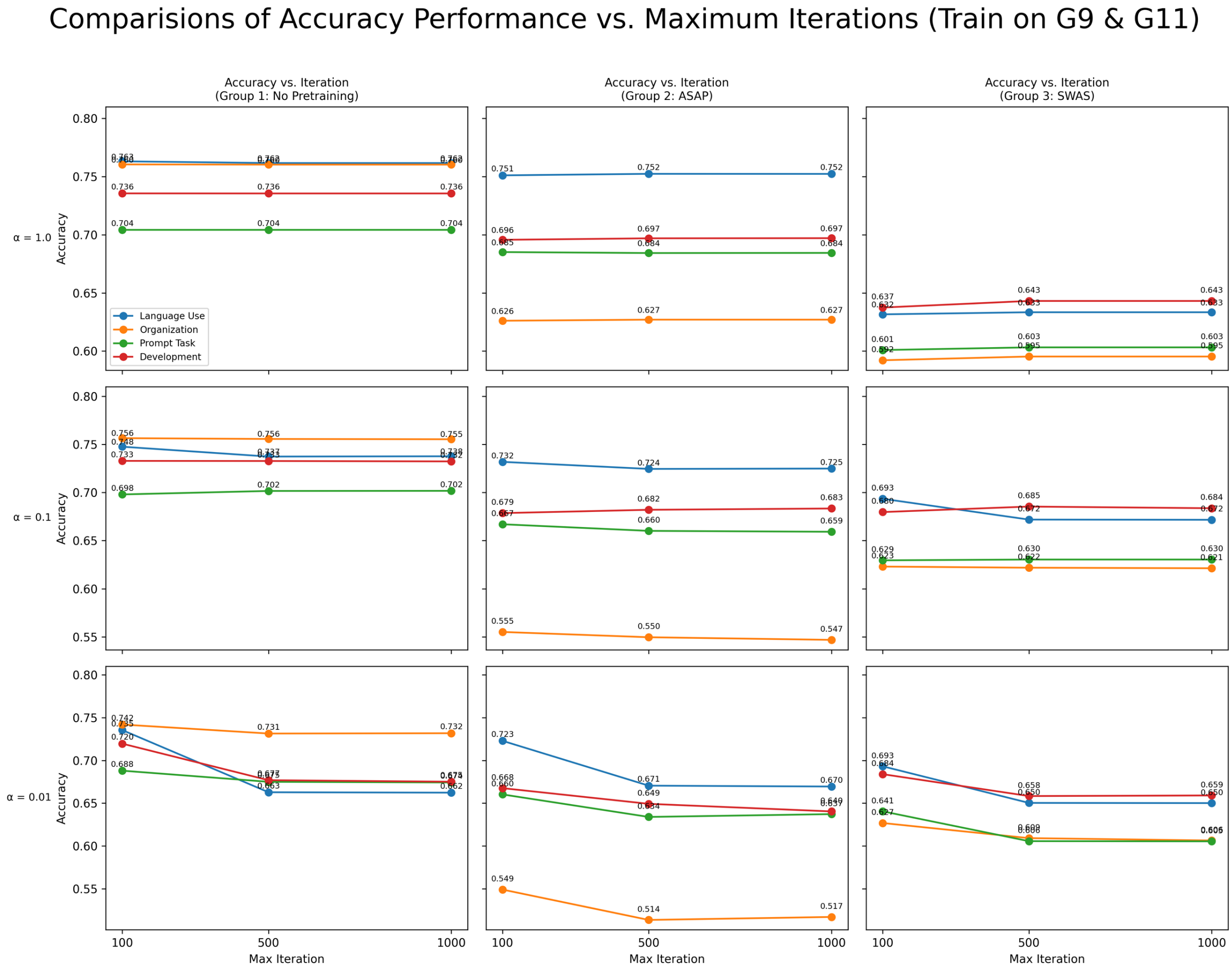


Figure 9: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 9&11 and Test on Grade 10

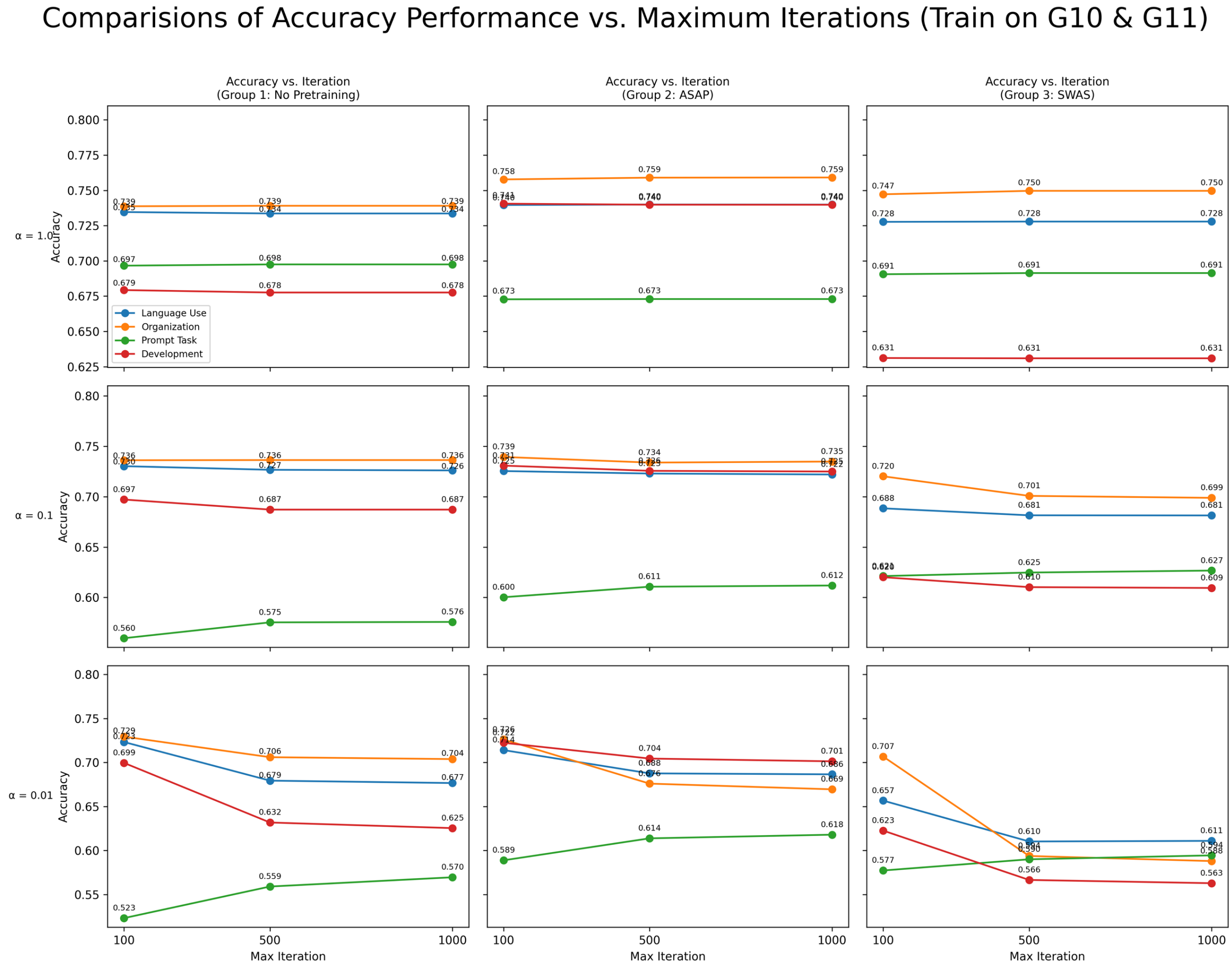


Figure 10: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 10&11 and Test on Grade 9

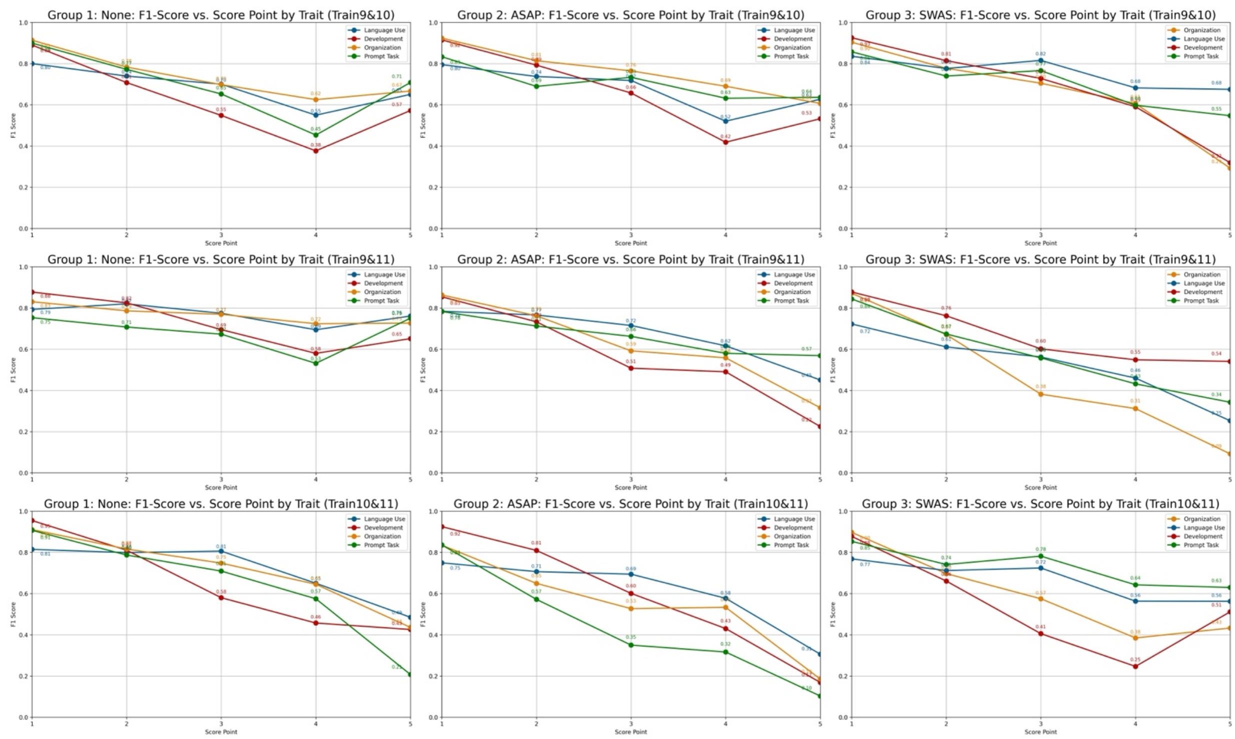


Figure 11: Mean F1-scores in All Score Levels of the 3 Groups across Trait Scores

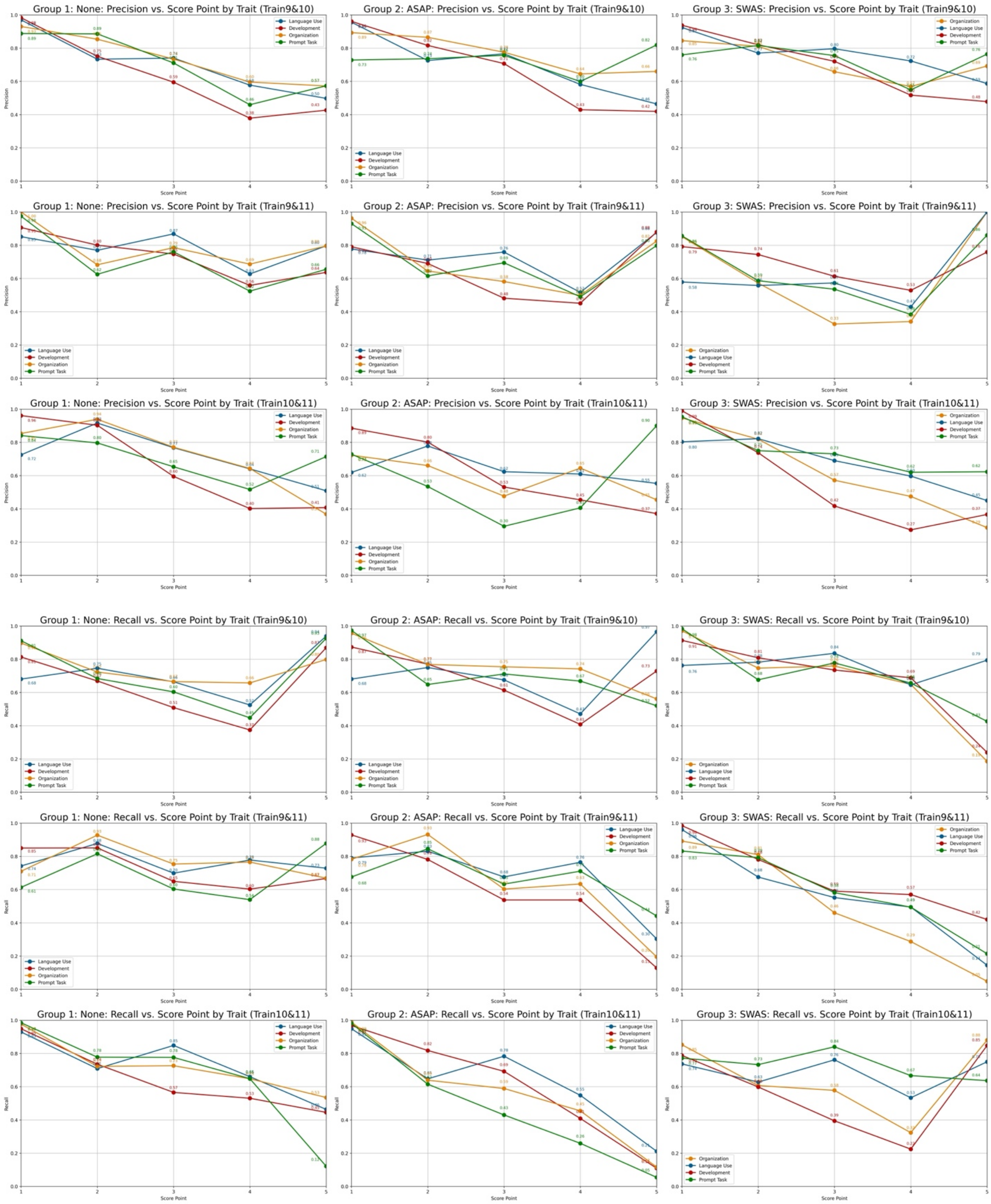


Figure 12: Mean Precision and Recall in All Score Levels of the 3 Groups across Trait Scores

References

Adepoju, A., and K. Adeleke. 2010. Ordinal logistic regression model: An application to pregnancy outcomes. *Journal of Mathematics and Statistics* 6:279–285.

Ajafabadi, M. M. N., F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2:1–21.

Allwright, S. 2022. What is a good F1 score and how do I interpret it? *stephenallwright.com*. Available at https://stephenallwright.com/good-f1-score/

American Educational Research Association, American Psychological Association, and NCME. 2014. Standards for educational and psychological testing. *American Educational Research Association*.

Anyoha, R. 2017. The history of artificial intelligence. Available at https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Attali, Y., and J. Burstein. 2006. Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment* 4(3).

Ayari, R. 2020. NLP: Word embedding techniques demystified—Bag-of-Words vs TF-IDF vs Word2Vec vs Doc2Vec vs Doc2VecC. *towardsdatascience.com*. Available at https://towardsdatascience.com/nlp-embedding-techniques-51b7e6ec9f92

Baccianella, S., A. Esuli, and F. Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications*, 283–287. IEEE.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3(Feb):1137–1155.

Buhl, N. 2023. F1 score in machine learning. *encord.com*. Available at https://encord.com/blog/f1-score-in-machine-learning/

Calvo, M. R. 2018. Dissecting BERT Part 1: The encoder. *medium.com*. Available at https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3

Cao, Y., H. Jin, X. Wan, and Z. Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1011–1020.

Center for Applied Linguistics. 2018. Annual technical report for ACCESS for ELLs 2.0 English language proficiency test, series 401 online, 2016–2017 administration (WIDA consortium annual technical report no. 13B). Washington, DC.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4):213–220.

Collobert, R., and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, vol. 1, 160–167.

Conneau, A., and D. Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1699–1704. European Language Resources Association.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Delua, J. 2021. Supervised versus unsupervised learning: What's the difference? Available at https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Devlin, J., and M.-W. Chang. 2018. Open sourcing BERT: State-of-the-art pre-training for natural language processing. *Google AI Language*. Available at https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019, Volume 1: Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dhami, D. 2020. Understanding BERT — word embeddings. *medium.com*. Available at https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca

Diederich, P. B. 1974. Measuring growth in English. *National Council of Teachers of English*.

Dodge, J., G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. 2020. Fine-tuning pre-trained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Dong, L., N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems* 32.

Donges, N. 2023. Introduction to natural language processing (NLP): The ultimate goal of natural language processing is to help computers understand language as well as we do. *builtin.com*. Available at https://builtin.com/data-science/introduction-nlp

Faigley, L. 1985. Assessing writers' knowledge and processes of composing. *Ablex Publishing Corporation*.

Feathers, T. 2019. Flawed algorithms are grading millions of students' essays. *Vice*. Available at https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays

Fiacco, J., E. Cotos, and C. Rosé. 2019. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 310–319. Association for Computing Machinery.

Fleiss, J., B. Levin, and M. Paik. 2004. *Statistical Methods for Rates and Proportions*, 3rd edn. Wiley-Interscience.

Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3):613–619.

Foltz, P. W., L. A. Streeter, K. E. Lochbaum, and T. K. Landauer. 2013. Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 68–88. Routledge.

Gere, A. R. 1980. Written composition: Toward a theory of evaluation. *College English* 42(1):44–58.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Google Inc. 2020. BERT. *Hugging Face*. Available at https://huggingface.co/docs/transformers/model_doc/bert

Graham, P., and R. Jackson. 1993. The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology* 46(9):1055–1062.

Hamner, B., and M. D. Shermis. 2013. Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 314–346. Routledge.

Hearst, M. A. 2000. The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications* 15(5):22–37.

High, P. 2017. Carnegie Mellon dean of computer science on the future of AI. *Forbes*. Available at https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/?sh=1e8c39b72197

Hinton, G., O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Houlsby, N., A. Giurgiu, S. Jastrzębski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2790–2799. PMLR.

Howard, J., and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Long Papers*, 328–339. Association for Computational Linguistics.

huggingface.co. 2020. Summary of the Tokenizers—transformers 4.3.0 documentation. *huggingface.co*. Available at https://huggingface.co/docs/transformers/tokenizer_summary

Huh, M., P. Agrawal, and A. A. Efros. 2016. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Katyal, N. K. 2003. The promise and precondition of educational autonomy. *Hastings Constitutional Law Quarterly* 31:557–613.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint abs/1609.04836*.

Khanna, C. 2021. WordPiece: Subword-based tokenization algorithm: Understand subword-based tokenization algorithm used by state-of-the-art NLP models — WordPiece. *towardsdatascience.com*. Available at https://towardsdatascience.com/wordpiece-subword-based-tokenization-algorithm-1fbd14394ed7

Komatsuzaki, A. 2019. One epoch is all you need. *arXiv preprint arXiv:1906.06669*.

Kortschak, H. 2020. Attention and transformer models: A complex algorithm, simply explained. *towardsdatascience.com*. Available at https://towardsdatascience.com/attention-and-transformer-models-fe667f958378

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Landauer, T. K., and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Landauer, T. K., D. Laham, and P. Foltz. 2000. The Intelligent Essay Assessor. *Intelligent Systems, IEEE* 15:27–31.

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller. 2002. Efficient backprop. In *Neural Networks: Tricks of the Trade*, 9–50. Springer.

Lee, C., K. Cho, and W. Kang. 2019. Mixout: Effective regularization to finetune large-scale pre-trained language models. In *International Conference on Learning Representations*. Available at https://arxiv.org/abs/1909.11299

Liu, J., Y. Xu, and L. Zhao. 2019. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Mayfield, E., and A. Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162. Association for Computational Linguistics.

Megumi, K.-M. 2003. E-rater software. *Association for Language Teaching*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* 26.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Muangkammuen, P., and F. Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. *AACL*.

Murphy, R. F. 2019. Artificial intelligence applications to support K-12 teachers and teaching. *Rand Corporation*.

Nadeem, F., H. Nguyen, Y. Liu, and M. Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications at ACL 2019*, 484–493.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common Core State Standards for English Language Arts. Washington, DC.

Oquab, M., L. Bottou, I. Laptev, and J. Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724. https://doi.org/10.1109/CVPR.2014.222

Page, E. B. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan* 47(5):238–245.

Pearson Inc. 2010. Reliable automated writing assessment. Available at https://mlm.pearson.com/northamerica/mywritinglab/educators/features/writing-practice/index.html

Pedregosa, F., F. Bach, and A. Gramfort. 2017. On the consistency of ordinal regression methods. *Journal of Machine Learning Research* 18(55):1–35.

Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Persing, I., and V. Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 543–552. Association for Computational Linguistics.

Persing, I., and V. Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1384–1394.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In M. Walker, H. Ji, and A. Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 2227–2237. Association for Computational Linguistics.

Peters, M. E., S. Ruder, and N. A. Smith. 2019. To tune or not to tune? Adapting pre-trained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 7–14. Association for Computational Linguistics.

Phang, J., T. Févry, and S. R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Rawat, W., and Z. Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation* 29(9):2352–2449.

Rennie, J., and N. Srebro. 2005. Loss functions for preference levels: Regression with discrete ordered labels. *IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, Edinburgh, Scotland.

Riley, J. C. 2019. The Massachusetts Board of Elementary and Secondary Education update on automated test scoring. Massachusetts Department of Elementary and Secondary Education. Available at https://www.doe.mass.edu/bese/docs/fy2019/2019-01/spec-item2.html

Rodriguez, P. U., A. Jafari, and C. M. Ormerod. 2019. Language models and automated essay scoring. *International Journal of Assessment Tools in Education* 10(3):149–163.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. 2015. ImageNet large-scale visual recognition challenge. *International Journal of Computer Vision* 115:211–252.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saravia, E. 2018. Deep learning for NLP: An overview of recent trends. *medium.com*. Available at https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d

Scherer, D. L. 1985. Measuring the measurements: A study of evaluation of writing: An annotated bibliography.

Schuster, C. 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement* 64(2):243–253.

Shermis, M., J. Burstein, D. Higgins, and K. Zechner. 2010. Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, 20–26.

Shermis, M. D., and J. Burstein (eds.). 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 1st edn. Routledge.

Shermis, M., H. Mzumara, and J. Olson. 2001. On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education* 26.

Sheshadri, A. K., A. R. Vijjini, and S. Kharbanda. 2021. Wer-BERT: Automatic WER estimation with BERT in a balanced ordinal classification paradigm. *arXiv preprint arXiv:2101.05478*.

Sun, C., X. Qiu, Y. Xu, and X. Huang. 2019. How to fine-tune BERT for text classification? In *Proceedings of Chinese Computational Linguistics: 18th China National Conference*, 194–206. Springer.

Taghipour, K., and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891.

Vanbelle, S. 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika* 81(2):399–410.

Vantage Learning. 2001. A preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments. *Vantage Learning*.

Vantage Learning. 2002. A study of IntelliMetric™ scoring for responses. *Vantage Learning*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems* 30.

Veal, L. R., and S. A. Hudson. 1983. Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English* 17(3):290–296.

White, E. M. 1985. Teaching and assessing writing: Recent advances in understanding and improving student performance. *ERIC*.

Williamson, D., X. Xi, and F. J. Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice* 31(1):2–13.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized autoregressive pre-training for language understanding. *Advances in Neural Information Processing Systems* 32.

Yannakoudakis, H., and R. Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 213–223.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 27.

Young, T., D. Hazarika, S. Poria, and E. Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13(3):55–75.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems* 32.

Towards assessing persistence in reading in young learners using pedagogical agents

Caitlin Tenison[1], Beata Beigman Klebanov[1], Noah Schroeder [2], Shan Zhang[2], Michael Suhan[1], Chuyang Zhang[2]
[1] Educational Testing Service, [2] University of Florida
Correspondence: ctenison@ets.org

Abstract

This pilot study investigated the use of a pedagogical agent to administer a conversational survey to second graders following a digital reading activity, measuring comprehension, persistence, and enjoyment. Analysis of survey responses and behavioral log data provide evidence for recommendations for the design of agent-mediated assessment in early literacy.

1 Introduction

Understanding how young learners respond to reading difficulties is important for supporting early literacy development. This paper presents a preliminary effort to use a pedagogical agent to elicit self-reflections during reading and examine how these reflections align with behavioral patterns during reading captured through process data.

2 Literature Review

2.1 Persistence and Reading

Persistence is generally defined as a student's sustained effort toward academic goals despite difficulty, confusion, or failure (Skinner et al., 2022; Wang et al., 2020). It is a key predictor of learning and long-term achievement, particularly in complex problem-solving tasks (Dweck et al., 2014; Farrington et al., 2012). In K-12 settings, persistence is often supported through instructional strategies such as productive failure (Kapur, 2008), erroneous examples (Richey et al., 2019), and motivational framing (Cook et al., 2019). Persistence is measured in various ways, including behavioral indicators (e.g., time on task, number of attempts) and self-report instruments (e.g., grit scales, mindset surveys) (Shute et al., 2013). For instance, Goh (2025) measured persistence by log-transformed time on task and found it was associated with improved performance.

In the reading context, persistence is a critical factor in reading achievement. A meta-analysis by Toste et al. (2020) reviewed 60 studies of K-12 students and found that motivational factors like self-efficacy, task value, and goal orientation positively relate to reading outcomes, especially when they foster sustained effort and engagement. This effect was even stronger in elementary schools, suggesting motivation is especially important in early literacy. Another study defined persistence in reading as sustained cognitive effort and behavioral engagement in the face of challenges, assessed through self-reports in middle and high school students (Reschly and Christenson, 2022). Maegi et al. (2018) found that effortful control and task persistence, rated by teachers and parents, predicted reading fluency and comprehension in sixth graders. These findings suggest that persistence enables students to stay focused and overcome reading challenges.

This prior research highlights the important role of persistence in reading achievement and outlines diverse ways it has been conceptualized and measured. However, most existing studies focus on upper elementary or older students, leaving a gap in understanding how persistence develops in younger learners. This may stem partly from a reliance on self-report measures to assess persistence, whose validity is limited with younger children who are still developing the metacognitive skills needed for accurate self-reflection (Craig et al., 2020). In response to these challenges, researchers have recommended age-appropriate adaptations to self-reports and combining evidence from surveys and observed behaviors (Gascoine et al., 2017; Desoete, 2008). That said, few studies in the reading context explicitly examine how persistence is being measured using process data (e.g., behavior logs, transcript) compared to traditional methods, such as self-report or teacher rating. This limited use of multimodal approaches limits our understanding of the behavioral and cognitive dimensions of persistence, especially in early literacy development.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 82–90
October 27-29, 2025

The current pilot study begins to address this gap by pairing age-appropriate self-reports with behavioral data in young learner's reading activities.

2.2 Pedagogical Agents

One approach to administer a survey or gain insights into a student's experience is through the use of virtual characters. Frequently called pedagogical agents when used in learning settings (Schroeder et al., 2025; Siegle et al., 2023), virtual characters have been widely used in K-12 settings to help students learn (Zhang et al., 2024a,b). However, pedagogical agents can play a wide variety of roles in a learning environment (Clarebout et al., 2002). Research has shown that it is very common to use pedagogical agents as an information source or for coaching and scaffolding, but is rare to use them to administer self report surveys (Schroeder and Gotch, 2015; Zhang et al., 2024a). While many critical questions must ultimately be addressed to validate agent-administered assessments, a foundational concern is whether students will engage meaningfully with the pedagogical agents in a testing setting. In this proof-of-concept study, we designed and piloted a pedagogical agent to administer a conversational survey to young learners. Given the complexity of surveying students at this age, our goal was to explore the feasibility and potential of integrating agent-mediated assessment within the learning process in an engaging manner.

2.3 Research Questions

We address the following research questions:

- To what extent can a pedagogical agent elicit a broad range of responses (i.e., responses are distributed across response options)?
- How do the behavioral data relate to the data collected via the pedagogical agent survey?

3 Methods

3.1 Participants

Participants were drawn from a second-grade classroom in a charter school in the Northeastern U.S.; the study was reviewed and approved by the Institutional Review Board and all participating students had documented parental consent. The students engaged in a reading activity with the App in six sessions for about 15 minutes each. In each session, the class read one story. The stories were grade-appropriate fictional narratives licensed from Cricket Media, which publishes literary magazines for young readers. After the 3rd reading session during which the students read the story titled "Happy, The Hearing Ear Dog,"[1] the students interacted with Adam, the virtual agent that delivered the survey. The activities took place as part of normal school programming during the first two weeks of April 2025 and were led by the teacher, with technical assistance from a member of the research team during the first reading session and during the survey activity.

Of the 25 students in the class, 20 had parental consent for their reading data to be used for research. Of these, 18 completed the survey. The primary analyses reported below focus on the 18 students who participated in both components. Information about student gender was provided by the school. Our final sample consisted of 7 female and 11 male students. Additional demographic information came from an optional demographic questionnaire completed by the students' parent or guardian. The average age was 7.5 years (SD=0.51). Eleven students' parents reported English as their child's first language, while five indicated English was not their child's first language (2 did not respond). Ten students were identified by parents as Hispanic or Latino and 6 as Black or African American; 2 parents did not respond to this item.

3.2 Instruments

3.2.1 Reading Application

Relay Reader[2] is a reading and listening app developed to support readers as they transition into fluent reading (Madnani et al., 2019). Readers take turns reading stories out loud with a pre-recorded model human narrator (audiobook). The target length of reading and listening turns can be configured by the reader; for this study both were set up at 70 words as default. None of the students changed the default settings. The transition between narrator and student turns occurs on paragraph breaks. The allocation of a passage into narrator or student turn happens dynamically, where paragraphs are added to the turn as long as adding the paragraph made the passage closer to the target length (from above or from below) than not adding it. After every other student turn, before the next narrator turn starts, the student is asked two multiple choice comprehension questions. Questions were created

[1] https://www.audible.com/pd/Happy-the-Hearing-Ear-Dog-Audiobook/B0DJ9SDW68

[2] https://relayreader.org

for approximately every 100 words of running text by researchers and research assistants experienced with developing such items, and reviewed by the senior members of the team as well as by the research institution's fairness review committee. The questions focus on the salient aspects of the plot, settings, and characters, and are surface-level, not requiring inferential reasoning. Thus, the two questions a reader would be presented with would refer to something that was mentioned within 200 words preceding the current bookmark.

As readers interact with the app, the app collects timestamped log data of the various activities (including the focal activities of reading, listening, answering questions, as well as other activities in the app such as looking at the reading history or changing the fonts or other settings). The audio recording of every student turn is processed using an in-house speech analysis system validated for this use case using data predominantly from students in grades 3-5 (Beigman Klebanov and Loukina, 2021; Loukina et al., 2019, 2017). The system produces estimates of reading accuracy and fluency for all scorable recordings. The app has been previously used with students in grades 2-8 in the US in school and summer camp contexts. Depending on the grade and the study goals, the app library – how many and which books each student has access to at any given time – is flexibly managed by the researchers. For this study, one story was put in the library for all participants for each of the six reading sessions. For sessions 1-5, all students got the same story; on day 6, we assigned a new story to students who read all five and assigned one of the stories they missed for students who missed some of the first five reading sessions. All 18 students were present during the 3rd reading session in which they engaged with the pedagogical agent and all read the same story about a dog. Students read on Kindle Fire HD 8 (12th generation) tablets provided by the researchers.

3.3 Reading Data

During the reading activity, the app collected data related to students' reading performance. We focused on four measures for analysis:

1. **RCQ:** Percent correct in the multiple choice reading comprehension questions embedded in the app (see previous section), both for the Dog story specifically and across all the six stories);

2. **SKT:** proportion of skipped turns, defined as the share of reading turns completed faster than the 90th percentile of oral reading fluency norm for the Spring of second grade[3] (148 words per minute) adjusted for text-based variation in fluency estimates using the method in Beigman Klebanov et al. (2019), resulting in the cut rate of 178 words per minute;

3. **ACC:** reading accuracy, calculated as the proportion of words in the passage assigned to the reading in the current turn that were recognized as pronounced correctly by the in-house speech analysis engine (see previous section);

4. **WCPM:** words read correctly per minute, a fluency measure calculated based on the automated recognition and scaled to account for systematic error in the automated measurement (Beigman Klebanov and Loukina, 2021). Following prior research, we restricted analysis of accuracy and WCPM to recordings with accuracy ≥ 0.7 (Liceralde et al., 2022), as lower scores often reflect issues with audio recording rather than students' bona-fide performance.

3.3.1 Pedagogical Agent Design

The pedagogical agent's role was positioned as that of a proof-of-concept conversational assessment. Specifically, the pedagogical agent delivered a series of questions in a conversational style that the learner replied to by answering a multiple-choice question. Questions were presented as on-screen text and simultaneously narrated by the agent. The agent was designed to appear as a teenager, with narration provided by a teenage male speaking English. The system was built using Unity and deployed on the same Kindle Fire tablets that the students used for the reading app. Student responses to the agent were stored locally on the device.

3.3.2 Survey

To assess students' reading attitudes and experience with the reading app, the pedagogical agent administered a brief conversational survey after the 3rd reading session. The survey was designed for second-grade students and emphasized simple language, limited response options, and concrete behavioral prompts. With a total of 14 items, the survey included a mix of comprehension checks,

[3] https://www.readingrockets.org/topics/fluency/articles/fluency-norms-chart-2017-update

attitudinal measures, and behavioral self-reports (see Appendix A for more detail).

We assessed basic reading comprehension and memory of the story using four multiple-choice items based on the story they just read (Happy, The Hearing Ear Dog). These items differed from the in-app RCQ items and asked students what characters the story was about and about a key characteristic of one of the characters. Our reading self-reflection items measured students' perception of reading ease, interest, and learning. These items used simple, developmentally appropriate 3 or 4-point Likert scales (such as 'not really', 'sometimes', 'definitely').

We measured reading persistence, the focal construct of this study, through two items that asked students how they typically respond when confronting challenges when reading, specifically unfamiliar written words or unknown word meanings. Response options reflected specific behaviors associated with persistence or quitting. Students could respond that they used effortful strategies such as 'sounding it out' or 'figuring it out' to indicate persistence or avoidance behaviors such as *skipping* or *stopping* which may suggest lower persistence. We deliberately framed these as concrete, first-person behavior reports (e.g., 'What do you do if you don't know what a word means?') rather than abstract or hypothetical self-assessments typical of measures designed for older students (e.g., "I leave things unfinished"; Chernyshenko et al. 2018; Sparks and Lehman 2025).

To gauge the students' reaction to the reading and agent activities, we asked 3 subjective questions. This included two questions about their preferences related to the stories and a final item which asked the students if they wanted to interact with the agent again.

4 Results

This study reflects an early-stage pilot investigation, and the small sample size (n = 18) limits our use of formal statistical tests or validation procedures.

4.1 Survey

In Figure 1, we present a heat map showing student response patterns across the multiple reading and persistence constructs assessed in our survey. Responses are grouped by construct and scaled such that higher values (and darker colors) reflect greater expression of the underlying construct. Basic reading comprehension and story recall were high, with 72% of students answering all four questions correctly. This is consistent with all of the students reporting that reading aloud was 'OK' or 'easy', with no one saying that it was 'hard' (Item 9); however, 27.7% responded that there were 'lots' of words they did not know how to say (Item 11) or did not know the meaning of (Item 13) on our two follow-up questions about reading ease. While no students reported that they hated reading, 27.8% reported they didn't enjoy reading and 16.7% reported the did not feel they learned much from reading.

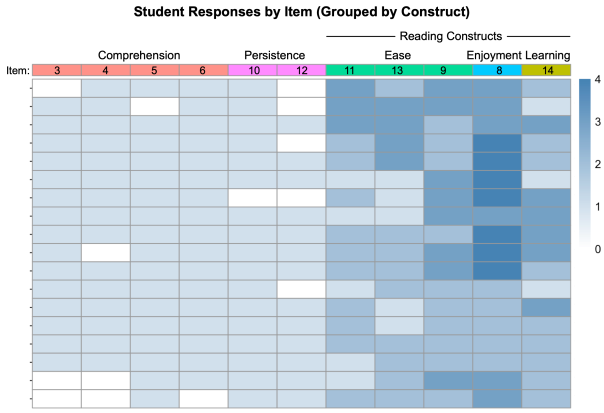


Figure 1: Heatmap of student responses to survey. Rows represent students; columns represent individual items grouped by construct. Darker colors indicate stronger expression of targeted construct. Comprehension and persistence items are coded as correct/incorrect or persistent/not persistent; reading self-reflection constructs use a Likert scale.

In response to the subjective questions about the activities, all students responded that they liked the story and when asked to recommend a story for the agent to read next, 4 students chose to not recommend any of the stories. At the end of the survey, the agent asked students if they wanted to interact again in the future. Only one student chose 'no', with 6 responding 'maybe' and 11 agreeing to a future interaction ('sure').

Student responses on our items measuring persistence showed limited variability. In response to 'What do you do if you don't know how to say a word?' 17 out of 18 students selected 'sounding it out', a behavior we would associate with persistence. Only one student selected 'I skip it and continue' while no students chose 'I ask someone' or 'I stop reading'. Similarly, the second item, 'What do you do if you don't know what a word means?', 13 students selected 'I figure it out' with the remaining 5 responses divided between asking for help or skipping the word. While these responses suggest high levels of self-reported persistence, the strong skew towards a single response option on each item limits our ability to differentiate students' behavior using these measures alone.

4.2 App Use

We first examined the pairwise correlations between students' RCQ scores and behavioral measures extracted from the logs of the app (Figure 2).[4] RCQ accuracy was positively correlated with reading accuracy (ACC; r(13)= .56, p=0.019) and negatively correlated with the proportion of skipped reading turns (SKT; r(16)=-.66, p=.003), but was not significantly associated with fluency (WCPM; r(13)=.2, p=.47).

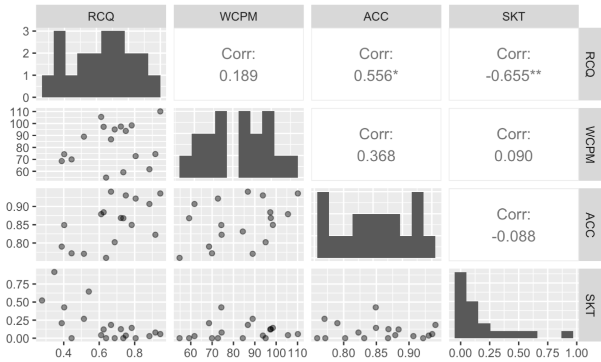


Figure 2: Pairwise correlations between RCQ scores and Relay Reader behavioral measures. Pearson r values shown above the diagonal; p$\leq$.05 (*), p $\leq$.005 (**).

4.3 Relating Survey Response to In-App Behaviors

We examined whether students' self-reported survey responses reflected the behavioral data collected about those same students during their reading activity. First, we tested whether comprehension scores on the survey related to performance on the RCQ questions presented during the reading activity for the same story. Using a Wilcoxon rank-sum test to compare students with perfect vs. non-perfect scores on the survey,[5] we found no significant difference in app RCQ scores (W = 23, p = .36). Next, we assessed whether students' ratings of reading ease related to measured WCPM or ACC. We found no association with fluency (R(16) = .15, p = .6)) and a marginally significant negative correlation with accuracy (R(16) = -.45, p = .089). Finally, we examined whether self-reported reading persistence was related to the proportion of skipped reading turns. A Wilcoxon rank-sum test revealed no significant difference in skipping behaviors between students who responded with 'sound it out' or 'figure it out' to both items versus those who did not (W = 16, p = .11).

5 Discussion

5.1 Relations Between Variables

For the reading data, we found that oral reading accuracy (ACC) and comprehension (RCQ) were positively correlated, while proportion of skipped reading turns and comprehension (RCQs) were negatively correlated. These relationships make sense: Reading what is on the page is necessary to answer questions based on the story, as the questions were designed so that one cannot consistently answer them correctly based on general knowledge, without actually reading the story. The negative correlation between the proportion of skipped turns and comprehension is consistent with findings with the app with older students (Beigman Klebanov et al., 2019).

While higher comprehension typically does correlate with higher fluency in the same readers in assessment data (Wise et al., 2010), it is in principle possible to read slowly but with good comprehension. This is likely in our case since students are not reading for a test and are not being urged to read fast. The scatterplot of RCQ and WCPM scores in Figure 2 shows some low fluency readers who nevertheless showed strong comprehension – there are three students with comprehension $\geq$ .80 who read with fluency around 60-75 WCPM (72 WCPM is the 25th percentile in fluency, according to norms). Further analysis of these individuals revealed that while one of these readers had near perfect reading accuracy, the other two struggled, with an average reading accuracy between 72-76% and all three very rarely skipped their turn ($\leq$ 8% turns skipped). All three reported using persistent reading strategies in the survey. Though a small sample, these students illustrate our hypothesis about persistent reading behavior; despite low fluency, they consistently complete their turns (low skip proportion), suggesting they may be encountering and overcoming difficulties reading.

[4]Only 15 of our 18 students had audio recordings of adequate quality to estimate an average ACC and WCPM.

[5]We chose to collapse scores into two categories because the items assessed basic recall and understanding. Given the simplicity of the questions, incorrect responses likely reflected substantial comprehension difficulties or inattention.

In terms of the relationship between the reading data and the survey data, there are some discrepancies. In particular, 17 out of 18 students said they would sound out a word they didn't know and most admitted that there were some or lots of words they didn't know. However, the process data suggests that there were 3 students who did not spend enough time on most of their turns to read to any substantive degree (SKT$\geq$0.5), let alone sound out difficult words. However, the interpretation may not be straightforward. First, students may have responded to the survey question as if it asked about a general habit rather than about reading in the app specifically; they may be sounding words out in other contexts but not in this reading context, which they may have perceived as more informal. Second, the teacher told us she explicitly instructs them to sound out unfamiliar words as they read, so their response to the survey may have reflected what they thought they should be doing rather than what they actually did.

5.2 Limitations and Future Research

The study was carried out in a single classroom with only 18 students who had parental consent and participated in all activities. The small sample size limited our ability to detect significant relationships between variables. Furthermore, adapting the surveys to this young population limited the number of questions and response options that we could include within a single survey administration.

Students generally reacted positively to the pedagogical agent-based activity, based on observations of the research team member who assisted the teacher, conversations with the teacher, and on the students' responses where they expressed readiness or tentative readiness to chat with the pedagogical agent again. We are thus encouraged to continue exploring the utility of the agent through co-design with teachers for expanding the agent's role beyond the self-report-based assessment function. In a preliminary focus group we conducted with three elementary school teachers to start addressing this issue, the clearest message was that the agent should try and encourage the students in their reading endeavor. In addition, we intend to implement a more flexible conversation with students, using automated speech recognition and an LLM that would help generate a larger variety of agent responses, with the caveat that strong guardrails would need to be implemented to ensure that the agent's conversation is appropriate. We envision using an LLM to create a diverse set of responses that would be vetted and placed into a database for the agent to choose from rather than allowing students to interact with the LLM directly.

Weak correlations between self-reports and behavioral indicators for constructs such as persistence, metacognition, and self-regulation are well documented in the literature (e.g., Craig et al., 2020). These discrepancies can point to limitations in the survey design (Desoete, 2008), but they may also yield meaningful insights into students' developmental trajectory in their own self-awareness (Andrade, 2019). Reflecting on our own findings, a fruitful direction for future research would be to explore when and how we prompt young learners to self-report, such as asking why they skipped some of their reading turns when those behaviors occur. That would require better personalization of the survey so that the question is only posed to students who did skip their turns as a matter of course. A conversational agent that is connected to the process data will be able to deliver such personalization. Embedding surveying within the activity would also support capturing students' motivations and metacognitive awareness in context, providing insight into how their behaviors and beliefs relate to their reading experience as it unfolds. This type of survey administration would help support future efforts to better understand how student persistence varies over time and across learning contexts. Students' overall positive response and receptivity to the agent-delivered survey suggest this is a promising approach for integrating surveys into the activity without disrupting engagement. However, these findings stem from a small proof-of-concept study and should not be interpreted as validation of the approach itself.

6 Conclusion

We investigated the extent to which a pedagogical agent can function appropriately in an assessment or surveying role with young learners in a classroom environment. The results of our study showed that pedagogical agents hold promise for engaging students in conversational surveys following a tablet-based reading task. However, our study was limited to one classroom, and thus more research is needed to understand in what learning scenarios and for what learners pedagogical agents are appropriately positioned in a testing or surveying role.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) and the Institute of Education Sciences (IES) under Grant DRL-2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF or IES. We are grateful to Jesse Sparks and Catherine Castellano for their feedback on drafts of this paper.

References

Heidi L Andrade. 2019. A critical review of research on student self-assessment. In *Frontiers in education*, volume 4, page 87. Frontiers Media SA.

Beata Beigman Klebanov and Anastassia Loukina. 2021. Exploiting structured error to improve automated scoring of oral reading fluency. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 76–81.

Beata Beigman Klebanov, Anastassia Loukina, Nitin Madnani, John Sabatini, and Jennifer Lentini. 2019. Would you? could you? on a tablet? analytics of children's ebook reading. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 106–110, Tempe, AZ, USA. ACM.

Oleksandr S Chernyshenko, Miloš Kankaraš, and Fritz Drasgow. 2018. Social and emotional skills for student success and well-being: Conceptual framework for the oecd study on social and emotional skills. *OECD Education Working Papers*, 173.

Geraldine Clarebout, Jan Elen, Lewis Johnson, and Erin Shaw. 2002. Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational multimedia and hypermedia*, 11(3):267–286.

David A Cook, Becca L Gas, David R Farley, Matthew Lineberry, Nimesh D Naik, Francisco J Cardenas Lara, and Anthony R Artino Jr. 2019. Influencing mindsets and motivation in procedural skills learning: two randomized studies. *Journal of Surgical Education*, 76(3):652–663.

Kym Craig, Daniel Hale, Catherine Grainger, and Mary E Stewart. 2020. Evaluating metacognitive self-reports: systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, 15:155–213.

Annemie Desoete. 2008. Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning*, 3:189–206.

Carol S Dweck, Gregory M Walton, and Geoffrey L Cohen. 2014. Academic tenacity: Mindsets and skills that promote long-term learning. *Bill & Melinda Gates Foundation*.

Camille A Farrington, Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W Johnson, and Nicole O Beechum. 2012. *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance–A Critical Literature Review*. University of Chicago Consortium on Chicago School Research.

Louise Gascoine, Steve Higgins, and Kate Wall. 2017. The assessment of metacognition in children aged 4–16 years: a systematic review. *Review of Education*, 5(1):3–57.

Tiong-Thye Goh. 2025. Learning management system log analytics: the role of persistence and consistency of engagement behaviour on academic success. *Journal of Computers in Education*, pages 1–24.

Manu Kapur. 2008. Productive failure. *Cognition and Instruction*, 26(3):379–424.

Van Rynald T Liceralde, Anastassia Loukina, Beata Beigman Klebanov, and John R Lockwood. 2022. Beyond text complexity: Production-related sources of text-based variability in oral reading fluency. *Journal of Educational Psychology*, 114(1):16.

Anastassia Loukina, Beata Beigman Klebanov, Patrick L Lange, Yao Qian, Binod Gyawali, Nitin Madnani, Abhinav Misra, Klaus Zechner, Zuowei Wang, and John Sabatini. 2019. Automated estimation of oral reading fluency during summer camp e-book reading with My Turn To Read. In *Proceedings of INTERSPEECH*, pages 21–25.

Anastassia Loukina, Beata Beigman Klebanov, Patrick Lange, Binod Gyawali, and Yao Qian. 2017. Developing speech processing technologies for shared book reading with a computer. In *Proc. WOCCI 2017: 6th International Workshop on Child Computer Interaction*, pages 46–51.

Nitin Madnani, Beata Beigman Klebanov, Anastassia Loukina, Binod Gyawali, Patrick L Lange, John Sabatini, and Michael Flor. 2019. My Turn to Read: An interleaved e-book reading tool for developing and struggling readers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146.

Katrin Maegi, Eve Kikas, and Piret Soodla. 2018. Effortful control, task persistence, and reading skills. *Journal of Applied Developmental Psychology*, 54:42–52.

Amy L Reschly and Sandra L Christenson. 2022. *Handbook of research on student engagement*. Springer.

J Elizabeth Richey, Juan Miguel L Andres-Bray, Michael Mogessie, Richard Scruggs, Juliana MAL Andres, Jon R Star, Ryan S Baker, and Bruce M McLaren. 2019. More confusion and frustration, better learning: The impact of erroneous examples. *Computers & Education*, 139:173–190.

Noah L Schroeder, Robert O Davis, and Eunbyul Yang. 2025. Designing and learning with pedagogical agents: An umbrella review. *Journal of Educational Computing Research*, 62(8):2127–2156.

Noah L Schroeder and Chad M Gotch. 2015. Persisting issues in pedagogical agent research. *Journal of Educational Computing Research*, 53(2):183–204.

Valerie J Shute, Matthew Ventura, and Yoon Jeon Kim. 2013. Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research*, 106(6):423–430.

Robert F Siegle, Noah L Schroeder, H Chad Lane, and Scotty D Craig. 2023. Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends*, 67(5):851–864.

Ellen A Skinner, Thomas A Kindermann, Justin W Vollet, and Nicolette P Rickert. 2022. Complex social ecologies and the development of academic motivation. *Educational Psychology Review*, 34(4):2129–2165.

J. R. Sparks and B. Lehman. 2025. Measuring persistence and academic resilience: Analytic review and operational definitions. In *Paper presented at the annual meeting of the American Educational Research Association. Denver, CO.*

Jessica R Toste, Lisa Didion, Peng Peng, Marissa J Filderman, and Amanda M McClelland. 2020. A meta-analytic review of the relations between motivation and reading achievement for K–12 students. *Review of Educational Research*, 90(3):420–456.

Ming-Te Wang, Jessica L Degol, Jamie Amemiya, Alyssa Parr, and Jiesi Guo. 2020. Classroom climate and children's academic and psychological wellbeing: A systematic review and meta-analysis. *Developmental Review*, 57:100912.

Justin C Wise, Rose A Sevcik, Robin D Morris, Maureen W Lovett, Maryanne Wolf, Melanie Kuhn, Beth Meisinger, and Paula Schwanenflugel. 2010. The relationship between different measures of oral reading fluency and reading comprehension in second-grade students who evidence different oral reading fluency difficulties. *Language, Speech, and Hearing Services in Schools*, 41(3):340–348.

Shan Zhang, Chris Davis Jaldi, Noah L Schroeder, and Jessica R Gladstone. 2024a. Pedagogical agents in K-12 education: a scoping review. *Journal of Research on Technology in Education*, pages 1–28.

Shan Zhang, Chris Davis Jaldi, Noah L Schroeder, Alexis A López, Jessica R Gladstone, and Steffi Heidig. 2024b. Pedagogical agent design for K-12 education: A systematic review. *Computers & Education*, page 105165.

A Appendix: Survey Interface and Script

To provide further detail on the pedagogical agent's role in survey administration, we include a screenshot (Figure A1) of the agent interface as it appeared during the survey session with students. The image depicts the agent presenting a question with response options rendered as on-screen buttons, designed to scaffold independent interaction for early elementary learners.

Figure A1: Screenshot of the pedagogical agent delivering a survey question.

In addition to the interface example, Table A1 provides the full conversational script used by the agent during survey delivery. This includes all administered items as well as transitional dialogue interludes designed to maintain a conversational tone and scaffold student engagement. The script also includes conditional statements (e.g., varied follow-up prompts) based on students' prior responses. This structure reflects our effort to position the survey as an interactive, child-appropriate experience rather than a traditional assessment.

See Table A1 (next page) for the complete set of agent interactions, response options, and associated constructs.

Table A1: Conversational Agent Survey Script

Q#	**Agent Prompt**	**Student Response Options**	**Construct**
Hi, I am Adam. Let's chat about the reading you've done on the Kindle. I'm curious how that went.			
1	Ready to begin?	Let's go!	–
2	Did you like the story you read today? Please tap on your answer for me.	Yes / No	Subjective Experience
Let's see. Which story was that? Hmmm…			
3	About a frog and a dog?	Yes / No	Comprehension
4	About a cat and a dog?	Yes / No	Comprehension
5	About a grandma and a dog?	Yes / No	Comprehension
6	What sort of grandma was she?	Very tired / Deaf / Upset / Singing	Comprehension
<TEACHER> told me you read stories about Willie the Donkey, Chippy's Birthday, and Grandma and the Dog.			
7	Which of these should I read next?	Willie / Chippy / Grandma / None	Subjective Experience
Ok, thank you! I was thinking about reading about Chippy's Birthday, because my friend's birthday is next week, you know? [If the student selected Willie or Grandma above: *"But now I think I'll read the <X> story first."]*			
8	Do you like reading?	Very much / It's OK / Not really / I hate reading	Reading Enjoyment
<TEACHER> said you read out loud today.			
9	How did reading out loud feel?	Easy / OK / Hard	Reading Ease
I see. [If response to Q9 = Easy: *"Easy-peasy!"*; if Hard: *"Tough going."]*			
10	What do you do if you don't know how to say a word?	Sound it out / Ask someone / Skip / Stop reading	Persistence
11	Were there lots of words in the story you didn't know how to say?	Yes, lots! / Only a few / None	Reading Ease
Sounds like reading out loud was [insert: *"pretty hard" / "not too hard" / "a breeze"* based on Q11].			
12	What do you do if you don't know what a word means?	Figure it out / Ask someone / Skip / Stop reading	Persistence
13	Were there lots of words in the story you didn't understand?	Yes, lots! / Only a few / None	Reading Ease
[If Q13 = lots or few: *"You'll figure it out!"*; if Q13 = none: *"You got it!"]*			
14	Do you feel like you learn a lot when you read?	Definitely / Sometimes / Not really	Learning
15	Thanks for chatting with me! It was fun, for me. Should we chat again sometime?	Sure / Maybe / No, thank you	Subjective Experience
Bye-bye!			

LLM-Based Approaches for Detecting Gaming the System in Self-Explanation

Jiayi Zhang
University of Pennsylvania
Philadelphia, PA, United States
joycez@upenn.edu

Ryan S. Baker
Adelaide University
Adelaide, Australia
ryanshaunbaker@gmail.com

Bruce M. McLaren
Carnegie Mellon University
Pittsburgh, PA, United States
bmclaren@andrew.cmu.edu

Abstract

Self-explanation supports deeper learning by prompting students to articulate their reasoning and connect new concepts with prior knowledge. Open-ended self-explanation questions promote elaborative processing and help address knowledge gaps. However, these benefits may be undermined when students game the system — a maladaptive learning strategy where students exploit the learning environment rather than engaging in meaningful learning. While previous studies have successfully detected this behavior in students' interactions with learning activities, this study focuses on identifying such behavior in students' open-ended responses within a math digital learning game. We evaluated two large language model (LLM)-based approaches: one using sentence embeddings and another using a prompt-based method. Both showed acceptable performance, but the embedding-based model outperformed the prompt-based one. Error analysis revealed the prompt-based model struggled with short, low-context responses and produced false positives when students referenced using hints. Consistent with earlier findings, we showed that higher rates of gaming behavior in open-ended responses negatively correlated with learning gains.

1 Introduction

Self-explanation, an important pedagogical strategy, has been frequently used in classrooms to facilitate learning. During this process, students articulate their reasoning, connect new information with prior knowledge, and identify gaps in their understanding (Fonseca and Chi, 2011; Wylie and Chi, 2014). Self-explanation can be self-initiated or externally prompted. Previous studies have shown that self-explanation leads to improved performance, deeper conceptual understanding, and better long-term retention (Bisra et al., 2018; VanLehn et al., 1992). In mathematics learning, students who engage in self-explanation are more likely to develop a more robust understanding of problems and improve their ability to transfer knowledge to novel situations (McEldoon et al., 2013; Rittle-Johnson, 2006).

Given these benefits, self-explanation questions have been increasingly integrated into digital learning platforms. However, due to the limitations of digital learning systems—which, until recently, had a limited ability to process natural language and provide feedback—self-explanation questions have often been designed in a closed-ended format, such as multiple-choice, fill-in-the-blank questions, or sentence builders (McLaren et al., 2022). Nonetheless, open-ended self-explanation questions "may invite elaborative processing better adapted to each learner's unique gaps in knowledge" (Bisra et al., 2018) and encourage deeper cognitive processing (Kwon et al., 2011). A recent study comparing three self-explanation formats (multiple-choice, fill-in-the-blank, and open-ended) found that students who answered open-ended self-explanation questions achieved the greatest learning gains (McLaren et al., 2022).

However, failing to engage meaningfully with these self-explanation questions can potentially diminish the positive effects. In gaming the system, a disengaged behavior and maladaptive learning strategy, students attempt to succeed by exploiting system properties rather than engaging in meaningful learning, resorting to behaviors such as systematic guessing or abusing hints (Baker et al., 2008). Gaming the system has been observed across platforms and is consistently associated with lower learning gains and long-term negative outcomes (e.g., Baker et al. (2006b); Cocea et al. (2009)). In a previous study, the negative effects of gaming have also been demonstrated within (non-open-ended) self-explanation questions, in which students who had a higher rate of gaming were associated with lower learning gain. Furthermore, the rate of gaming in the self-explanation moderated the differences

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 91–98
October 27-29, 2025

in learning between boys and girls (Baker et al., 2024).

As such, to support interventions, gaming detectors have been developed in the past to identify instances when students game the system (Li et al., 2022; Xia et al., 2020). However, most of these detectors are designed for close-ended questions, which identify gaming based on interaction patterns with learning activities. A few gaming detectors for text-based open-ended responses have primarily focused on response patterns (e.g., detecting repetition in open-ended responses) rather than analyzing the semantic content of the inputs (Darvishi et al., 2022). For example, identifying instances where students game the system by cycling through answers, entering responses such as "It will be 7.1", "It will be 7.2", "It will be 7.3". As a result, a significant gap remains in detecting gaming behaviors in the open-ended responses.

The advancement of large language models (LLMs) presents an opportunity for this use case. Trained on vast amounts of text data, these models have demonstrated capabilities in processing, understanding, and generating natural language with high accuracy (Brown et al., 2020). As a result, LLMs have been increasingly used to analyze and categorize textual data, presenting an opportunity to perform classification tasks such as assessing the correctness or relevance of self-explanations (Nguyen et al., 2023) or identifying the presence or absence of gaming in open-ended responses. One common approach to leveraging LLMs for classification tasks is through sentence embeddings, where text inputs are transformed into high-dimensional vectors that capture semantic meaning. These embeddings can then be input into machine learning models to categorize responses. Alternatively, prompt-based methods (e.g. Generative Pre-trained Transformer; GPT) frame classification tasks as text-generation problems, allowing pre-trained LLMs to infer labels based on contextual prompts. Several studies have found that classifying embeddings outperforms prompt-based approaches in various classification tasks (Liu et al., in press; Hutt et al., 2024). Recent studies have explored prompt engineering, examining how one-shot (providing one example), few-shot (providing a few examples), adding context (Xiao et al., 2023), modifying prompt structure (White et al., 2023), and defining roles influence model performance (Hou et al., 2024). However, less research has explored where the two approaches diverge and under what conditions or context one approach is more effective than the other, evaluating and comparing the validity and reliability of the two approaches for classification tasks.

In this study, we explored the use of large language models (LLMs) to detect gaming the system in open-ended responses to self-explanation questions within a math digital learning game. We identified gaming behavior using both an embedding-based and a prompt-based approach and compared their performance. To understand where the two approaches diverge, we conducted an error analysis examining the types of errors each approach is prone to, highlighting the context under which one approach might be more efficient than the other. Lastly, we applied the best-performing model to the full dataset and conducted analyses to examine the relationships between gaming during the self-explanation step and learning gains within this learning system. By detecting gaming the system in this additional context, we enhance our understanding of how broadly this phenomenon occurs and enable learning technologies to intervene in a wider range of contexts. Additionally, the comparison between the two approaches contributes to the growing body of research on leveraging LLMs for text classification.

2 Methods

2.1 Learning Platform and Data

Student log data were collected from Decimal Point, a single-player web game designed to motivate middle-school students to learn decimal concepts (McLaren, 2024; McLaren et al., 2017). Students wander through a virtual amusement park and play a variety of mini-games that incorporate decimal challenges, such as sorting decimals. In the version of the game where the data was collected, students were first asked to solve a problem (problem-solving step) and then prompted to reflect on how they solve the problem and explain their reasoning with an open-ended self-explanation question (self-explanation step) (McLaren et al., 2022). To assure that students expend at least minimal effort in answering the self-explanation questions, the response needed to contain at least four words with at least one of the words from a relevant list (including common misspellings) that would legitimately be found in a correct explanation. Students could make multiple attempts and could only move to the next question once the response meets these criteria.

To investigate LLM's ability at detecting gaming in open-ended responses, we collected the text-based responses submitted by 212 students and delineated them into clips, with each clip containing all the attempts (responses) a student submitted at answering a self-explanation question. In total, 2553 clips were extracted. We also collected students' pre-test, post-test, and delayed post-test scores.

2.2 Coding Gaming the System

Text replay coding was conducted to establish ground truth. In text replays, human coders examine each clip and determine the presence or absence of gaming the system using a codebook (Baker et al., 2006a). The codebook was developed through an iterative process to ensure that the behaviors classified as gaming aligned with previous conceptualizations (e.g. as defined in Baker et al. (2008)) and were salient in the dataset. Through this process, we developed a codebook consisting of three criteria: (1) a low degree of semantic difference between consecutive responses – e.g. changing between highly related alternatives, (2) systematically cycling through modifications to responses or potential multiple answers, and (3) making a conceptual or functional change between responses (e.g., identifying a concept versus suggesting an action, trying to figure out what category of response is needed without thinking through the question) in conjunction with the previous two criteria. The gaming criteria and examples are presented in Table 1.

Using the codebook, two coders first independently coded the same set of data to establish inter-rater reliability (κ = 0.8). Once consensus was reached, the coders proceeded to code a total of 1,465 clips from 116 students, of which 8.9% were positive (gaming) clips.

2.3 Approach 1: Detecting Gaming with Sentence Embeddings

To train models that automatically detect gaming, we first con-catenated textual responses from all attempts within a clip, separating each attempt with a period. We then vectorized the text using two sentence embedding models: the Universal Sentence Encoder Large v5 (USE) developed by Google, which generates a 512-dimensional vector for each entry (Cer et al., 2018), and sentence-embedding-3-short developed by OpenAI, which produces a 1,536-dimensional vector (Neelakantan et al., 2022).

For each set of embeddings, we trained a neural network model with one hidden layer to predict the presence or absence of gaming. The models were evaluated using 5-fold student-level cross-validation. Model performance was evaluated using the average Area Under the Receiver Operating Characteristic Curve (AUC) and Kappa.

2.4 Approach 2: Detecting Gaming using Prompt-Based Model

For prompt-based methods, we leveraged both zero-shot and one-shot prompting techniques, providing the GPT-4-turbo model with the definition of gaming the system and the three criteria from the codebook for zero-shot prompting, and the corresponding examples (as listed in the codebook) for one-shot prompting. The exact prompt used for zero-shot prompting is presented below. For one-shot prompting, examples were added to the prompt. The temperature was set to 0 to minimize randomness. To account for the stochastic nature of GPT, we ran the prompt three times to assess consistency across iterations. The final prediction was determined using majority voting across the three outputs. The predictions were evaluated against the ground truth using AUC and Kappa.

> "Review the provided text and code it based on the construct: gaming the system. The definition of this construct is: a maladaptive learning strategy where students attempt to succeed by exploiting properties of a learning environment. Some criteria of gaming the system in open-ended responses include: 1) a low degree of semantic difference between responses, 2) cycling through multiple answers/ modifications to their responses, or 3) conceptual or functional change between responses (e.g., identifying a concept versus suggesting an action) accompanied by the previous two criteria. After reviewing the text, assign a code of '1' if you believe the text exemplifies gaming the system, or a '0' if it does not. Your response should only be '1' or '0'. TEXT TO BE REVIEWED: [TEXT]"

Gaming Criteria	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Minor Semantic Difference	I need to move it vertically	Move side to side	–	–
Cycling through Modifications	It will be 7.1	It will be 7.2	It will be 7.3	–
Conceptual or Functional Change	It is 1.7	It is 1.9	By adding	By subtracting

Table 1: Examples of gaming behaviors across multiple attempts.

3 Results

3.1 Model Performance

As shown in Table 2, with 5-fold student-level cross-validation, the neural network model built using sentence embeddings from the Universal Sentence Encoder achieved an average AUC of 0.902 and a Kappa of 0.535. The neural network model using sentence-embedding-3-short as the encoder performed better, reaching an average AUC of 0.935 and a Kappa of 0.564. In contrast, the prompt-based model with zero-shot prompting achieved an AUC of 0.699 and a Kappa of 0.345, and an AUC of 0.754 and a Kappa of 0.358 with one-shot prompting. We also recorded the number of false positive and false negative cases for each model, which is discussed in the next section.

3.2 Error Analysis

To examine differences in prediction accuracy across the models, we conducted an error analysis using both quantitative and qualitative methods, counting the number of type I and type II errors as well as reviewing the responses the models misclassified.

As shown in Table 2, both sentence embedding models were more likely to make Type II errors (false negatives) than Type I errors (false positives), meaning they incorrectly assessed the student as not gaming when the response actually demonstrated gaming behaviors. In contrast, the prompt-based models were more prone to Type I errors (false positives) than Type II errors (false negatives), predicting gaming when the student was not actually gaming. Additionally, Type I errors were twice as frequent for the prompt-based models compared to the sentence embedding models.

To better understand where the models failed to make accurate predictions, we examined the misclassified cases, analyzing responses in which there was a discrepancy between the sentence embedding approach and the prompt-based approach. Of the 1,465 responses, 178 were correctly classified by both sentence embedding models (Universal Sentence Encoder and sentence-embedding-3-short) but misclassified by at least one of the prompt-based approaches. Among these, 25 had a true label of gaming, and 153 had a true label of not gaming.

Upon examining these cases, we identified several patterns. One common pattern among **false positives** for the prompt-based models was responses that are not considered gaming in this particular dataset but could be considered gaming if gaming were defined more broadly. For example, some responses mentioned the use of hints. In one instance where the prompt-based model falsely classified the behavior as gaming, the student (somewhat oddly) stated "Always remember to use the hint button. It gives you the answer if you click until it doesn't say 'next,' and you should get the answer correct if you follow what it says." Another false positive example is when a student said, "22.0. You have to add. 22.0. You can look at the hints to find the answer. You can find this answer by adding 17.6 + 4.4." In these cases, the prompt-based model flagged the responses as gaming likely due to mentions of hints, but they may not strictly align with the definition of gaming behavior in this specific context.

Comparing between zero-shot and one-shot prompting for the false positive cases, we noticed that the majority of cases misclassified by the one-shot model were responses that repeated themselves without any semantic changes. For example, when asked, "Is 0.2 bigger or smaller than 0.22? How do you know?", a student responded, "It is smaller. It is smaller." The model with one-shot prompting misclassified this as gaming, whereas the zero-shot model correctly classified it as not gaming, as there was no semantic difference between the two entries, and it didn't imply a cycling behavior.

A common pattern among **false negatives** for the prompt-based model with zero-shot prompting was that shorter responses lacked sufficient con-

Model	AUC (stdev)	Kappa (stdev)	False Positive	False Negative
Universal sentence encoder	0.902 (0.038)	0.535 (0.087)	52	73
sentence-embedding-3-short	0.935 (0.026)	0.564 (0.088)	44	70
Prompt-based zero-shot	0.699	0.345	112	68
Prompt-based one-shot	0.754	0.358	169	48

Table 2: Classification results and total errors.

text for the model to accurately interpret gaming behavior. For example, when asked the same question, "Is 0.2 bigger or smaller than 0.22? How do you know?" a student responded, "Smaller. Bigger. Bigger. Bigger. Smaller. Smaller because 0.22 has an extra digit than 0.2." Due to the brevity of the response, the model may have struggled to contextualize it properly, leading to a misclassification. However, this is less frequent with one-shot prompting, possibly because of the brevity in the examples provided.

Altogether, these patterns suggest that the prompt-based model may struggle with nuanced cases where gaming behaviors depend on context, leading to predictions that are not context-specific. Specifically, it tends to misclassify responses that mention hints or shortcuts as gaming, even when they might not strictly fit the definition based on the current operationalization. Compared to zero-shot, one-shot prompting is also more prone to Type II errors, misclassifying cases where students repeat responses as gaming rather than as recycling responses with minimal semantic changes. Conversely, prompt-based approach struggles to detect gaming in shorter responses that lack sufficient context, especially when examples are not provided.

The same qualitative approach was conducted to evaluate the predictions of the embedding-based models, focusing on responses that were correctly classified by both prompt-based models but misclassified by at least one of the embedding-based models. Of the 1,465 responses, 84 were correctly classified by both prompt-based models (zero-shot and one-shot) but misclassified by at least one of the embedding-based models. Among these, 32 had a true label of gaming, and 52 had a true label of not gaming.

By analyzing the **false negative** cases, we found that, similar to zero-shot prompting, sentence-embedding models are prone to Type II errors when responses are brief and seemingly disjointed. This issue is especially apparent when key explanatory words (such as “because”) are missing. For example, when asked, “Is 0.456 to the left of 0 or to the right of 0 on the number line? How do you know?”, one student responded, “Right. 0.5. Left. 0.45. 0.45 to the right.” Another example comes from the question, “Is 6.5 bigger or smaller than 6.41? How do you know?”, to which a student responded, “6.5 is smaller. 6.41 is smaller. 6.41 is bigger.” These responses clearly reflect cycling behavior, even though they lack explanatory words (such as “because”) that directly address the question’s explanatory prompt. Sentence-embedding models failed to detect gaming in such cases possibly because they rely on overall semantic similarity to the example cases (e.g., frequent usage of explanatory terms) and lack the contextual understanding needed to recognize patterns like repetitive guessing or cycling.

3.3 Gaming the System and Learning Gains

After applying the best model (the model trained using embeddings derived from sentence-embedding-3-short) to the full dataset (2,553 clips), we found that students’ detected frequency of gaming was negatively correlated with the pre-test ($r = -0.233$, $p = 0.058$), post-test ($r = -0.312$, $p = 0.010$), and delayed post-test ($r = -0.355$, $p = 0.003$). We found that gaming frequency was not correlated with normalized learning gains between the pre-test and post-test ($r = -0.121$, $p = 0.329$), but was negatively and significantly correlated with normalized learning gains between the pre-test and delayed post-test ($r = -0.247$, $p = 0.044$).

4 Discussion and Conclusion

4.1 Main Findings

Self-explanation promotes deeper learning by helping students articulate their reasoning and connect new information with prior knowledge. Open-ended self-explanation questions, in particular, foster more elaborative processing, allowing students to address their unique knowledge gaps. However, these benefits can be undermined when students disengage and attempt to game the system. This study addresses this challenge by introducing an automated approach to detect gaming in open-ended responses using large language models (LLMs). Specifically, we compare a sentence embedding-based method with a prompt-based approach. By identifying gaming behavior in real time, this method can support targeted interventions, such as adaptive feedback, to help students re-engage and maximize the benefits of self-explanation.

Our results show that while all models demonstrate reliable performance in detecting gaming in open-ended responses, the sentence embedding-based approach, particularly the OpenAI sentence-embedding-3-short model, outperformed the prompt-based method, achieving an AUC of 0.935 and a Kappa of 0.564. While the prompt-based model was easier to implement, it was more prone to false positives, frequently misclassifying responses that mentioned hints or repeated responses as gaming. These results highlight the challenges of using prompt-based models for nuanced classification tasks, particularly when the definition of the target behavior is context-dependent

Additionally, both prompt-based and sentence-embedding-based models struggled with shorter, context-poor responses, leading to false negatives. However, this issue can be attenuated with one-shot prompting.

Overall, the comparison between the two approaches suggests that sentence-embedding is more conservative in detecting gaming, making it more prone to Type II than Type I errors, at least for this application. On the other hand, the prompt-based approach—possibly due to access to additional contextual information provided in the prompt is more liberal and less context-specific, making it more prone to Type I than Type II errors. These findings may suggest a direction for future study to explore the possibility of combining the two approaches and leveraging them for their strengths. It is also possible to adapt the model selection based on the data as well as the desired outcomes.

Furthermore, we found that the frequency of detected gaming behavior was negatively correlated with students' pre-test, post-test, delayed post-test scores, and delayed learning gains, suggesting that gaming the system in this context is also associated with lower learning outcomes. This aligns with previous research that has consistently linked gaming behavior with reduced learning gains (Baker et al., 2008; Cocea et al., 2009).

4.2 Future Work

We acknowledge the following limitations. First, it is possible that the prompt-based model's performance may have been constrained by the limited prompt engineering employed in this study, for instance, not providing more specific context information for self-explanations. Future work could explore more sophisticated prompting strategies, such as few-shot learning, where the model is provided with more than one labeled example to improve its performance. Additionally, fine-tuning the LLM on a domain-specific dataset could further enhance its ability to detect gaming especially in contexts where nuanced semantic understanding is critical.

Second, the generalizability of our findings may be limited by the specific context in which gaming is being operationalized. Future studies should validate these approaches in other learning environments and with more diverse datasets. This would help determine whether the observed patterns hold across different digital learning contexts and student populations.

Finally, while our study focused on detecting gaming behavior, future research could explore the possibility of distinguishing specific gaming behaviors (e.g., minor semantic differences or cycling through modifications) and examine whether they impact learning outcomes differentially.

4.3 Conclusion

In contrast to previous gaming detectors based on interaction data, this study demonstrates the potential of using LLMs to detect gaming behavior in open-ended self-explanation responses by identifying gaming based on the semantic meaning of text-based responses. Our findings suggest that sentence embedding-based approaches are more effective than prompt-based methods for this task, possibly because the definition of gaming the system is context-dependent. Consistent with prior research, we found that gaming in open-ended self-explanation questions is also negatively correlated with learning gains, emphasizing its detrimental impact and the need for intervention. The ability to detect gaming in open-ended responses opens new possibilities for intervention and support in digital learning environments, helping ensure that students engage meaningfully with self-explanation tasks and achieve better learning outcomes.

References

R.S. Baker, A.T. Corbett, and A.Z. Wagner. 2006a. Human classification of low-fidelity replays of student actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, pages 29–36.

R.S. Baker, J.E. Richey, J. Zhang, S. Karumbaiah, J.M. Andres-Bray, H.A. Nguyen, J.M.A.L. Andres, and B.M. McLaren. 2024. Gaming the system mediates the relationship between gender and learning outcomes in a digital learning game. *Instructional Science*.

R.S.J. d. Baker, A.T. Corbett, K.R. Koedinger, S. Evenson, I. Roll, A.Z. Wagner, M. Naim, J. Raspat, D.J. Baker, and J.E. Beck. 2006b. Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems*, pages 392–401. Springer Berlin Heidelberg.

R.S.J.D. Baker, A.T. Corbett, I. Roll, and K.R. Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3):287–314.

K. Bisra, Q. Liu, J.C. Nesbit, F. Salimi, and P.H. Winne. 2018. Inducing self-explanation: a meta-analysis. *Educational Psychology Review*, 30(3):703–725.

T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, and T. Henighan. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R.S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder. *arXiv preprint*, arXiv:1803.11175.

Mihaela Cocea, Arnon Hershkovitz, and Ryan S.J.d. Baker. 2009. The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In *Frontiers in Artificial Intelligence and Applications*, pages 207–514. IOS Press.

A. Darvishi, H. Khosravi, S. Sadiq, and D. Gašević. 2022. Incorporating ai and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, 53(4):844–875.

B.A. Fonseca and M.T. Chi. 2011. Instruction based on self-explanation. In *Handbook of Research on Learning and Instruction*, pages 310–335. Routledge.

C. Hou, G. Zhu, J. Zheng, L. Zhang, X. Huang, T. Zhong, S. Li, H. Du, and C.L. Ker. 2024. Prompt-based and fine-tuned gpt models for context-dependent and -independent deductive coding in social annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 518–528, Kyoto, Japan.

S. Hutt, A. DePiro, J. Wang, S. Rhodes, R.S. Baker, G. Hieb, S. Sethuraman, J. Ocumpaugh, and C. Mills. 2024. Feedback on feedback: Comparing classic natural language processing and generative ai to evaluate peer feedback. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 55–65, Kyoto, Japan.

K. Kwon, C.D. Kumalasari, and J.L. Howland. 2011. Self-explanation prompts on problem-solving performance in an interactive learning environment. *Journal of Interactive Online Learning*, 10(2).

Y. Li, X. Zou, Z. Ma, and R.S. Baker. 2022. A multi-pronged redesign to reduce gaming the system. In *International Conference on Artificial Intelligence in Education*, pages 334–337.

X. Liu, A.F. Zambrano, R.S. Baker, A. Barany, J. Ocumpaugh, J. Zhang, M. Pankiewicz, N. Nasiar, and Z. Wei. in press. Qualitative coding with gpt-4: Where it works better. *Journal of Learning Analytics*.

K.L. McEldoon, K.L. Durkin, and B. Rittle-Johnson. 2013. Is self-explanation worth the time? a comparison to additional practice. *British Journal of Educational Psychology*, 83(4):615–632.

B.M. McLaren. 2024. Decimal point: A decade of learning science findings with a digital learning game. In P. Ilic, I. Casebourne, and R. Wegerif, editors, *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, pages 145–203. Springer Nature Switzerland.

B.M. McLaren, D.M. Adams, R.E. Mayer, and J. Forlizzi. 2017. A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning*, 7(1):36–56.

B.M. McLaren, J.E. Richey, H.A. Nguyen, and M. Mogessie. 2022. Focused self-explanations lead to the best learning outcomes in a digital learning game. In *Proceedings of the 16th International Conference on Learning Science*, pages 36–56.

A. Neelakantan and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint*, arXiv:2201.10005.

H.A. Nguyen, H. Stec, X. Hou, S. Di, and B.M. McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*, pages 278–293, Cham.

B. Rittle-Johnson. 2006. Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, 77(1):1–15.

K. VanLehn, R.M. Jones, and M.T.H. Chi. 1992. A model of the self-explanation effect. *Journal of the Learning Sciences*, 2(1):1–59.

J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D.C. Schmidt. 2023. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint*, arXiv:2302.03459.

R. Wylie and M.T. Chi. 2014. The self-explanation principle in multimedia learning. In *The Cambridge Handbook of Multimedia Learning*, pages 413–432.

M. Xia, Y. Asano, J.J. Williams, H. Qu, and X. Ma. 2020. Using information visualization to promote students' reflection on "gaming the system" in online learning. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, pages 37–49, Virtual Event USA.

Z. Xiao, X. Yuan, Q.V. Liao, R. Abdelghani, and P.-Y. Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *28th International Conference on Intelligent User Interfaces*, pages 75–78, Sydney, NSW, Australia.

Evaluating the Impact of LLM-guided Reflection on Learning Outcomes with Interactive AI-Generated Educational Podcasts

Vishnu Menon[1], Andy Cherney[1], Elizabeth B. Cloude[2], Li Zhang[1], Tiffany D. Do[1]

[1]Drexel University, [2]Michigan State University

Correspondence: {vishnu.v.menon|harry.zhang|tiffany.do}@drexel.edu, cloudeel@msu.edu

Abstract

This study examined whether embedding LLM-guided reflection prompts in an interactive AI-generated podcast improved learning and user experience compared to a version without prompts. Thirty-six undergraduates participated, and while learning outcomes were similar across conditions, reflection prompts reduced perceived attractiveness, highlighting a call for more research on reflective interactivity design.

1 Introduction

What if educational content could not only speak to learners, but listen, adapt, interact, and assess learning processes – like *reflection* – in real-time? As learners increasingly disengage from traditional materials like textbooks (Baron and Mangen, 2021), large language models (LLMs) offer new opportunities to deliver content in more engaging, interactive, and personalized formats, such as AI-generated podcasts (Jin et al., 2025). Emerging tools like NotebookLM[1] illustrate growing public interest in generative AI for learning.

Personalized learning with AI has been shown to support self-regulated learning by encouraging learners to plan, monitor, and evaluate their progress (Shemshack and Spector, 2020; Molenaar et al., 2023). Prior work demonstrates that personalized AI-generated podcasts based on college textbooks (tailored to learners' majors, interests, and instructional preferences) can enhance learning and enjoyment compared to both textbooks and non-personalized content (Do et al., 2025). However, most AI-generated podcasts remain *passive*: learners can ask questions, but the system does not initiate interaction or assess learning to guide deeper engagement. This represents a missed opportunity, as structured interactivity has been shown to enhance engagement and active learning in other domains (Laban et al., 2022).

More importantly, reflection is a critical component of learning – it helps learners draw meaningful and construct understanding in connection with learning goals. Embedding structured, reflection prompts into AI-generated podcasts could enhance engagement and learning, but may also disrupt learners' concentration and flow, possibly reducing their effectiveness. Design trade-offs remain unclear: when should reflection be prompted, and how can learners' responses be assessed in real-time with AI-generated podcasts?

We investigated these questions in a controlled experiment with a sample of 36 undergraduates, comparing two conditions: Reflection, where an AI-generated podcast periodically prompted learners to reflect and responded based on their input, and Standard, where no reflection prompts were prompted by the system. This study specifically investigates an AI-generated podcast featuring a single host, using two research questions: (1) Do interactive—in this case, meaning a model that can be freely interrupted, conversed with and asked questions—reflection prompts improve learning outcomes when incorporated into AI-generated podcasts compared to standard AI-generated podcasts? and, (2) Do interactive reflection prompts improve user experience when incorporated into AI-generated podcasts compared to standard AI-generated podcasts?

2 Related Work

Reflection is a key self-regulatory process that supports deeper learning and metacognitive awareness by promoting learners to contemplate their understanding and connect it with previous learning experiences. McAlpine et al. (1999) conceptualize reflection as a goal-driven process in which learners continuously integrate knowledge and action.

[1]https://notebooklm.google/

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 99–106
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

Building on this, recent work has explored whether digital learning environments can scaffold reflection to improve learning outcomes. (Cloude et al., 2021) examined the impact of reflective prompts in a game-based learning environment with 120 adolescents. Learners received one of three types of reflection prompts during learning: progress planning, solution strategy, and different problem approaches. Findings showed that the quantity and quality of reflections influenced learning, but their effects varied depending on the learner's goals.

Carpenter et al. (2021) further investigated reflection quality in middle-school students, using a rubric to assess written responses on a 5-point scale ranging from non-reflective to highly reflective. Higher-quality reflections – those including hypotheses, planning, and reasoning – were more predictive of learning gains. However, these reflections were scored post hoc, underscoring a key limitation in current research: the inability to evaluate and respond to reflection *in real time*. Cloude et al. (2021) highlight the lack of theoretical clarity around when and how to prompt reflection during learning, a gap this work aims to address by embedding structured, interactive prompts into AI-generated podcasts. In our system, an LLM-driven agent prompts learners to reflect and evaluates their spoken responses to guide real-time support.

While prior studies have explored personalized AI-generated podcasts for education, they have not addressed reflection. Do et al. (2025) compared generalized and personalized podcasts – generated from textbook chapters using LLMs – to traditional textbook reading on learning outcomes. Personalized podcasts, tailored to learners' majors and interests, improved enjoyment and learning outcomes. Their systems used a multi-stage generation pipeline with Gemini 1.5 Pro to convert textbook content into conversational podcast scripts, which were then synthesized using text-to-speech models.

Other work has explored AI-generated podcasts outside of education. Yahagi et al. (2025) showed that transforming academic papers into AI-generated podcasts lowered barriers to engaging with academic literature. Similarly, Laban et al. (2022) examined AI-generated podcasts for news delivery and found that it enhanced enjoyment. However, unlike our system, these podcasts lacked interactive, reflective components and were not designed for structured learning contexts. Together, these lines of research inform our approach: we integrate structured reflection prompts – adapted in real-time by an LLM – into AI-generated podcasts to support engagement, reflection, and learning during listening.

3 Interactive Podcast Architecture

The technical implementation of our AI-generated podcast system involves ingesting textbook content and delivering an interactive podcast on demand. The system supports two modes of interaction (Figure 1):

- *Standard*: The system delivers audio content continuously from the textbook and allows learners to interrupt at any time with questions or comments. This interaction style is similar to existing consumer podcast systems, such as NotebookLM's Interactive mode.

- *Reflection*: The system delivers audio content and incorporates structured reflection prompts, periodically pausing after key concepts are introduced, and requiring the learner to demonstrate understanding of the content before continuing in their spoken reflection.

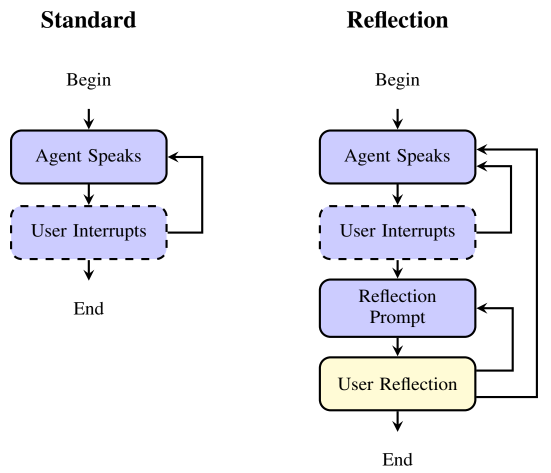


Figure 1: Standard vs Reflection Interaction Modes.

The system architecture is shown in Figure 2. The system consists of a Python backend for content generation, which hosts a LiveKit[2] room and creates an agent for speech synthesis. LiveKit is a platform for building AI-voice applications that can interact with users over the web.

The frontend is built with React, Next.js, and TailwindCSS, and connects to the backend to enable real-time communication with the system.

[2] https://livekit.io/

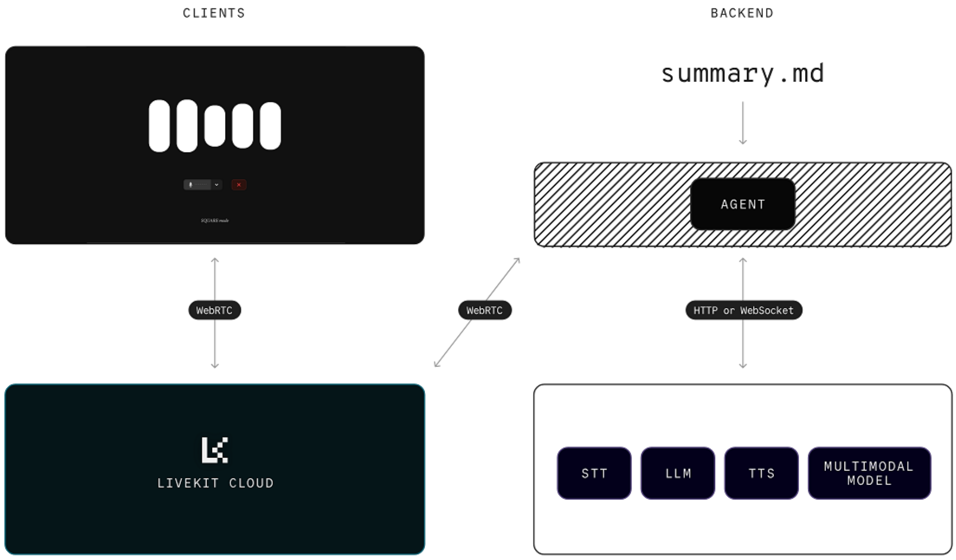


Figure 2: A diagram of our system, adapted from LiveKit's Agents Overview.

Based on these components, our system achieves an average response latency of 300ms (d'Sa, 2024), closely matching the pace of natural human conversation (Stivers et al., 2009). This low-latency architecture, which we further detail below, enables controlled comparisons between the two interaction paradigms. To support reproducibility, we have open-sourced the system; the code repository is available on GitHub[3].

3.1 Structured Summary

The system first ingests Chapter 1, Section 1.1 of OpenStax's textbook *Introduction to Philosophy* (Smith, 2022). Then, using GPT-4 Turbo, it converts academic text into a structural and summarized skeleton for the podcast, following the summarization procedure outlined by Laban et al. (2022). We found that using Laban et al.'s structured summarization approach to generate podcasts addressed several challenges, such as material omission. When generating podcasts directly from source text, we found that the model often omitted important material and produced outputs constrained by its context window, regardless of input length. This created an artificial ceiling on content length and limited scalability for longer educational materials. By using a structured summary, where each section corresponds to a paragraph from the original source, we were able to generate each segment independently, ensuring content coverage and improving quality, similar to skeleton-of-thought (Ning et al., 2023). The structured summary also improves interpretability, providing transparency into the generation process and facilitating easier debugging. It serves as a reference to track content coverage during the learner-facing conversation.

3.2 Podcast Generation

The structured summary is first divided into segments according to its outline and then processed by GPT-4o-mini. Each segment is used to generate corresponding portions of the podcast. Using OpenAI's GPT-4o text-to-speech (TTS) model with the *Alloy* voice, the podcast is synthesized as natural-sounding speech, incorporating appropriate pacing and intonation based on the skeleton structure.

3.3 User Interaction and Reflection

We serve the content to the learner differently depending on their current interaction context, tracked via state machine. In the Reflection mode, it monitors learner responses to assess their knowledge of the topic and prompt reflections by asking: "So, what is the most important thing you've learned so far?" at the end of each section, following similar prompts by (Cloude et al., 2021). After the learner responds to the reflection prompt, we use a one-shot evaluation to determine whether the response is suitable using a binary assessment (1 =

[3]https://github.com/DU-DIVALab/tutorflow

demonstrates understanding, 0 = does not demonstrate understanding). The learner's answer is considered satisfactory if the prompt "demonstrates awareness of their own knowledge" This is to ensure the learner cannot proceed with the audio session simply by restating a keyword the model used back to it, as previous work showed that domain-specific words in reflections were not indicative of the quality of reflection, but rather reflective depth (Cloude et al., 2021).

We employed in-context learning (Dong et al., 2022) using examples in the prompt to guide the agent's judgment. An example was if a learner is listening to content about Confucius, and they respond to a reflection prompt as "Confucius" to the model, this—while technically not incorrect, we guided learners to provide a more, detailed response to demonstrate synthesis of their knowledge to demonstrate a suitable reflection. For example, a response to a prompt with "Confucius' teachings would be considered patriarchal by modern standards" demonstrates a learner's understanding by combining a facet of what the content learned with how they contextualize the subject to present day.

It is important to note that the reflection prompts are different from quizzes. The learner does not need to mention everything they learned about the topic, only by demonstrating their ability to synthesize new understanding, the model deems their engagement and reflection on the material. The binary satisfactory/unsatisfactory classification acts as an elegant gate to guiding learner's progress in real-time to capture reflective depth, as opposed to relying on keyword matching, while avoiding the complexity of multi-dimensional rubrics. In the Standard mode, our system continuously listens for interruptions but otherwise continues speaking until one occurs, and does not prompt the learner to reflect. During learner interactions, the system uses Deepgram for learner speech transcription, and Silero VAD [4] to detect when learners were speaking. Additionally, a fine-tuned SmolLM v2 model (Allal et al., 2025) predicts speech boundaries to support smooth turn-taking.

3.4 Podcast Interface

As shown in Figure 2, the web application displays a decorative wave, an abstract animated visualization of the generated speech that animates based on the volume and cadence of the AI podcaster's voice. When the podcaster is silent, the wave appears as a flat line of dots. Learners begin the session with their microphone automatically turned on after granting permission through their browser.

[4] https://github.com/snakers4/silero-vad

4 Methods

4.1 Sample

To build on the methods used by Do et al. (2025), we designed our study as an extension of their work and used the same source material and measurements. This study was approved by an Institutional Review Board and a total of 36 (n=36; 42% female) college students enrolled at universities in the United States were recruited through the Prolific online marketplace. Participants were pre-screened for English fluency and minimal prior knowledge of the subject (Introductory Philosophy). We also screened participants for technical requirements to ensure they had a working microphone and speaker for audio. One participant was excluded from our analysis due to adversarial responses to reflection prompts (e.g., "I hate bots and I hate them in the work place [sic]"). Due to the added length introduced by the interaction in the Reflection condition, we limited the scope to a single textbook, Introduction to Philosophy (Chapter 1) (Smith, 2022), and focused only on one subsection (Chapter 1.1), to ensure a manageable session duration while maintaining consistency with the original study design.

4.2 Procedure

The study consisted of a single 40-minute remote session. Before the session, participants were randomly assigned to the 1) Reflection or 2) Standard condition. Next, participants completed a brief demographic survey and were informed they would interact with an AI-generated podcast to learn about philosophy, after which we collected informed consent. Participants were then directed to a web application (described in Section 3) and guided through an interactive AI-generated podcast lasting approximately 15 minutes. Upon completion, the agent provided a verbal code and displayed a popup in the browser, enabling participants to proceed. They were redirected to a survey, where they completed the learning outcomes test and the User Experience Questionnaire (UEQ). Participants were compensated with $10 USD after finishing the study.

4.3 Dependent Variables

4.3.1 Learning Outcomes

Learning was assessed using items from the post-chapter test bank from the OpenStax textbook[5]. The assessment items included seven multiple-choice questions from the Section 1.1 test bank. Due to the small number of multiple-choice questions for the single subsection, we included adapted three open-response items into multiple-choice questions, known as "Review Questions" (Smith, 2022), resulting in a total of 10 questions (see adapted items in Appendix B). Statistical analysis revealed no significant score differences between the original and supplemental items ($ps > .05$).

4.3.2 User Experience

User experience was measured using the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008) immediately after the audio session, specifically the *Attractiveness* and *Stimulation* subscales (see Appendix A). These subscales gauge the overall appeal and engagement of user's experience during the interaction, respectively. The UEQ employs a 7-point anchored Likert scale using adjective pairs such as "Annoying–Enjoyable" for *Attractiveness* and "Demotivating–Motivating" for *Stimulation*. We measured *Attractiveness* and *Stimulation* by averaging item scores within each subscale.

5 Results

Before conducting statistical analysis, we assessed whether our data adhered to a normal distribution using histograms and the D'Agostino-Pearson test, which suggested that our data were normally distributed across all variables ($K^2 = 2.56, p = .28$).

5.1 Research Question 1

To address our first research question, do interactive reflection prompts improve learning outcomes when incorporated into AI-generated podcasts compared to standard AI-generated podcasts, we calculated a two-sample t-test to compare whether there were differences in learning outcomes between Reflection and Standard conditions. The results suggested that there were no differences in learning outcomes between Reflection and Standard conditions, $t(34) = 0.89, p = 0.38, D = 0.29$.

[5]We do not provide the questions and answers for the knowledge retention questionnaires due to OpenStax policy. Verified educators from academic institutions may access test banks directly through OpenStax.

5.2 Research Question 2

To address our second research question, do interactive reflection prompts improve user experience when incorporated into AI-generated podcasts compared to standard AI-generated podcasts, we calculated 2 separate two-sample t-tests to compare whether there were differences in user experience subscales: attractiveness and simulation. The results suggested there was a significant difference in *Attractiveness*, $t(34) = 2.26, p = 0.03, D = 0.75$, where the Standard condition rated the experience more favorably than the Reflection condition (Figure 3). Conversely, there were no significant differences in *Stimulation*, $t(34) = 1.31, p = 0.20, D = 0.44$, between the Standard and Reflection conditions. Descriptive statistics are in Table 1.

Table 1: Dependent variable descriptive statistics by condition.

	Learning	Attractiveness	Stimulation
Reflection	5.89 (1.94)	26.22 (4.58)	21.22 (3.68)
Standard	6.50 (2.06)	29.56 (4.00)	23.17 (4.87)

Note. Means and (standard deviations) are provided.

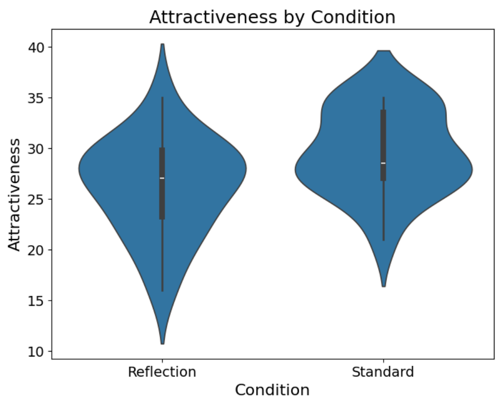


Figure 3: Attractiveness ratings across conditions.

6 Discussion

Personalized learning via interactions and reflection prompts are both recognized as valuable tools to enhance engagement and active learning (Sahronih et al., 2019) (Zhai et al., 2023). We implemented reflection prompts guided by (McAlpine et al., 1999)'s model of reflection and empirical literature suggesting that deeper reflection enhances learning, as supported by prior research that included no evaluation of responses in real-time (Cloude et al., 2021; Carpenter et al., 2021).

To build on this, our system applied a one-shot evaluation to judge learners' understanding based on their spoken reflections. Our study sought to explore whether integrating reflection prompts within an interactive, AI-generated educational podcast would improve user experience and learning outcomes compared to a standard, interactive AI-generated podcast.

Our first research question revealed that interactive, AI-generated podcasts with reflection prompts did not significantly improve learning outcomes compared to those without reflection prompts. This result contrasts with previous research, which found that reflection enhanced learning outcomes in game-based environments (Cloude et al., 2021; Carpenter et al., 2021). One explanation could be the nature of the prompt, which asked learners to recall factually relevant information rather than promoting reflection in relation to learning goals or planning, which encourages more thorough reflection and which is required by a game-based environment. Another possibility is that while reflection may be a useful tool for AI-generated podcasts, simply relying on the LLM to guide the reflection process without accounting for the learning goals or knowledge state of the learner may not be effective for promoting reflection with interactive AI-generated podcasts.

In our second research question, we found that the reflection prompts with an interactive, AI-generated podcast significantly reduced *Attractiveness* ratings compared to the standard AI-generated podcast condition. This indicated that the interactive elements in the Reflection condition may have disrupted the learners' flow and enjoyment during audio-based learning. This is different than previous research (Wang et al., 2025), which found that perceived interactivity and reflection boosted enjoyment and facilitated more active learning. The lack of significant differences for *Stimulation* suggests that the reflection intervention, despite its theoretical foundation in enhancing learning through reflective scaffolding (McAlpine et al., 1999), did not measurably improve user experience or learning outcomes in our study. Effective podcast-based reflections with LLMs likely require more detailed scaffolding, such as fine-tuning the model to provide automatic, tailored feedback that is based on learners' individual goals and current knowledge state. Future work should focus on developing stronger guidance methods to support reflection. This addresses a limitation identified by Do et al., who reported that participants desired "opportunities for active engagement" with AI-generated podcasts (2025), and suggests that while the specific implementation of reflection prompts may have detracted from the user experience, the general concept of interactive learning remains appealing to learners. The challenge appears to be finding the right balance between maintaining content flow and providing meaningful opportunities for reflection that effectively support learning.

6.1 Limitations

This study has important limitations to consider. First, our sample size of 36, which may limit the statistical power and generalizability of our results. Moreover, the focused scope of the content being taught (one section of a chapter) may not fully represent how reflection impacts learning across different subjects. Furthermore, our reflection responses were evaluated using a binary metric (understood/not understood) rather than evaluating the depth of reflection. This methodological constraint, though appropriate for our specific learning context, may have reduced the potential effectiveness of the reflection intervention compared to more elaborate implementations with different types of prompts. There are likely degrees to understanding which learning tools often fail to capture. Perhaps a human learner may feel more inclined to skip content they aren not understanding only to return to it later. Finally, the learner was exposed to the content for only 15 minutes, which may have reduced their learning and reflection due to the short nature of that task.

6.2 Future Work

Future research should explore alternative approaches for incorporating reflection and interaction in AI-generated podcasts. Developing adaptive reflection systems using LLMs that dynamically adjust based on learner engagement and metacognition would be a promising direction. Future work should investigate the use of LLMs for more fine-grained grading approaches for reflection quality, moving beyond binary assessments to evaluate responses with greater nuance. Additionally, investigating whether multi-modal data could better inform interaction and how to prompt reflections (e.g., eye movements, physiology, facial expressions, prior reflection quality, etc.) to enhance understanding.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Naomi Baron and Anne Mangen. 2021. Doing the reading: The decline of long long-form reading in higher education. *Poetics Today*, 42:253–279.

Dan Carpenter, Elizabeth Cloude, Jonathan Rowe, Roger Azevedo, and James Lester. 2021. Investigating student reflection during game-based learning in middle grades science. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 280–291, New York, NY, USA. Association for Computing Machinery.

Elizabeth Cloude, Dan Carpenter, Daryn A. Dever, James Lester, and Roger Azevedo. 2021. Game-based learning analytics for supporting adolescents' reflection. *Journal of Learning Analytics*, 8(2):51–72.

Tiffany D. Do, Usama Bin Shafqat, Elise Ling, and Nikhil Sarda. 2025. Paige: Examining learning outcomes and experiences with personalized ai-generated educational podcasts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–10.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Russ d'Sa. 2024. Openai and livekit partner to turn advanced voice into an api.

Fangzhou Jin, Chin-Hsi Lin, and Chun Lai. 2025. Modeling ai-assisted writing: How self-regulated learning influences writing outcomes. *Computers in Human Behavior*, 165:108538.

Philippe Laban, Elicia Ye, Srujay Korlakunta, John Canny, and Marti Hearst. 2022. Newspod: Automatic and interactive news podcasts. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 691–706, New York, NY, USA. Association for Computing Machinery.

Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*, pages 63–76. Springer.

Lynn McAlpine, Cynthia Weston, Catherine Beauchamp, C Wiseman, and Jacinthe Beauchamp. 1999. Building a metacognitive model of reflection. *Higher education*, 37:105–131.

Inge Molenaar, Susanne de Mooij, Roger Azevedo, Maria Bannert, Sanna Järvelä, and Dragan Gašević. 2023. Measuring self-regulated learning and the role of ai: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139:107540.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *Proceedings ENLSP-III*.

Siti Sahronih, Agung Purwanto, and M. Syarif Sumantri. 2019. The effect of interactive learning media on students' science learning outcomes. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, ICIET 2019, page 20–24, New York, NY, USA. Association for Computing Machinery.

Atikah Shemshack and Jonathan Michael Spector. 2020. A systematic literature review of personalized learning terms. *Smart Learning Environments*, 7(1):33.

Nathan Smith. 2022. *Introduction to Philosophy*. OpenStax, Houston, Texas.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Feifei Wang, Alan C. K. Cheung, Ching Sing Chai, and Jin Liu. 2025. Development and validation of the perceived interactivity of learner-ai interaction scale. *Education and Information Technologies*, 30(4):4607–4638.

Yuchi Yahagi, Rintaro Chujo, Yuga Harada, Changyo Han, Kohei Sugiyama, and Takeshi Naemura. 2025. Paperwave: Listening to research papers as conversational podcasts scripted by llm. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, page 10, New York, NY, USA. ACM.

Na Zhai, Yong Huang, Xiaomei Ma, and Jingchun Chen. 2023. Can reflective interventions improve students' academic achievement? a meta-analysis. *Thinking Skills and Creativity*, 49:101373.

A User Experience Questionnaire

All items were assessed on a 7-point scale, with the terms as anchors, adapted from (Laugwitz et al., 2008).

Attractiveness
Annoying – Enjoyable
Bad – Good
Unlikeable – Pleasing
Unpleasant – Pleasant
Unattractive – Attractive
Unfriendly – Friendly
Stimulation
Inferior – Valuable
Boring – Exciting
Not interesting – Interesting
Demotivating – Motivating

B Review Questions

We adapted three additional questions from the free-response items in (Smith, 2022) into multiple-choice format, alongside the chapter questions.

1. What characteristics are essential for being identified as a "sage"?
a) Upholding social norms and exercising political power
b) Seeking profound understanding through critical inquiry and providing foundational insights
c) Mastering persuasive rhetoric and accumulating significant wealth
d) Adhering to religious doctrines and conducting spiritual rituals
2. What does it mean for philosophy to "have an eye on the whole"?
a) Rejection of traditional narratives through empirical investigation
b) Fusion of mystical beliefs with systematic logical analysis
c) Skeptical inquiry into established wisdom and foundational explanations of reality
d) Emphasis on practical skills for societal and technological advancement
3. Which philosopher held that moral behavior and social harmony were linked to the natural order?
a) Confucius
b) Pythagoras
c) Thales
d) Yajnavalkya

Generative AI in the K–12 Formative Assessment Process: Enhancing Feedback in the Classroom

Michael Maksimchuk, EdD
Kent Intermediate School District, Grand Rapids, Michigan

Edward Roeber, PhD
Michigan Assessment Consortium, Lansing, Michigan

Davie Store, PhD
Kent Intermediate School District, Grand Rapids, Michigan

Abstract

This paper explores how generative AI can enhance formative assessment practices in K–12 education. It examines emerging tools, ethical considerations, and practical applications to support student learning, while emphasizing the continued importance of teacher judgment and balanced assessment systems.

1 Introduction

The rapid evolution of generative artificial intelligence (AI) tools, such as ChatGPT, Microsoft Copilot, and Perplexity AI, has catalyzed significant opportunities in education. While student adoption of these tools has grown swiftly, many educators remain inexperienced in their use (University of Illinois, Urbana-Champaign, 2024). This disparity underscores the urgency of examining how AI can responsibly enhance teaching and learning.

Formative assessment, understood as an ongoing process of gathering and using evidence to inform instruction, presents a promising domain for AI integration (Hopfenbeck et al., 2023). Persistent challenges—such as large class sizes, variability in teacher expertise, and limited time for individualized feedback—suggest that AI could serve as a valuable partner in extending teachers' capacity. At the same time, integrating AI raises issues of bias, equity, accessibility, and privacy.

2 Defining the Formative Assessment Process

Formative assessment is not a product or event but a planned, ongoing process in which teachers and students collaboratively use evidence of learning to improve understanding and guide instruction (Michigan Assessment Consortium, 2017; Renaissance, 2021). Distinct from summative assessment, which evaluates learning at the end of instruction, formative assessment occurs continuously during instruction, is low-stakes, and prioritizes descriptive feedback to support improvement (Michigan Assessment Consortium, 2017, 2018, 2024a).

Key elements include clarifying learning goals and success criteria, eliciting and analyzing evidence of student thinking, providing actionable feedback, engaging students in peer and self-assessment, and adjusting instruction based on emerging evidence (Michigan Assessment Consortium, 2021).

This process-orientation positions students as active agents of their own learning, co-constructing goals, monitoring progress, and making decisions about next steps.

3 Opportunities and Realities in Implementing the Formative Assessment Process (FAP)

Despite broad support in the literature (Black & Wiliam, 1998; Hattie & Timperley, 2007), several barriers exist in the widespread and effective use of formative assessment. These include:

Time and Workload: Providing high-quality, individualized feedback for large classes is often untenable (Gamlem & Vattoy, 2023). Teachers resort to general or delayed comments, undermining formative intent.

Variability in Teacher Assessment Literacy: Many educators lack adequate training in assessment design and data interpretation (Wylie & Lyon, 2015). Misunderstandings persist, with some equating formative assessment only to ungraded quizzes.

Equity and Contextual Barriers: In some settings, cultural norms, oversized classes, or

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 107–110
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

limited resources inhibit practices such as peer feedback and student-centered dialogue (Halai et al., 2023).

Sustainability: Designing rigorous tasks, interpreting evidence, and maintaining feedback cycles require expertise and planning time that teachers often lack (Schmoker, 2011).

Without adequate support, formative assessment struggles to scale beyond isolated classrooms. These challenges create fertile ground for AI assistance (Swiecki et al., 2022; Zhai & Nehm, 2023).

4 The Role of Generative AI in the FAP

AI can provide immediate, descriptive, and individualized feedback, increasing both timeliness and frequency (Maksimchuk & Pentón Herrera, 2025). Studies show AI feedback can align well with rubric criteria and reduce teacher burden, though human feedback remains superior in accuracy and tone (Steiss et al., 2024). AI works best in partnership with teachers—offering preliminary feedback that educators review and adapt.

Dialogic interaction is a unique advantage: students can query AI for clarification, examples, or alternative explanations, fostering self-regulation and deeper learning (Mahapatra, 2024). Yet concerns persist about accuracy, tone, and potential bias, underscoring the importance of a "human-in-the-loop" approach (Mollick, 2024).

4.1 AI as a Tool for Designing Prompts

Teachers can use AI to generate formative tasks, unpack standards, and create authentic prompts aligned with learning goals (Black & Wiliam, 1998). AI serves as a co-designer, producing first drafts of questions, rubrics, or feedback stems, which teachers refine. Tools like the Kent ISD "AI for Assessment" prompt library exemplify efforts to guide teachers in effectively harnessing AI (Maksimchuk, 2025). Importantly, AI can also flag potential cultural biases in assessment materials.

4.2 AI as a Student Partner in Reflection and Peer Feedback

AI can support student self-regulation by prompting metacognitive reflection and providing personalized explanations. It may also function as a "peer" in giving feedback or serve as material for critique—students assess AI-produced responses, sharpening their understanding of success criteria (Wang & Fan, 2025). Proper training is essential so students engage with AI as a learning aid rather than a shortcut.

4.3 AI for Teachers' Growth

Using AI requires teachers to articulate learning targets and success criteria clearly, reinforcing assessment literacy. Teachers can leverage AI for rubric creation, item analysis, or exploring alternative formative strategies, effectively turning the technology into embedded professional learning (Michigan Assessment Consortium, 2024a). Over time, AI can act as a coaching tool, offering guidance on question quality, instructional adjustments, and data interpretation.

5 Ethical and Equity Considerations

Integrating AI into assessment requires attention to fairness, accessibility, and privacy.

AI outputs may privilege dominant cultural or linguistic norms, disadvantaging English language learners or misinterpreting diverse perspectives (University of Illinois, Urbana-Champaign, 2024; University of Texas at Austin, 2025). Teachers must review outputs critically and guide students in recognizing potential bias.

Accessibility: AI must be inclusive for students with disabilities and multilingual learners, ensuring equitable participation.

Data Privacy: Compliance with FERPA and ethical data practices is essential. Student work and learning data must be safeguarded.

Equity Lens: The Michigan Assessment Consortium's Components of Equitable Assessment Systems (2024b) framework stresses centering equity in AI use. Educators should ensure AI augments, rather than undermines, fairness in feedback and instructional decisions.

6 Case Studies and Practical Applications

Several examples illustrate AI's formative potential:

- High School English: AI-generated feedback on student writing increased revision cycles and student engagement, though teacher review was still critical.
- Mathematics: Teachers co-designed assessments with AI, generating varied question types aligned with learning targets and identifying misconceptions.
- Science Inquiry: Students engaged AI as a partner in developing and refining hypotheses, receiving iterative feedback during investigations.

Across cases, AI supported timely feedback, diversified assessment strategies, and fostered greater student ownership of learning. Teachers emphasized the importance of guidance, critical evaluation, and contextual adaptation.

7 Recommendations

For School Leaders

- Provide professional development that pairs AI tool use with deepening assessment literacy.
- Ensure equitable access to AI-supported learning tools across all schools and communities.
- Establish clear ethical guidelines for AI use in classrooms.

For Teachers

- Use AI to supplement, not replace, human feedback and professional judgment.
- Involve students in critiquing AI feedback to foster critical thinking.
- Collaborate with colleagues to share effective prompts and strategies.

For Policymakers

- Incorporate AI literacy into educational standards.
- Fund research and pilot programs evaluating AI's impact on formative assessment and equity.
- Address infrastructure gaps so underserved schools can access AI resources.
- Adapt assessment and accountability policies to encourage responsible AI use in classrooms.

8 Conclusion

Generative AI offers a powerful means to strengthen formative assessment by making feedback more immediate, personal, and interactive; supporting teachers in prompt and rubric design; and building assessment literacy among educators. Yet, the promise of AI is balanced by risks related to bias, privacy, and equity.

The future lies in a human-driven, AI-augmented classroom where teachers retain responsibility for instructional judgment, empathy, and relational pedagogy, while AI expands opportunities for feedback, reflection, and differentiation. As the field moves forward, iterative, evidence-based implementation will ensure that AI in formative assessment fulfills its potential to inform and improve learning for every student.

References

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.

Gamlem, S. M., & Vattoy, K.-D. (2023). *Feedback and classroom practice*.

Halai, A., Sarungi, V., & Hopfenbeck, T. N. (2023). Teachers' perspectives and practice of assessment for learning in classrooms in Tanzania. In *International Encyclopedia of Education(Fourth Edition)* (pp. 63–72). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.09039-4

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). *Challenges and opportunities for classroom-based formative assessment and AI: a perspective article*. https://doi.org/10.3389/feduc.2023.1270700

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning*

Environments, *11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Maksimchuk, M. T. (2025, May 11). *AI for Assessment: Prompt Library*. AI For Assessment. https://docs.google.com/document/d/1hlbYg-lzqNjG1LfyUGwJg8ukx04fkdUFLscPpwLYfTE/edit?tab=t.0#heading=h.dptxlv92mndr

Maksimchuk, M. T., & Pentón Herrera, L. J. (2025, April 24). *The E5 Model of Equitable Assessment* [Keynote Address]. 5th International Conference: A Person in the Language Space: Historical Heritage, Problems and Development Prospects, Berdiansk State Pedagogical University and Kharkiv Humanitarian Pedagogical Academy.

Michigan Assessment Consortium. (2017, September). *What do we mean by Formative Assessment?* https://www.michiganassessmentconsortium.org/wp-content/uploads/LP_FORMATIVE-ASSESSMENT-1.pdf

Michigan Assessment Consortium. (2018, May). *What conditions are necessary for successful implementation of formative assessment?* https://www.michiganassessmentconsortium.org/wp-content/uploads/2018_May_Conditions_Necessary_for_Implementation_0.pdf

Michigan Assessment Consortium. (2021, February). *Formative assessment(s) or formative assessment? The "s" makes a difference.* https://www.michiganassessmentconsortium.org/wp-content/uploads/LP-FORMATIVE-ASSESSMENT-VS-ASSESSMENTS.pdf

Michigan Assessment Consortium. (2024a). *Assessment Literacy Standards: A National Imperative*. Michigan Department of Education. https://www.michigan.gov/mde/-/media/Project/Websites/mde/OEAA/Formative-Assessment-Process/MAC-Assessment-Literacy-Standards.pdf?rev=cf153fb987c74afe9ccd40e8b7f2344d&hash=55B879EB948EE10EED727D53803E2DC7

Michigan Assessment Consortium. (2024b, June). *Components of an Equitable Assessment System*. https://www.michiganassessmentconsortium.org/wp-content/uploads/MAC_CEAS_Brief.pdf

Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.

Renaissance. (2021, December 9). Formative assessment: What is it and why use it? *Renaissance*. https://www.renaissance.com/2021/12/09/blog-formative-assessment-what-is-it-and-why-use-it/

Schmoker, M. J. (2011). *Focus: Elevating the essentials to radically improve student learning*. ASCD.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

University of Illinois, Urbana-Champagne. (2024, 24). *AI in Schools: Pros and Cons*. College of Education. https://education.illinois.edu/about/news-events/news/2024/10/24/ai-in-schools--pros-and-cons

University of Texas at Austin. (2025). *Generative AI in Teaching and Learning: Biases and Risks | Center for Teaching & Learning*. Center for Teaching and Learning. https://ctl.utexas.edu/generative-ai-teaching-and-learning-biases-and-risks

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 1–21. https://doi.org/10.1057/s41599-025-04787-y

Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 140–160. https://doi.org/10.1080/0969594X.2014.990416

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390–1398. https://doi.org/10.1002/tea.21885

Using Large Language Models to Analyze Students' Collaborative Argumentation in Classroom Discussions

Nhat Tran and Diane Litman and Amanda Godley
University of Pittsburgh
Pittsburgh, PA, USA
{nlt26, dlitman, agodley}@pitt.edu

Abstract

Collaborative argumentation enables students to build disciplinary knowledge and to think in disciplinary ways. We use Large Language Models (LLMs) to improve existing methods for collaboration classification and argument identification. Results suggest that LLMs are effective for both tasks and should be considered as a strong baseline for future research.

1 Introduction

Collaborative argumentation is a key mechanism through which students engage in critical thinking and co-construct knowledge during classroom discussions (Larson, 2000; Reznitskaya and Gregory, 2013). Because well-facilitated discussions are a cornerstone of effective instruction, they are frequently a target of measurement (Matsumura et al., 2008; Hill et al., 2008; Reznitskaya and Wilkinson, 2021; Bouton and Asterhan, 2023). However, large-scale human evaluation is costly and challenging due to the complexity of collaboration and argumentation in multi-party dialogue. Thus, AI methods – across a range of measurement frameworks – are being developed to automatically assess classroom dialogue quality (Wang and Demszky, 2023; Xu et al., 2024; Kelly et al., 2018), and to develop tools for improving dialogic teaching aimed at teachers (Lugini et al., 2020; Suresh et al., 2021), coaches (Wang and Demszky, 2023), and learning scientists (Tran et al., 2024b).

The tasks of computationally analyzing students' *collaboration* and *argumentation* in a classroom discussion are challenging (Olshefski et al., 2020; Lugini and Litman, 2020; Wang and Chen, 2024; Shiota and Shimada, 2022). For our dataset (example in Table 3 and details in Section 3), collaboration analysis involves classifying every student turn as relevant to collaborative argumentation (e.g., initiating a new idea or challenging another student's claim) or not (non-argumentative). Argumentation analysis can be further divided into a pipeline of two subtasks. The first involves identifying spans of text consisting of argument discourse units (ADUs), i.e., argument component detection (ACD). The next subtask, argument component classification (ACC), focuses on assigning a label (Claim, Evidence, Warrant) to each ADU [1].

While computational argument mining is an active research area (Stede and Schneider, 2019; Lawrence and Reed, 2020), relatively little work has been done on collaborative discussions. Also, prior work often omits ACD and takes already identified argument components as input, and thus focuses on only argument component classification (ACC) rather than on end-to-end argument mining (Deguchi and Yamaguchi, 2019; Tran and Litman, 2021). Finally, argument component classification is often treated as a sequence labeling task, but it needs extensive finetuning and offers limited control over the output, especially when capturing relationships between components (Schulz et al., 2019; Alhindi and Ghosh, 2021).

To address these challenges, we leverage Large Language Models (LLMs) for two key tasks in assessing collaborative argumentation in classroom discussions. LLMs offer strong generative capabilities, enabling effective classification and sequence labeling with minimal annotated data. For *collaboration classification*, we replace traditional classifiers with LLMs and compare multi-class versus binary prompting strategies. For end-to-end argumentation identification, we use LLMs to jointly segment and classify argument components. Our study aims to answer the following questions:

RQ_1 Is LLM effective for collaboration classification?

RQ_2 Can we use LLM to perform end-to-end argument identification, and how good is it?

[1] We use ADU and argument component interchangeably.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 111–125
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

Our contributions are two-fold. First, we show that few-shot prompting enables LLMs to outperform a BERT-based collaboration classifier trained on significantly more annotated data, with binary prompting proving more effective than multi-class classification. Second, we show that LLMs can perform end-to-end argument identification, with our structure-focused evaluation highlighting their effectiveness under a simplified argument scheme (i.e., at most one Claim, Evidence, or Warrant).

2 Related Work

Much of the prior work on argument mining addressed the problems of argument segmentation (i.e., identifying ADU boundaries), component classification, and relation identification modeled in a pipeline of subtasks (Potash et al., 2017; Niculae et al., 2017). However, many of them assume the availability of segmented argumentative units and do the subsequent tasks such as classification of argumentative component types (Lugini and Litman, 2018, 2020; Garcia-Gorrostieta et al., 2018), and argument relation identification (Ghosh et al., 2016; Gemechu et al., 2024; Contalbo et al., 2024). We perform argument component segmentation and classification simultaneously by utilizing LLMs.

Previous work on argument segmentation includes approaches that model the task at a surface level by classifying sentences as argumentative or non-argumentative (Ajjour et al., 2017; Chakrabarty et al., 2019). At a more fine-grained level, there are studies that use heuristics to identify argumentative segment boundaries (Wachsmuth et al., 2016). Prior work also treats the task as a sequence labeling task by performing token-level classification to directly identify the type of the argument component and achieves promising results (Schulz et al., 2019; Alhindi and Ghosh, 2021). Additionally, multi-task learning, which utilizes other NLP tasks such as part-of-speech tagging or datasets from other domains, is a widely used tool to further boost performance of argument component classification (Daxenberger et al., 2017; Schulz et al., 2018; Mensonides et al., 2019). Unlike these approaches, we do not formulate the argument identification task as a token-level sequence labeling task. Instead, we consider it a text generation task by leveraging LLMs, which have been shown to be effective at text span extraction (Tran et al., 2024a; Wang et al., 2025).

Since LLMs such as GPT-4 (OpenAI et al., 2024), Llama (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023) have outperformed pretrained language models (PLMs) such as BERT (Devlin et al., 2019) in many natural language processing (NLP) tasks, there has been growing interest in leveraging them for argument mining and text extraction. Kashefi et al. (2023) uses GPT-3 for claim and premise detection, but it is only a classification task on the sentence level. Chen et al. (2024b) explores the potential of LLMs in many argument computation tasks, but does not cover joint tasks such as end-to-end argument identification. Pichler et al. (2025) and Lin and Koedinger (2024) demonstrate that LLMs are effective in sequence labeling if they are prompted appropriately, but they do not test them in the context of argument mining. Our work leverages LLM for analyzing collaborative argumentation, focusing on 2 tasks: collaboration classification and argument identification.

3 Data

We use Discussion Tracker (DT)[2], publicly accessible classroom discussion data annotated for collaborative argumentation (Olshefski et al., 2020), for our experiments. The DT data comprises 90 transcribed multi-party discussions conducted in American high school English Language Arts classes. We use two subsets from the corpus. They were collected in 2019 (29 transcripts) and 2022 (61 transcripts) using the same annotation guidelines, so we refer to them as DT_19 and DT_22, respectively.

We use the data for two tasks: *collaboration* code classification and *argumentation* identification. Students' talk at the turn level was annotated for *collaboration*, and talk at the argument discourse unit (ADU) level was annotated for *argumentation*. Specifically, argumentative turns were annotated with one of four collaboration codes: *New Idea*, *Agreement*, *Extension*, and *Challenge*; turns that contained no substantive argumentation were labeled with the collaboration code *None*. Argumentative turns were further segmented into argument discourse units (ADUs), which were labeled for argument types: *Claim*, *Evidence*, or *Warrant*. Annotators were instructed not to segment turns into multiple claims or multiple units of evidence, and every word belongs to one ADU (i.e., no gaps between ADUs). As a result, each segmented ADU is considered an argument component.

Definitions of collaboration and argumentation

[2]https://discussiontracker.cs.pitt.edu

coding are in Tables 1 and 2. Table 3 shows an annotated transcript, while statistics are in Table 4.

4 Method

We use few-shot prompting to instruct a LLM to tackle the tasks, using the prompts in Tables 5 and 6. The few-shot examples are not from the test set; they are either from the training set (cross-validation) or from a different DT corpus (e.g., using examples from DT_19 to test on DT_22).

4.1 Collaboration Classification

The collaboration task involves classifying a student's turn into 1 of 5 classes: Non-Argumentative (None), New Idea, Agreement, Extension, Challenge. We utilize LLMs in two approaches.

LLM-multi. We treat the task as standard multi-class classification. Specifically, we ask the LLM which of the 5 classes it thinks the turn belongs to. The prompt includes the instruction, definitions of the 5 classes, and 10 few-shot examples. Each few-shot example consists of a turn and its correct class. We have 2 examples for each of the 5 classes.

LLM-binary. Although not specifically focused on collaboration classification, prior work has shown that utilizing LLM is more effective at binary classification compared to multi-class classification on classroom discussion data (Tran et al., 2024b). Thus, we perform 4 binary classification tasks for each student's turn. For an argumentative class X (New Idea, Agreement, Extension, and Challenge), we ask the LLM a yes/no question about whether the turn is considered X by providing it with X's definition. We call the set of the remaining argumentative classes except X as S. For instance, if X is New Idea, S = {Agreement, Extension, Challenge}. For few-shot examples, we provide 5 examples where a turn should be predicted as X (positive examples) and 5 examples where it should not be (negative examples). In the 5 negative examples, we use 1 example where the turn's gold-standard class is s_i for all $s_i \in S$ and 2 examples where the turn's class is None. For the final turn-level prediction, if the LLM predicts 'no' for all of the 4 argumentative classes, it is a non-argumentative turn (None). If there is more than one class predicted as 'yes', we select one with the highest probability, p(yes|X).

4.2 Argumentation Identification

This task is typically approached as a two-step pipeline applied to argumentative student turns. The first step, argument component detection (ACD), involves identifying spans of text that constitute argument discourse units (ADUs). The second step, argument component classification (ACC), assigns a label (Claim, Evidence, or Warrant) to each identified ADU. One way to solve two subtasks simultaneously is to treat them as a sequence labeling task using the BIO scheme (Beginning, Inside, or Outside) (Schulz et al., 2019; Alhindi and Ghosh, 2021). Specifically, instead of segmenting the text into ADUs first, we can conduct a token-level[3] classification task to identify the type of the argument component (e.g., B/I tokens from claim, evidence, and warrant) directly by joining the first and the second sub-tasks in a single task (i.e., B-Claim, I-Claim, B-Evidence, ...). See Figure 1 for an illustration of the BIO conversion. However, since LLMs are potent tools for following human instructions, prior work utilizing LLMs for sequence labeling employs generative approaches instead of performing the traditional token-level classification task (Lin and Koedinger, 2024; Wang et al., 2025). Also, due to the nature of the dataset, every word in an argumentative turn belongs to either Claim, Evidence, or Warrant (i.e., no O labels are present). Thus, we treat the task as a text generation task for the LLMs to perform both ACD and ACC tasks simultaneously.

LLM-auto. We let the LLM extract non-overlapping text spans of the target turn into C, E, and W. Because 95% of turns had a collaborative relationship with turns within the previous four turns (Olshefski et al., 2020), we provide four previous turns for the dialogue context, along with the definitions of C, E, and W for reference. The output is formatted as Claim: {claim_span}, Evidence: {evidence_span}, Warrant: {warrant_span}. Because an argumentative turn does not necessarily consist of all three segments (e.g., only C and E), the output text spans can be empty. We also ensure that all segmentation scenarios are covered in the few-shot examples by including at least one example for each class combination. Specifically, if we consider a scenario as a combination of C, E, and W, along with their order of appearance in the text from left to right, there are 10 scenarios in the dataset: (C), (E), (C, E), (E, C), (C, W), (E, W), (C, E, W), (C, W, E), (E, C, W), and (E, W, C). We provide one example for each of the scenarios.

LLM-refine. Previous studies show that (i)

[3]We use words as tokens.

LLM is more effective with more detailed instructions (Tran et al., 2024b) and (ii) LLM is good at judging LLM's generated answers (Chen et al., 2024a; Huang et al., 2025). We assume that LLM is better at the task when the correct combination of C, E, and W is provided. In other words, if the LLM knows that the turn only contains C and E, it provides a better segmentation than it does without that information. First, we use LLM to generate multiple argument segmentations (**LLM-gen**) given the combinations of C, E, and W (e.g., segment the text into Claim and Evidence). We ignore the ordering of C, E, and W in the combination, but instead provide different orderings in the few-shot examples. Therefore, each turn will be segmented into one of the following six combinations: (C), (E), (C, E), (C, W), (E, W), (C, E, W). Since (C) and (E) simply require marking the entire turn as C or E, we do not need LLM to do so. As a result, each turn will be segmented into four different ways by the LLM. The prompt for the first step consists of four previous turns as the dialogue context, the definitions of C, E, and W, and a specific argument combination we want to split the text into. The second step, the refinement step (**LLM-judge**), involves selecting the most suitable segmentation from the six generated options. To do so, we consult another LLM to select the best segmentation from the six options.

LLM-acc. Since prior work often only focused on the argument component classification (ACC) task (Lugini and Litman, 2020; Kashefi et al., 2023; Garcia-Gorrostieta et al., 2018; Hidayaturrahman et al., 2021) and assumed that the correct segmentation is given, we additionally conduct an experiment on using LLM specifically for argument component classification. For an ADU, we prompt the LLM to classify it as C, E, or W. Similar to other LLM approaches, we provide the 4-turn dialogue context, definitions of C, E, and W, along with 9 few-shot examples (3 of each type C, E, and W).

5 Experimental Setup

5.1 Baseline Models

Collaboration. We train a **BERT** model to predict whether a turn is either a New Idea, Agreement, Extension, Challenge, or Non-argumentative.

Argumentation. For the *argument component classification* task in which the correct argument component segmentation is provided, we compare our LLM's results (**LLM-acc**) with results from a BERT-based model utilizing local context and speaker context from Lugini and Litman (2020) (**BERT-context**). For the downstream *argument identification task* (argument segmentation + classification), we follow prior work and use BERT for sequence labeling as a baseline (Schulz et al., 2018; Kashefi et al., 2023). We call it **BERT-BIO**, which employs a BIO classification scheme to identify and classify argument components. We use BERT as the base transformer model and train a token-level classifier head on top. This baseline aims to label each token as B-Claim, I-Claim, B-Evidence, I-Evidence, B-Warrant, or I-Warrant.

We note that the BERT-context's results are from a publication using an older version of the DT_19 data (Lugini and Litman, 2020), which is no longer available. Our DT_19 version, which is corrected for better consistency, has 10 more ADUs (3145 versus 3135) compared to their version. However, because the difference is small, we still use the previously published BERT-context results to compare with our models' performance on the DT_19 data.

All BERT models are `bert-base-uncased`[4].

5.2 Experiment and Evaluation

We compare the performance of LLM and baseline approaches to answer the two research questions. mentioned in Section 1.

For collaboration prediction and argument component classification, we use the F_1 score as our evaluation metric since it is a standard multi-class classification task. We also report results in predicting Argumentative and Non-Argumentative turns.

For argument identification (segmentation + classification), due to our limited resources, we only conduct experiments on the larger corpus DT_22. After converting LLM's outputs to the word-level BIO format (see Figure 1 for an example), we can treat the task as a word-level classification task and compute the weighted F_1 score. We decided to use weighted F_1 because finding the exact boundaries of each segment is not essential empirically.

We also propose a new metric for argument identification on the component level. In a real-world application of an automated argument identification system (e.g., creating teacher dashboard analytics such as how many student claims were supported by evidence (Lugini et al., 2020)), it is more crucial to capture the structure of argument components within a single turn than to find the exact

[4]https://huggingface.co/google-bert/bert-base-uncased

splits. This is applicable to our data, as there are at most three argument components that cover every word in a turn. The word-level F_1 score does not consider component-level matching, whereas metrics like seqeval (Nakayama, 2018), which are popular for sequence labeling tasks such as named entity recognition, only consider strict matching between boundaries. We want to know whether the automated segmentation and classification have the same argument components, while not too strict in finding the boundaries between them (e.g., it is fine to have the two last words from Evidence identified as part of Warrant). To do so, we modify the metric from SemEval-2013 (Segura-Bedmar et al., 2013). Given a threshold K, a true positive (TP) is counted when the predicted span (*pred_span*) has the same label as the gold-standard span (*gold_span*) and they overlap at least $K\%$. The overlapping is calculated as $\frac{|\text{pred\_span} \cap \text{gold\_span}|}{max(|\text{pred\_span}|, |\text{gold\_span}|)}$, where $|\cdot|$ denotes the number of words in a span. Then, we can calculate Precision, Recall and F_1 normally.

The value of K controls how strictly we want the spans to match. At K = 100, we require an exact match between the two spans (i.e., same boundaries and same label) for a TP. At K = 0, we only compare predicted labels (C, E, W) with the gold-standard ones for a given turn. For example, if we predict a turn has one C and one E, as long as the gold-standard consists of exactly one C and one E, it is a correct prediction. We call this new metric Argument Component Score at K (**ACS@K**).

All experiments, including the baselines, are conducted using the same 10-fold cross-validation split provided by the DT corpus. Due to our limited resources, we utilize LLama3-8B (Grattafiori et al., 2024) as our LLM for all tasks [5].

6 Results and Discussion

6.1 Collaboration Results (RQ_1)

Table 7 shows the macro-F_1 over 10-fold cross-validation for the collaboration prediction task on both DT_19 and DT_22. Both LLM approaches significantly outperform the BERT baseline on both datasets. Additionally, LLM-Binary is significantly better than LLM-multi in all categories ($p < 0.05$), suggesting that using multiple LLMs as binary classifiers is an effective approach (Tran et al., 2024b). On the other hand, LLM approaches are not significantly better than BERT in classifying Argumentative and Non-Argumentative turns (except for LLM-binary in DT_22). It implies that the BERT model is not inferior in identifying argumentative turns, but struggles to predict the correct labels among the four collaboration codes. Using Cohen's kappa as the metric (Table 8), we get similar observations as the two LLM approaches constantly outperform BERT, and LLM-binary consistently achieves the best performance.

Looking into Table 9, the higher weighted F_1 scores compared to macro F_1 (Table 7) indicate that the models perform better on more frequent classes. We observe that the LLM approaches significantly outperform BERT in New Idea, Extension, and Challenge. Among these three classes, New Idea and Challenge are consistently the bottom 2 for all models. We also witness opposite cases for the two minority classes that take up less than 10 % of the data on both datasets, Agreement and Challenge. For Agreement, while LLM-binary is superior compared to BERT, BERT is not significantly worse than LLM-multi, and it even surpasses LLM-multi on DT_22. We hypothesize there are lexical clues (e.g., "I agree ...") for Agreement, and the increase in training data for Agreement in DT_22 (177 versus 38 instances) helps BERT learn to recognize the pattern of this type of collaboration. On the other hand, both BERT and LLM approaches struggle with Challenge, suggesting that the difficulty does not come from the scarcity of the class (i.e., LLM models need no training data). For Extension, while LLM-multi and LLM-binary's results suggest that it is easier than Agreement, BERT finds the opposite, and the largest performance gap between BERT and LLM approaches also falls in this category. This implies that BERT is not as effective in distilling knowledge to identify Extension after training as an LLM with few-shot prompting.

6.2 Argumentation Results (RQ_2)

For Argument Component Classification (ACC) on DT_19 [6], BERT-context (Lugini and Litman, 2020) and LLM-acc achieve 77.4 and 80.2 macro-F_1 scores, respectively. This suggests that LLM is not particularly better than BERT in classifying C, E, and W when the correct ADUs are provided.

However, when we have to perform the full Argument Identification task from scratch, which includes ACD (segmentation) and ACC (classification), we observe some performance gaps between LLM and the BERT-BIO baseline. In terms of

[5] https://huggingface.co/meta-llama/Llama-3.1-8B

[6] The two DT_19 corpora are slightly different as mentioned at the end of Section 5.1.

token-level F_1 score (Table 10), both LLM approaches outperform BERT-BIO, suggesting that utilizing the generative capability of LLM has advantages over sequence labeling with a transformer like BERT. On one hand, it is not feasible to control the tagging process of BERT-BIO at inference time. As a result, there are cases in which it provides more than one Claim, Evidence, or Warrant for a turn, which violates the nature of the DT corpus used for testing. On the other hand, we can restrict the output of LLM approaches by giving the instructions in the prompts and few-shot examples, which prevents them from violating the aforementioned data constraint. Thus, this can be one reason for the inferior performance of BERT-BIO in terms of word-level F_1 scores.

The score of the Beginning of a segment (B-C/E/W) is always lower than the Inside counterparts (I-C/E/W), which implies that it is hard to find the exact segmentation boundaries. However, B-E has higher results compared to B-C and B-W, demonstrating that the models are more effective at finding the beginning of Evidence. We hypothesize that certain words (e.g., 'because') can signal the start of evidence, making it easier to detect when students begin providing it. Among the C, E, and W, W appears to be the most challenging class to correctly identify, as the results of B-W and I-W are lower than those of the other two. Furthermore, LLM-refine significantly outperforms LLM-auto in average weighted F_1 ($p = 0.03$), suggesting that LLM is good at judging argument identification.

Figure 2 presents the proposed metric ACS@K with various K. Similar to the average weighted F_1 score (Table 10), LLM-refine beats LLM-auto and BERT-BIO. While the results of LLM-auto and LLM-refine are quite close, the BERT-BIO baseline yields noticeably lower performance. The discrepancies between BERT-BIO and the two LLM models are also larger compared to Table 10. In other words, LLM approaches are even more effective when evaluated on the argument component level. When the argument is simplified (i.e., only one C, E, and W), lacking control over the output by treating the task as a sequence labeling task (BERT-BIO) makes the argument identification results less desirable. In addition, LLM approaches are more robust when the threshold K is varied. We observe most increases in ACS@K score for LLM approaches until about K = 40, after which the curve remains more stable. Based on that observation, we hypothesize that LLMs might not be effective at finding exact segmentations, but are good at identifying argument components in the correct order. For example, assume the gold-standard labels for the turn from left to right are C, E, and W. If the model predicts a different order (e.g., E, C, W), it is considered correct when K = 0. As we increase K, that answer becomes incorrect because the overlaps between text spans do not satisfy the increased threshold. However, the graph shows that there are no big differences between K = 40 and K = 0 for the LLM approaches. This implies that the models get the argument components in the correct order. Lowering K after 60 does not show noticeably higher ACS@K scores, which further implies that the predicted argument components already have good overlap with the gold standard.

7 Conclusion

In this work, we experimented with LLMs in two classroom discussion assessment tasks: turn-level collaboration classification and end-to-end argument identification. The results show that LLMs outperform the BERT baselines in both tasks. For collaboration classification, we observe that different ways of formulating the task (binary versus multi-class classification) have an impact on performance, as the former yields better results. For argument identification, instead of dividing the task into two individual subtasks of ACD and ACC, we utilize LLMs to perform text generation to solve them simultaneously and achieve promising results.

Our results show that LLMs are robust under ACS@K, indicating they capture the correct order of argument components. Instruction following further allows finer control over argument constraints in LLMs, unlike sequence labeling with models like BERT. Future work includes fine-tuning LLMs, exploring diverse prompting strategies (e.g., Chain-of-Thought (Wei et al., 2022), example-retrieval (Wang et al., 2024), zero-shot methods), and applying these assessments downstream.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant # 1917673. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The authors would like to thank Yang Zhong and the anonymous reviewers for their valuable feedback on this work.

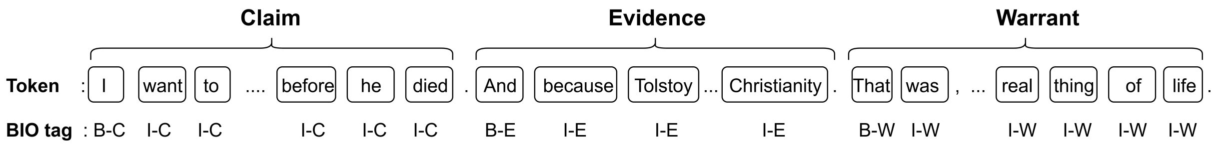


Figure 1: Conversion to BIO format. Each token is tagged as X-Y, where X is either B (Beginning) or I (Inside), and Y is either C (Claim), E (Evidence), or W (Warrant).

Code	Definition
New Idea	An initiating turn is the expression of a new idea in the discussion. This does not have to be a new topic, but should be a new idea, concept, or perspective. It usually does not reference ideas in prior turns at talk, or it does so only superficially. Turns that build on ideas in previous turns at talk are coded as "Extension". New student questions posed to the whole class that do not probe or question a previous answer are uncoded.
Extension	A turn is an extension if it builds off another student's ideas. Extension turns must extend one of the preceding four codeable student turns unless a turn prior to those 4 is specifically referenced. Extension turns include at least 2 key ideas or terms that were voiced by another student. Key ideas/terms may be textual, topical or conceptual terms. Textual terms may include characters and places from a text under discussion (like "Macbeth" or "Birnam Wood"), but do not include titles of texts. Topical terms may include disciplinary topics (like theme, metaphor, symbol, etc.). Conceptual terms may include abstract ideas (like "culture," "domination," "regret"). Extensions sometimes (but not always) include terms like "also, another, too"; or indicators of agreement/alignment (such as, "like X said…") Extensions can also include a self extension which is a turn of talk that adds information to or re-words one's own idea that was shared without acknowledging the idea of other speakers in close proximity.
Challenge	Challenge turns challenge or question a prior idea. Challenges should reference another student's turn in the preceding four codeable student talk turns. Challenges to points made further back are considered "New Ideas". A turn is considered a challenge if it includes both (1) keywords/concepts from previous turns (such as "culture," "domination," or "regretful") and (2) some indication of disagreement. Note that indications of disagreement can be very subtle (such as "still" or "actually" or "he did tell his sister") or more explicit (such as "I disagree", "No," "but," "however," "though"). A turn is considered a challenge if it challenges or requests more information, detail, elaboration, or clarification/explanation in the form of a question ("Why do you think that?" "You really think Macbeth wasn't crazy?" or "What do you mean?"). Will often include second person pronoun or direct address. Does not include procedural questions like "Wait what was his question?". Turns sometimes contain what may appear to be indications of disagreement (e.g., "however" "isn't") but are actually referring to ideas within the turn—these would likely fall under the category of "Extension".
Agreement	Turns that either express almost the exact thing in one of the preceding four coded student turns OR affirm the previous statement with a short response like "yeah" or "I agree with what she said.". When a turn seems like it should be coded as an Extension but lacks two clear key terms or ideas, it is likely to be coded as an Agreement.

Table 1: Definitions of the collaboration codes.

Code	Definition
Claim	An arguable statement that presents a particular interpretation of a text or topic. DOES: often (but not always) precedes evidence and warrants. States something that can more or less be contested—infers, predicts, hypothesizes, considers possibilities. DOES NOT: simply recount details from text that are accessible to all readers (everyone knows Macbeth became king)
Evidence	Talk used to support, justify, or back a claim. DOES: includes facts, textual references, anecdotes. Often (but not always) follows a claim. Always proximal to a claim (within 1 or 2 turns) . DOES NOT: does not exist without a claim.
Warrant	Move that provides explanation for why evidence supports the claim. DOES: Always proximal to evidence supporting a claim (almost always follows evidence). DOES NOT: It rarely occurs before claim/ evidence that it is explaining.

Table 2: Definitions of the argumentation codes.

Turn	Speaker	Talk	Collaboration	Argumentation
1	St 22	I think it's completely understandable, obviously because of what happened on his father's final day. But I feel like he doesn't deserve necessarily to feel guilty	New Idea	Claim
		because he was put through so much. Whenever you're in that situation, he's been worn down so much and everything has been taken from him. I feel like in that moment, he couldn't really think of anything he could do because he's already done so much, and so many people are telling him like, "There's nothing you can do."		Evidence
		I don't necessarily think he deserves to feel guilty, but I understand why he would.		Warrant
2	St 20	I agree with St 22. He shouldn't feel guilty because it's not his fault. But at the same time, you can't control how you feel. I guess, that's it.	Agreement	Claim
3	Teacher	When he asked himself about, did he pass the test about Rabbi Eliahu, do you guys think that he passed the test or he failed the test, in your opinion?		
4	St 3	I can almost say he passed the test, in a sense.	New Idea	Claim
		But you have to consider that whatever his father thinks [...]. He never wanted to lose his father. He always tried to help his father until the last moment. But then he was in shock. I feel like in general, he passed the test.		Evidence
5	St 6	Yeah	None	
6	St 1	Sorry, go ahead	None	
7	St 6	Okay. I think a big difference between the rabbi and his son, and Elie is that the rabbi's son acted on it and he deliberately did it. But Elie only had a subconscious thought about it and he never really intended on acting on it. He still gave his rations to him. He didn't take him away. He still felt bad. He tried to protect his father as best he could. He never really wanted him to die. It was more something he thought in the moment. Again, the cancer was getting to his head, too.	Challenge	Claim
		I think he passed his test. I don't think it's a big issue if you just thought about it for a second.		Warrant
		...		
11	St 1	Speaking on that note, someone mentioned talking about the "Free at last" part. The way I interpret it personally was that I thought that he felt his father was also free at last because he didn't have to deal with his suffering, which also shows that he did pass the test.	New Idea	Claim
12	St 13	Yeah. I also think whenever Elie talks about his father being a burden, it might not be he feels that his father coming around with him, brings him down,	Extension	Claim
		which I think it certainly does when he was thinking about that on the run. But I think that going back to Robbie's point, I think that it also could mean burden of his father's state and how his father is probably going to die is probably a burden on him mentally, as well as how his father is maybe making his chance to death.		Evidence

Table 3: A sample transcript with annotations for students' turns from DT_22 (T1.5.DT_2022.1.Night).

	Annotation	DT_19		DT_22	
		Count	Percentage	Count	Percentage
Collaboration	New Idea	802	24.59%	1585	20.87%
	Extension	1014	31.09%	2584	34.03%
	Agreement	38	1.17%	177	2.33%
	Challenge	271	8.31%	401	5.28%
	None	1136	34.84%	2847	37.49%
	Total	3261	100.00%	7594	100.00%
Argumentation	Claim	2054	65.31%	4724	62.99%
	Evidence	764	24.29%	1922	25.63%
	Warrant	327	10.40%	854	11.39%
	Total	3145	100.00%	7500	100.00%

Table 4: Descriptive statistics of the two corpora: DT_19 and DT_22.

Approach	Prompt
LLM-multi	Below are the definitions of 4 collaboration classes: New Idea, Extension, Challenge, and Agreement. # Definition of the 4 collaboration classes New Idea: {Definition of New Idea} Extension: {Definition of Extension} Challenge: {Definition of Challenge} Agreement: {Definition of Agreement} You are given a 5-turn conversation in a multi-party classroom discussion. Using the provided definition, your task is to classify the last turn into New Idea, Extension, Challenge, Agreement, or None if it does not belong to the four mentioned classes. # Example 1 {Example conversation 1} Output (New Idea, Extension, Challenge, Agreement, or None): {gold standard answer} ... # Example 10 {Example conversation 10} Output (New Idea, Extension, Challenge, Agreement, or None): {gold standard answer} # Your task {5-turn conversation} Output (New Idea, Extension, Challenge, Agreement, or None):
LLM-binary	Below are the definitions of 4 collaboration classes: New Idea, Extension, Challenge, and Agreement. # Definition of the 4 collaboration classes New Idea: {Definition of New Idea} Extension: {Definition of Extension} Challenge: {Definition of Challenge} Agreement: {Definition of Agreement} You are given a 5-turn conversation in a multi-party classroom discussion. Using the provided definitions, your task is to identify if the last turn is {One targeted class (New Idea, Extension, Challenge, or Agreement)}. Only answer yes or no. # Example 1 {Example conversation 1} Output (yes/no): {gold standard answer} ... # Example 10 {Example conversation 10} Output (yes/no): {gold standard answer} # Your task {5-turn conversation} Output (yes/no):

Table 5: Prompts used for collaboration classification. { } is a placeholder. Definitions of collaboration classes are from Table 1.

Approach	Prompt
All	Below are the definitions of 3 argumentation classes: Claim, Evidence, and Warrant. # Definition of the 3 argumentation classes Claim: {Definition of Claim} Evidence: {Definition of Evidence} Warrant: {Definition of Warrant}
LLM-auto	You are given a 5-turn conversation in a multi-party classroom discussion. Using the provided definitions, your task is to segment the last turn into one or more of the following argumentation components: Claim, Evidence, and Warrant. The segmentation must include at least one of these components, but it is not required to include all three. Every word in the last turn must belong to one category. Format your output as follows: Output Claim: {} Evidence: {} Warrant: {} # Example 1 (C) {Example conversation 1} Output Claim: {gold standard claim} Evidence: {gold standard evidence} Warrant: {gold standard warrant} ... # Example 10 (E, W, C) {Example conversation 10 with gold standard output} # Your task {5-turn conversation} Output
LLM-gen	You are given a 5-turn conversation in a multi-party classroom discussion. Using the provided definitions, your task is to segment the last turn into {one specific combination of Claim, Evidence, and Warrant}. Every word in the last turn must belong to one category. Format your output as follows: Output (Optional) Claim: {} (Optional) Evidence: {} (Optional) Warrant: {} # Example 1 {one specific combination of Claim, Evidence, and Warrant} {Example conversation 1} Output (Optional) Claim: {gold standard claim} (Optional) Evidence: {gold standard evidence} (Optional) Warrant: {gold standard warrant} ... # Example 10 {Example conversation 10 with gold standard output} # Your task {5-turn conversation} Output
LLM-judge	You are given a 5-turn conversation in a multi-party classroom discussion and different ways to segment the last turn to Claim, Evidence, and Warrant based on the provided definitions. Your task is to pick the most reasonable segmentation. Answer only one number between 1 and 6. Options: 1. {(C) segmentation} 2. {(E) segmentation} ... 6. {(C, E, W) segmentation} The best option is (a number between 1 and 6):
LLM-acc	You are given a 5-turn conversation in a multi-party classroom discussion. Using the provided definition, your task is to classify the last turn into Claim, Evidence, or Warrant. # Example 1 {Example conversation 1} Output (Claim, Evidence, or Warrant): {gold standard answer} ... # Example 10 {Example conversation 10 with gold standard answer} # Your task {5-turn conversation} Output (Claim, Evidence, or Warrant):

Table 6: Prompts used for argument identification. {} is a placeholder. Definitions of argumentation classes are from Table 2. All approaches share the first row to provide the definitions of the classes to the LLM.

Model	DT_19		DT_22	
	Arg vs Non-arg	All 5 labels	Arg vs Non-arg	All 5 labels
BERT	79.6	65.9	79.1	66.8
LLM-multi	80.1	69.1*	80.5	69.9*
LLM-binary	**84.1**	**73.7***	**86.1***	**73.5***

Table 7: Macro (unweighted) F_1 scores of the Collaboration classification task on the two DT corpora. Bold numbers highlight the best results. * means the number is statistically significant compared to its counterpart in the BERT baseline ($p < 0.05$) based on a Wilcoxon signed-rank test.

Model	DT_19	DT_22
BERT	62.3	62.8
LLM-multi	65.5*	68.2*
LLM-binary	**69.8***	**70.2***

Table 8: Cohen's kappa of the Collaboration classification task on the two DT corpora on all 5 labels. Bold numbers highlight the best results. * means the number is statistically significant compared to its counterpart in the BERT baseline ($p < 0.05$) based on a Wilcoxon signed-rank test.

Label	DT_19			DT_22		
	BERT	**LLM-multi**	**LLM-binary**	**BERT**	**LLM-multi**	**LLM-binary**
New Idea	58.2	61.3*	**65.8***	57.1	61.2*	**64.2***
Extension	67.3	74.7*	**79.1***	68.7	74.1*	**79.9***
Challenge	60.5	62.4*	**66.7***	57.3	60.7*	**65.1***
Agreement	70.1	71.5	**79.6***	73.0	72.3	**78.3***
None	73.4	75.6	**77.3***	78.1	**81.3***	80.1
Weighted F_1	67.3	71.3*	**75.1***	69.8	73.7*	**76.3***

Table 9: F_1 score for each collaboration class on DT_19 and DT_22 data. * means the number is statistically significant compared to its counterpart in the BERT model based on a Wilcoxon signed-rank test. Bold numbers highlight the best results for each label per dataset.

Model	B-C	I-C	B-E	I-E	B-W	I-W	Weighted F_1
BERT-BIO	61.5	73.2	68.3	75.7	60.6	69.3	68.6
LLM-auto	66.4	81.2	70.2	81.2	**64.3**	73.4	**71.4**
LLM-refine	**67.3**	**83.1**	**71.9**	**85.4**	62.3	**76.3**	**73.3**

Table 10: Per-label F_1 scores and average weighted F_1 scores of the argument identification task on DT_22. The labels are B/I-Arg, where B/I represents Beginning/Inside and Arg represents one of the three classes: Claim (C), Evidence (E), Warrant (W). Bold numbers show the best results for each label. All numbers are statistically significant compared to their counterparts in the BERT-BIO ($p < 0.05$), as determined by a Wilcoxon signed-rank test.

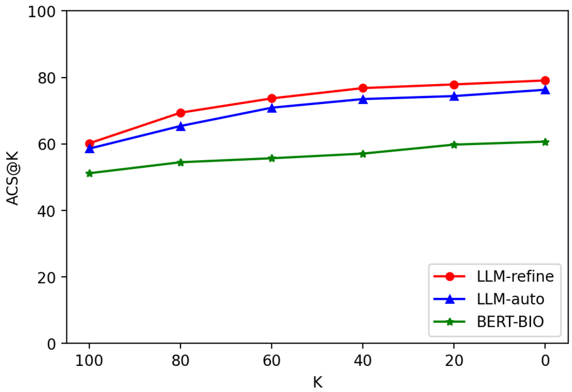


Figure 2: ACS@K with different values of threshold K.

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Tariq Alhindi and Debanjan Ghosh. 2021. "sharks are not the threat humans are": Argument component segmentation in school student essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–222, Online. Association for Computational Linguistics.

Edith Bouton and Christa SC Asterhan. 2023. In pursuit of a more unified method to measuring classroom dialogue: The dialogue elements to compound constructs approach. *Learning, Culture and Social Interaction*, 40:100717.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024b. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Michele Luca Contalbo, Francesco Guerra, and Matteo Paganelli. 2024. Argument relation classification through discourse markers and adversarial training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18949–18954, Miami, Florida, USA. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Mamoru Deguchi and Kazunori Yamaguchi. 2019. Argument component classification by relation identification by neural network and TextRank. In *Proceedings of the 6th Workshop on Argument Mining*, pages 83–91, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesús Miguel Garcia-Gorrostieta, Aurelio López-López, David Pinto, Vivek Kumar Singh, Aline Villavicencio, Philipp Mayr-Schlegel, and Efstathios Stamatatos. 2018. Argument component classification in academic writings. *J. Intell. Fuzzy Syst.*, 34(5):3037–3047.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hidayaturrahman, Emmanuel Dave, Derwin Suhartono, and Aniati Murni Arymurthy. 2021. Enhancing argumentation component classification using contextual language model. *Journal of Big Data*, 8(1):103.

Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. *Preprint*, arXiv:2403.02839.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Omid Kashefi, Sophia Chan, and Swapna Somasundaran. 2023. Argument detection in student essays under resource constraints. In *Proceedings of the 10th Workshop on Argument Mining*, pages 64–75, Singapore. Association for Computational Linguistics.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

Bruce E Larson. 2000. Classroom discussion: a method of instruction and a curriculum outcome. *Teaching and Teacher Education*, 16(5):661–677.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Jionghao Lin and Kenneth R. Koedinger. 2024. Haror: A system for highlighting and rephrasing open-ended responses. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 553–555, New York, NY, USA. Association for Computing Machinery.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium. Association for Computational Linguistics.

Luca Lugini and Diane Litman. 2020. Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Luca Lugini, Christopher Olshefski, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. Discussion tracker: Supporting teacher learning about students' collaborative argumentation in high school classrooms. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 53–58, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Lindsay Clare Matsumura, Helen E Garnier, Sharon Cadman Slater, and Melissa D Boston. 2008. Toward measuring instructional interactions "at-scale". *Educational Assessment*, 13(4):267–300.

Jean-Christophe Mensonides, Sébastien Harispe, Jacky Montmain, and Véronique Thireau. 2019. Automatic detection and classification of argument components using multi-task deep neural network. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 25–33, Trento, Italy. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The discussion tracker corpus of collaborative argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1033–1043, Marseille, France. European Language Resources Association.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Axel Pichler, Janis Pagel, and Nils Reiter. 2025. Evaluating LLM-prompting for sequence labeling tasks in computational literary studies. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics.

Alina Reznitskaya and Maughn Gregory. 2013. Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, 48(2):114–133.

Alina Reznitskaya and Ian A.G. Wilkinson. 2021. The argumentation rating tool: Assessing and supporting teacher facilitation and student argumentation during text-based discussions. *Teaching and Teacher Education*, 106:103464.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Claudia Schulz, Christian M. Meyer, and Iryna Gurevych. 2019. Challenges in the automatic analysis of students' diagnostic reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6974–6981.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tsukasa Shiota and Kazutaka Shimada. 2022. Annotation and multi-modal methods for quality assessment of multi-party discussion. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 175–182, Manila, Philippines. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2019. *Argumentation mining*. Springer.

Abhijit Suresh, Jennifer Jacobs, Charis Clevenger, Vivian Lai, Chenhao Tan, James H Martin, and Tamara Sumner. 2021. Using ai to promote equitable classroom discussions: The talkmoves application. In *International Conference on Artificial Intelligence in Education*, pages 344–348.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024a. Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 500–510, Atlanta, Georgia, USA. International Educational Data Mining Society.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024b. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2):304–335.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Deliang Wang and Gaowei Chen. 2024. On the interpretability of deep learning models for collaborative argumentation analysis in classrooms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 506–516, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4375–4389.

Evaluating Generative AI as a Mentor Resource: Bias and Implementation Challenges

Jimin Lee* and **Alena G. Esposito**
Department of Psychology, Clark University
Worcester, MA, USA
jmlee@clarku.edu aesposito@clarku.edu

Abstract

This study explored how students' perceptions of helpfulness and caring skew their ability to identify AI versus human mentorship responses. Emotionally resonant responses often lead to misattributions, indicating perceptual biases that shape mentorship judgments. The findings inform ethical, relational, and effective integration of AI in student support.

1 Introduction

Mentorship in higher education is widely recognized as a developmental relationship in which mentors offer academic, psychosocial, and emotional guidance to support students' success and growth (Nuis et al., 2023). Through sharing expertise and personal experience, mentors help students expand their knowledge base and pursue individual goals (Köbis and Mehner, 2021). As Generative Artificial Intelligence (GenAI) tools become increasingly integrated into academic settings, their role is expanding beyond academic support and research assistance to include potential contributions to mentoring relationships.

Upon entering college, students often encounter a combination of formal mentorship, typically through faculty advisors, and informal mentoring through peers or other institutional contacts (Rhodes et al., 2000; Jacobi, 1991). Understanding how these relationships form and function is critical to fostering positive and developmental outcomes. The rise of GenAI tools, such as ChatGPT (OpenAI, 2024a), prompts renewed reflection on how students engage with mentoring and what constitutes meaningful support in both human and machine-mediated contexts. Early evidence suggests that GenAI may function as a mentoring-like resource, offering students guidance and feedback that mimics the conversational tone of a human tutor (Le et al., 2025; Javaid et al., 2023). This insight highlights the need to examine how and why students may turn to GenAI for informal support and guidance.

GenAI tools can serve not only as tutors but also as supportive companions, helping reduce feelings of isolation and disconnection in academic environments (Farrelly and Baker, 2023). This growing interest in AI as a mentor-like resource is also shaped by broader concerns about burnout and mental health in higher education, which affect not only students but also faculty mentors who must balance teaching, research, and administrative demands (Hammoudi Halat et al., 2023). As institutions seek solutions to these overlapping pressures, GenAI presents both opportunities and challenges.

While GenAI facilitates academic learning by assisting with writing, problem-solving, and research tasks (Baidoo-Anu and Owusu Ansah, 2023; Le et al., 2025; Montenegro-Rueda et al., 2023; Schönberger, 2023), it still lacks the nuanced relational and developmental depth of human mentorship (Dempere et al., 2023). Ethical concerns and AI literacy are essential components of its responsible implementation, but so too is understanding students' lived perceptions of these tools. For GenAI to be effectively integrated into mentorship, educators and AI designers must understand how students evaluate its usefulness and trustworthiness. This factor is especially important in light of evidence that AI systems can unintentionally amplify human biases, especially in emotionally or socially sensitive domains, and that users may not always be aware of AI's influence on their perceptions and judgments (Glickman and Sharot, 2025).

Our prior work has explored these questions by examining how students interpret and engage with both AI-generated and human-authored responses in simulated mentorship scenarios. Drawing on the Perceptual Bias Activation (PBA) framework (Lee and Esposito, 2025b), we investigated whether stu-

*Corresponding author.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 126–133
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

dents' evaluations of response quality and accuracy of source identification were shaped by cognitive biases when the authorship of response sources differed across contexts, with the lowest accuracy in personal, mental-health-related scenarios. This finding may suggest that GenAI tools blend seamlessly into mentorship roles in mental health contexts but also raise concerns about overreliance on AI. Follow-up analyses in the personal domain further demonstrated that responses perceived as AI-authored were consistently rated as less helpful and caring, regardless of their actual source. However, when examined by actual authorship, AI-generated responses were rated as more caring than human responses. To further explore this discrepancy, we conducted an Inductive Content Analysis (ICA) of participants' open-ended explanations (Lee et al., 2025). The analysis revealed that source attributions were influenced by features such as tone, language, and perceived emotional depth, highlighting that students' interpretations were guided more by their perceptions and assumptions than by the intrinsic qualities of the response, which points to a lack of familiarity with GenAI tools for mental health support.

We also examined individual-level factors that might influence source accuracy and evaluation in all domains (Lee and Esposito, 2025a). Prior experience using GenAI was positively associated with more accurate source identification, suggesting that familiarity with GenAI tools may reduce perceptual bias. On the other hand, students' mentorship background (e.g., having a faculty mentor, peer mentor, or mental health counselor) did not predict improved source recognition. Using the Unified Theory of Acceptance and Use of Technology (UTAUT; (Venkatesh et al., 2003)), we found that students who rated GenAI responses as more useful, easier to use, and socially acceptable were more likely to evaluate them favorably, but only when they believed the response was AI-generated. These findings point to the need for greater transparency and intentional AI literacy efforts within higher education.

Our prior work reveals how perceptual biases can influence students' engagement with GenAI tools, often leading them to undervalue these resources, including in situations where the information is more readily available than from a human mentor. This raises two key questions: To what extent do perceptual biases limit the integration of Generative AI as a mentorship resource? And what factors, if any, mitigate this bias?

The current study seeks to address these two questions by reversing the analytical lens. Instead of examining how perceived or actual authorship affects evaluations, we ask: Are students more accurate in identifying the source of mentorship responses when they find those responses more helpful or caring? In other words, do positive evaluations enhance or cloud students' source discernment? We combine quantitative and qualitative analyses to explore this question. Specifically, we investigate whether students' ratings of helpfulness and caring predict their accuracy in identifying response sources, and how these patterns differ across personal, social, and academic mentorship contexts. We also analyze open-ended explanations from students to better understand the features that inform their judgments. This mixed-methods approach deepens our understanding of how perceptual biases shape students' interactions with human and AI mentorship. Furthermore, the findings of this study will have critical implications for the design and implementation of GenAI in higher education, particularly as institutions seek to balance technological innovation with relational and developmental support for students.

2 Methods

2.1 Participants

Our dataset stems from a larger project (Lee and Esposito, 2025a; Lee et al., 2025) that explored students' perceptions of GenAI and faculty mentorship in higher education. The study received approval from the college's Institutional Review Board (IRB Protocol #546). Although these data have previously been analyzed and published, the current study addresses new research questions and employs extended analytical approaches.

A total of 147 undergraduate students (M_{age} = 19.34 years, SD_{age} = 1.33 years, 105 female, 37 male, 2 non-binary, and 3 prefer not to answer) were recruited from a small liberal arts college in the northeastern United States. The sample was racially and ethnically diverse: 14. 97% Asian, 6. 80% Black, 67. 35% White, and 10. 88% Hispanic. All participants were at least 18 years old and provided informed consent prior to participation.

2.2 Procedure

The secure Qualtrics survey, which took approximately 30 minutes to complete, began with de-

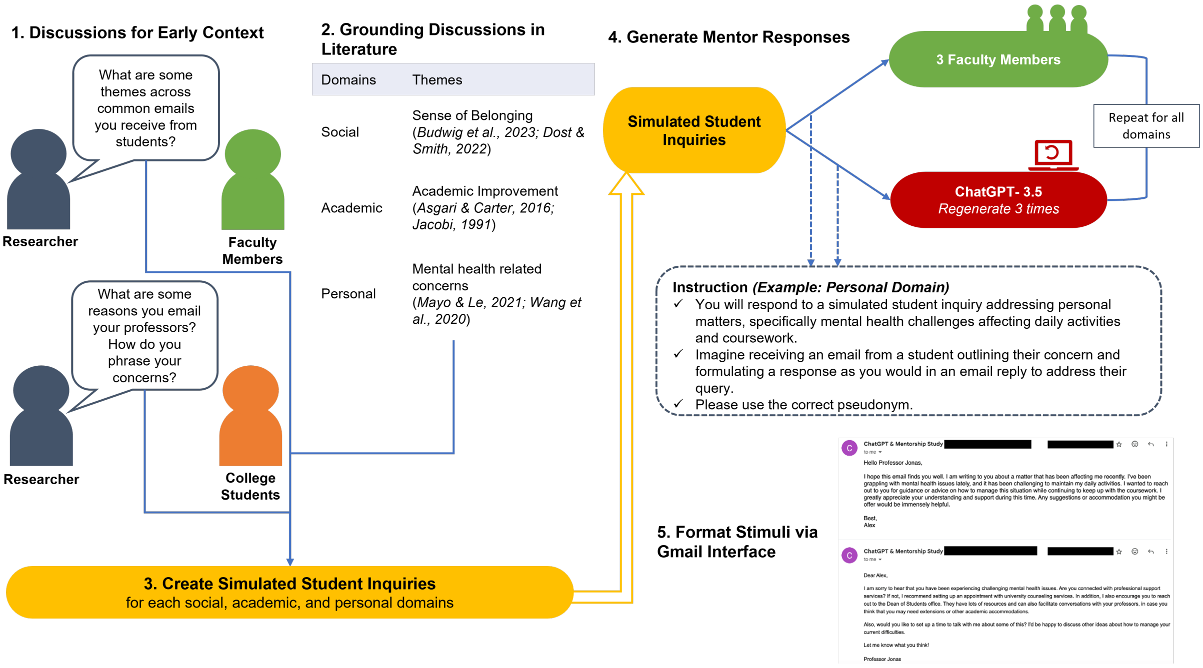


Figure 1: Contextualizing and creating stimuli.

mographic questions, followed by participant evaluations of mentorship interactions. Among the scenarios, the personal domain focused on mental health-related issues (Mayo and Le, 2021; Wang et al., 2020), the social domain focused on sense of belonging (Budwig et al., 2023; Dost and Mazzoli Smith, 2023), and the academic domain focused on academic improvement (Asgari and Carter, 2016; Jacobi, 1991). These domains were selected to reflect a realistic and broadly relevant context in which students seek support from their mentors (see Figure 1).

To explore perceptions of AI-generated versus human responses, participants were presented with three randomized and masked responses drawn from a pool of 18 responses (nine from ChatGPT version 3.5 (OpenAI, 2024b) and nine from human faculty members from three different academic disciplines who had received institutional awards or recognition for mentorship excellence within the past five years). Both ChatGPT and human faculty received identical prompts simulating student inquiries.

For the AI-generated responses, we regenerated three responses for each domain to maintain parity across conditions. Faculty members provided their responses based on previous mentoring experiences and did not use GenAI tools in drafting their replies. All responses were then reformatted to resemble the Gmail interface, reflecting the standard communication format used in many higher education settings.

Participants were instructed to identify whether each response was AI- or human-generated, without receiving feedback on their accuracy (see Table 1). This identification task was designed to activate perceptual biases. Once participants formed an impression of the source, this initial judgment could influence their subsequent evaluation of the response's quality and characteristics.

	AI	**Human**
Domain	%	%
Social	75.81	75.81
Academic	72.03	74.14
Personal	56.57	76.76

Table 1: Accuracy percentage of AI and human responses by domain.

After each identification, participants rated the response on a 5-point Likert scale (1= Not at all, 5= Extremely) across dimensions of helpfulness and caring (see Table 2).

They also provided written explanations for why they believed the response was from AI or a human, and why they rated it as they did, which served as our qualitative data. Following this evaluation task, participants completed additional survey measures assessing their broader perceptions of mentorship and AI in academic contexts.

		Perceived Source				Actual Source			
		AI		Human		AI		Human	
Domain	*Scales*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Social	Helpful	2.90	1.04	3.92	0.78	3.13	1.03	3.67	1.00
	Caring	2.67	1.01	4.00	0.85	2.91	1.12	3.74	1.02
Academic	Helpful	2.91	1.06	3.66	1.03	3.19	1.07	3.40	1.14
	Caring	2.73	1.08	3.56	1.16	3.06	1.17	3.25	1.21
Personal	Helpful	3.04	1.03	3.82	0.99	3.52	1.11	3.48	1.04
	Caring	2.93	1.03	3.77	1.06	3.54	1.11	3.33	1.14

Table 2: Ratings of helpfulness and caring by domain for perceived and actual source. *SD*= standard deviation.

3 Method of Analysis

To address our research question, we used both quantitative and qualitative measures. Quantitative data were analyzed using R (R version 4.3.2, R version 4.4.2) and RStudio (RStudio version 2024.09.1+394, RStudio version 2024.12.1+563) (R Core Team, 2024). To examine whether the helpfulness (Model 1) and care (Model 2) ratings predict the accuracy of the source across domains, we first performed binary logistic linear regression analyses. The reference category for our domain variable was set to Personal, where ratings were consistently higher across all scales.

We then used Inductive Content Analysis (ICA), a qualitative method used to identify patterns in textual data and support exploratory findings. ICA is particularly appropriate in contexts where prior research is limited, as it allows researchers to derive insights directly from the data through systematic coding and theme identification (Vears and Gillam, 2022). Given its applicability to various forms of written text, ICA was especially suitable for our study's purpose of exploring human perceptions and experiences, independent of the specific mode of data collection (Elo and Kyngäs, 2008). The final thematic structure consisted of six overarching categories and 17 subthemes (Table 3).

Using the finalized codebook, we independently coded the qualitative responses. We have previously presented partial results in the personal domain (Lee et al., 2025), but we extended the ICA coding to include the social and academic domains for this study. To ensure analytic consistency and rigor, coding discrepancies were reviewed through a collaborative resolution process (Kyngäs, 2020). When disagreements arose, we held structured consensus-building sessions in which coders explained their rationale for coding decisions (Forman and Damschroder, 2008). Final coding decisions were reached through negotiated agreement.

Main Category	Generic Categories	Sub-Categories
Students' Perceptions of Human vs. AI Mentorship	Tone of Response	Sincerity & Empathy Warmth & Approachability Professionalism & Formality
	Language	Authentic & Natural Clarity & Simplicity Structure & Format
	Information and Resources	Specific Information Resource Guidance Campus Knowledge
	Individualized Support & Actionable Advice	Personalized & Applicable Contextualized Understanding & Support
	Personal Connection	Emotional Connection Establishing Direct Connection in Person Genuine Investment in Student
	Holistic Student Support	Sense of Support Mental Health Overall Well-being and Growth

Table 3: Codebook developed and used for inductive content analysis.

4 Results

We investigated whether students' ratings of Helpfulness (Model 1) and Caring (Model 2) predict their accuracy across the domains.

4.1 Helpfulness and Domain Predicting Accuracy

A binary logistic regression was conducted to examine whether students' ratings of helpfulness predicted their ability to accurately identify the source of mentorship responses (human vs. AI) and whether this relationship differed across personal, social, and academic domains (see Table 4).

	Accuracy			
Predictors	*Odds Ratios*	*SE*	*CI*	*p*
(Intercept)	4.53	1.68	2.22–9.52	**<.001***
Helpfulness	0.79	0.08	0.65–0.96	**.019***
Domain [Social]	0.46	0.24	0.16–1.28	.137
Domain [Academic]	0.64	0.32	0.24–1.71	.374
Helpfulness×Domain [Social]	1.45	0.21	1.09–1.94	**.011***
Helpfulness×Domain [Academic]	1.24	0.17	0.94–1.63	.126
Observations	1289			
Tjur's R^2	0.015			

Table 4: Accuracy predicted by helpfulness and domain. *Indicates p <.05; SE = standard error; CI = 95% confidence interval.

Model 1 revealed a significant main effect of Helpfulness (*OR*= 0.79, p = 0.019, 95%CI [0.65, 0.96]), suggesting that higher helpfulness ratings were associated with lower odds of correctly identifying the response source. There was no significant main effect of domain (Social: p = .137; Academic: p = .374).

Qualitative responses indicated that when students misattributed authorship, it was due to tone, language, personalization, and informativeness. For example, a human-written response was misidentified as AI because it felt "too formal and robotic" (Tone of Response, P67), while another participant described a different human-generated response as "pretty basic without many specific details" (Information and Resource, P68). Several AI-generated responses were perceived as human due to emotionally resonant or personalized phrasing, such as showing "genuine appreciation and understanding of students' struggles" (Tone of Response, P55). These patterns highlight how AI responses can be anthropomorphized, while human mentors may also provide rigid or impersonal responses that fail to meet students' relational expectations.

Furthermore, there was a significant interaction between Helpfulness and the Social domain, indicating that in contexts related to sense of belonging, higher helpfulness ratings were positively associated with identification accuracy. Qualitative insights help explain this result. In the Social domain, participants were more likely to correctly identify human responses when they involved personal outreach, such as "offers to talk with students personally to give suggestions" (Personalized Guidance, P32), or when the tone conveyed compassion while respecting autonomy ("compassionate yet prioritizes the students' autonomy, privacy, and space," Language, P4). In contrast, responses perceived as checklist-like or impersonal were correctly identified as AI, as in comments like "feels incredibly impersonal and provides a checklist more than someone trying to communicate" (Language, P111) or "the advice would work for any university" (Personalized Guidance, P135). The interaction between Helpfulness and the Academic domain was not statistically significant (p = .126).

4.2 Caring and Domain Predicting Accuracy

A second logistic regression tested whether perceived Caring ratings predicted source identification accuracy, and whether this relationship varied across domains (see Table 5).

	Accuracy			
Predictors	*Odds Ratios*	*SE*	*CI*	*p*
(Intercept)	3.83	1.32	1.98–7.62	**<.001***
Caring	0.83	0.08	0.69–0.99	**.043***
Domain [Social]	0.72	0.35	0.28–1.88	.501
Domain [Academic]	1.46	0.69	0.58–3.71	.427
Caring×Domain [Social]	1.28	0.17	0.98–1.67	.073
Caring×Domain [Academic]	0.97	0.13	0.75–1.25	.794
Observations		1289		
Tjur's R^2		0.017		

Table 5: Accuracy predicted by caring and domain. *Indicates p <.05; SE = standard error; CI = 95% confidence interval.

The results showed a significant main effect of Caring, indicating that higher caring ratings were also associated with lower odds of accurate source identification. Though domain effects were not significant (Social: p = .501; Academic: p = .427), nor were the interactions between Caring and Domain (Social: p = .073; Academic: p = .794), our qualitative data illustrate perceptual bias towards responses.

Participants interpreted emotionally validating or well-phrased AI responses as human-authored. Participants reported "[the response] indicated the importance of our well-being" (Holistic Student Support, P58) and "used thoughtfully placed words to show validation and support" (Language, P43). These examples illustrate how AI's capacity to mimic affective tone can lead to over-attribution of caring intent and misidentification. Conversely, human responses perceived as distant or overly formal were misattributed as AI. One participant stated the response "felt a bit cold" (Tone of Response, P64), while another described it as a "scripted response" (Language, P77). Even when human mentors intended to convey care, lack of emotional language or concrete support diminished perceived authenticity: "appears to want to be supportive but does not provide the support in any tangible way" (Personalized Guidance, P64).

Interestingly, when participants correctly identified AI responses, they acknowledged that AI could simulate sympathy or concern, albeit with limitations. Though lacking personal depth, one student remarked that an AI response "did express sympathy regardless of how lackluster it seemed" (Personal Connection, P105), and another noted that "it could have been more to act on, like meeting up, but they did provide other options for help" (Personal Connection, P97). In contrast, accurately identified human responses were seen as invested in student success, but not necessarily emotionally expressive: "polite and invested in the student's success but not super emotionally supportive" (Personal Connection, P4).

5 Discussion

Our study investigated whether students' ratings of helpfulness and caring predicted their accuracy in identifying the source of mentorship responses (human vs. AI) across different domains. We found that higher ratings of both helpfulness and caring were associated with lower accuracy in source identification. This finding suggests that students equate warmth and supportiveness with human authorship, activating perceptual biases that limit their recognition of the potential support AI could provide. The more an AI-generated response resembles a human response, the less likely students are to recognize that it was written by AI.

These findings have critical implications for the integration of AI in mentorship contexts. Students may undervalue or distrust AI-generated guidance when it contradicts their assumptions about what AI can do. Even when AI performs well (e.g., offering emotional validation, supportive tone, or actionable advice), it is often misattributed or dismissed if its source is known. This bias may undermine trust in AI, particularly when relational authenticity is expected. While we previously attributed low source accuracy in the personal domain to students' unfamiliarity with using AI in such contexts, our mixed-methods findings offer a more refined explanation. Students appear to use emotional tone as a heuristic for authorship, interpreting both overly emotional and insufficiently emotional responses as AI-generated. In contrast, they associate human-authored responses with balanced, relationally calibrated communication. Thus, when a response deviates from this midpoint, either too cold or too warm, it violates expectations and is more likely to be attributed to AI.

Furthermore, these insights raise several considerations for AI-supported mentorship in higher education. First, students often expect relational support from humans and assume that AI has its limitations in providing it. This calls for educational interventions to demystify what AI is capable of, especially in relational contexts, and promote informed AI literacy. Second, institutions and AI-designers may develop AI systems tailored to the institution's specific policies, resources, and cultural context, similar to how businesses invest in AI chatbots. This solution could help AI resources feel more as an ethical and trustworthy resource. Third, AI could handle surface-level information requests or initial support, freeing human mentors from trivial tasks or duties to provide deeper and emotionally nuanced engagement. Rather than replacing human mentorship, AI could enhance it when used as a complementary resource. Lastly, just as students misperceive AI as human based on warmth, they also misperceive humans as AI when their tone is rigid, detached, or overly formal. Institutions might consider providing training for their faculty and staff members on relational communication strategies, especially in email or digital interactions, to ensure that students are supported, even in brief exchanges.

Our study is not without limitations. First, the explanatory power of our models was weak (Model 1 Tjur's R^2 = .015, Model 2 R^2 = .017). These values suggest that, while the predictors were statistically significant, they account for only a small proportion of the variance in the accuracy of the source identification. Future research should replicate these findings using a larger sample size, a greater variety of stimuli, and more diverse educational contexts to improve generalizability. Second, the participant pool was limited to undergraduate students from a single liberal arts college in the northeastern United States. As such, the findings may reflect institution-specific dynamics and should be interpreted as exploratory or case-based. Expanding this research to include participants from multiple institutions and institutional types (e.g., community colleges, large public universities) would provide a more comprehensive understanding of students' perceptions of AI and human mentorship. Third, although all faculty responses in this study were entirely human-authored without any AI assistance, it is possible that participants may have assumed that the faculty used AI tools to help craft their replies. Furthermore, the format in which responses were presented, modeled after email-based communication, may have influenced how participants perceived both the content and the source. AI responses framed as email replies may have appeared more human-like than if they were delivered through a chatbot or system-generated interface. This framing could have unintentionally blurred distinctions between human and AI authorship. Future research should investigate how different presentation formats (e.g., email, chatbot, forum post) shape students' assumptions about authorship and credibility, and compare perceptions of human-authored, AI-assisted human-authored, and AI-authored mentorship responses across these contexts. Lastly, future studies would benefit from examining demographic variables such as race, ethnicity, and gender. Understanding how students from diverse backgrounds interpret and engage with AI-generated versus human mentorship may yield important insights, particularly as institutions strive to promote equity and culturally responsive mentorship practices.

Perceptual bias is not simply a barrier to AI adoption, but is a lens through which students interpret support and relational intent. Our results show that emotionally resonant, helpful responses are often mistaken for humans regardless of authorship, while detached or impersonal responses are perceived as AI. Emotional tone and personalization appeared to be more influential than the actual source in shaping students' evaluations. Yet, these biases are not fixed. As students gain more exposure to AI and as these tools become more embedded in academic settings, their ability to discern source and engage with AI more responsibly and meaningfully may improve. Our findings and recommendations provide a reflection of deeper sociocultural expectations about relational care, authenticity, and the boundaries between human and machine. The future of AI mentorship depends not just on technical capability, but on thoughtful, human-centered design that attends to the cognitive and relational dynamics in higher education settings.

Acknowledgments

We thank our participants for their time in taking part in this study. We also extend many thanks to faculty members who contributed their time to generating stimuli. Additionally, we acknowledge ChatGPT (version 3.5) for its role in stimulus generation.

References

Shaki Asgari and Frederick Carter. 2016. Peer mentors can improve academic performance: A quasi-experimental study of peer mentorship in introductory courses. *Teaching of Psychology*, 43(2):131–135.

Dominic Baidoo-Anu and Linda Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN Electronic Journal*.

Nancy Budwig, Jimin Lee, and Raquel Jorge Fernandes. 2023. A developmental and sociocultural approach to the transition from high school to college: The importance of understanding student meaning-making. *Human Arenas*.

Joshua Dempere, Kiran Modugu, Ahmed Hesham, and Lakshmi K. Ramasamy. 2023. The impact of ChatGPT on higher education. *Frontiers in Education*, 8.

Gökhan Dost and Laura Mazzoli Smith. 2023. Understanding higher education students' sense of belonging: A qualitative meta-ethnographic analysis. *Journal of Further and Higher Education*, 47(6):822–849.

Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115.

Tom Farrelly and Nigel Baker. 2023. Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences*, 13(11):1109.

Jane Forman and Laura Damschroder. 2008. Qualitative content analysis. In Laura Damschroder, editor, *Empirical methods for bioethics: A primer*, volume 11 of *Advances in Bioethics*, pages 39–62. Emerald/Elsevier.

Mark Glickman and Tali Sharot. 2025. How human–AI feedback loops alter human perceptual, emotional, and social judgements. *Nature Human Behaviour*, 9:345–359.

Diala Hammoudi Halat, Amir Soltani, Rachid Dalli, Lujain Alsarraj, and Ahmed Malki. 2023. Understanding and fostering mental health and well-being among university faculty: A narrative review. *Journal of Clinical Medicine*, 12(13):4425.

Mary Jacobi. 1991. Mentoring and undergraduate academic success: A literature review. *Review of Educational Research*, 61(4):505–532.

Mohd Javaid, Abid Haleem, Rajiv P. Singh, Shujaat Khan, and I. H. Khan. 2023. Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(2):100115.

Nils Köbis and Christian Mehner. 2021. Ethical questions raised by AI-supported mentoring in higher education. *Frontiers in Artificial Intelligence*, 4:624050.

Helvi Kyngäs. 2020. Inductive content analysis. In *The application of content analysis in nursing science research*, pages 13–21. Springer.

Hanh Le, Yulong Shen, Zhaoyuan Li, Minghui Xia, Linying Tang, Xiaoyu Li, Jia Jia, Qiyun Wang, Dragan Gašević, and Yubo Fan. 2025. Breaking human dominance: Investigating learners' preferences for learning feedback from generative ai and human tutors. *British Journal of Educational Technology*, 56:1758–1783.

Jimin Lee and Alena G. Esposito. 2025a. ChatGPT or human mentors? student perceptions of technology acceptance and use and the future of mentorship in higher education. *Education Sciences*, 15(6):746.

Jimin Lee and Alena G. Esposito. 2025b. A comparative study between college student perception of an AI learning tool and human mentoring responses. Manuscript submitted for publication.

Jimin Lee, Si Wang, and Alena G. Esposito. 2025. Content analysis of college student perceptions: Mental health-related support from generative AI versus faculty mentors. *American Journal of Qualitative Research*. In press.

Douglas Mayo and Benjamin Le. 2021. Perceived discrimination and mental health in college students: A serial indirect effects model of mentoring support and academic self-concept. *Journal of American College Health*, 71(4):1184–1195.

Marta Montenegro-Rueda, José Fernández-Cerero, José M. Fernández-Batanero, and Eloy López-Meneses. 2023. Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, 12(8):153.

Wendy Nuis, Mien Segers, and Simon Beausaert. 2023. Conceptualizing mentoring in higher education: A systematic literature review. *Educational Research Review*, 41:100565.

OpenAI. 2024a. Chatgpt. `https://chat.openai.com/`. Conversational AI system; accessed March 2024.

OpenAI. 2024b. Gpt-3.5 (gpt-3.5-turbo) model card and usage. `https://platform.openai.com/docs/models#gpt-3-5`. Model version used in this study; accessed March 2024.

R Core Team. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jean E. Rhodes, Jean Baldwin Grossman, and Nancy L. Resch. 2000. Agents of change: Pathways through which mentoring relationships influence adolescents' academic adjustment. *Child Development*, 71(6):1662–1671.

Michael Schönberger. 2023. ChatGPT in higher education: The good, the bad, and the university. In *Proceedings of the 9th International Conference on Higher Education Advances (HEAd'23)*, pages 331–338.

Danya F. Vears and Lynn Gillam. 2022. Inductive content analysis: A guide for beginning qualitative researchers. *Focus on Health Professional Education: A Multi-Professional Journal*, 23(1):111–127.

Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478.

Xueming Wang, Shweta Hegde, Chiho Son, Bethany Keller, Alexander Smith, and Farzan Sasangohar. 2020. Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional survey study. *Journal of Medical Internet Research*, 22(9):e22817.

AI-Based Classification of TIMSS Items for Framework Alignment

Ummugul Bezirhan and **Matthias von Davier**
TIMSS & PIRLS International Study Center at Boston College
{bezirhan, vondavim}@bc.edu

Abstract

Large-scale assessments rely on expert panels to verify that test items align with prescribed frameworks, a labor-intensive process. This study evaluates the use of GPT-4o to classify TIMSS items to content domain, cognitive domain, and difficulty categories. Findings highlight the potential of language models to support scalable, framework-aligned item verification.

1 Introduction

International large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) play a critical role in monitoring educational outcomes across diverse systems. The validity argument of such assessments lies in the rigorous alignment of test items with the underlying assessment framework, which defines key content and cognitive domains that the assessment purports to measure. TIMSS assessment development is guided by the principles of Evidence-Centered Design (Mislevy et al., 2003), ensuring that each item serves as meaningful evidence for the targeted constructs. This process involves multiple rounds of expert review and collaboration with participating countries to verify item alignment and maintain the validity of measurement across contexts.

While effective, this expert-driven validation process is labor-intensive and time-consuming, particularly in the context of ongoing item development and reuse. As AI technologies continue to evolve, they offer new ways for automating or supporting some of these processes. One such approach is the use of large language models (LLMs) for automated item classification. If reliable, these tools could significantly reduce the burden on subject matter experts, streamline assessment development cycles, and enhance scalability without compromising psychometric quality.

This study explores the potential of GPT-4o to perform classification of TIMSS 2019 mathematics items. Specifically, we evaluate the model's ability to assign items to their appropriate content domain, cognitive domain, and difficulty level, based on the given TIMSS assessment framework. The items have already been reviewed and validated by expert panels and are used operationally, their classifications can be considered reliable benchmarks.

To assess alignment, AI-generated classifications are compared against expert-coded categories, analyzing agreement patterns and identifying systematic divergences. For difficulty, we define three difficulty regions using percent correct values derived from empirical item performance data and evaluate the model's capacity to approximate these classifications. The findings of this study contribute to ongoing discussions about the role of AI in assessment development and offer preliminary evidence on the feasibility of LLMs as tools to support item verification within established assessment frameworks.

2 Background

Construct validity has long been a central concern in educational assessment, particularly in international large-scale assessments such as TIMSS. A key aspect of evidence for validity is the alignment between test items and the assessment framework, that is the extent to which each item's

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 134–141
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

content and cognitive demands reflect the intended constructs of the study. Alignment in the context of ILSAs supports meaningful score interpretation, facilitates cross-national comparability, and provides assurance that assessment inferences are based on systematically defined learning goals. This also helps minimize the construct irrelevant variances.

Foundational work on test design and validity, such as Messick's (1990) unified validity framework and ECD of Mislevy et al. (2003), emphasizes that the validity argument must include an explicit evidentiary chain connecting item features to well-articulated domain models. Alignment research is one way to establish this chain by evaluating the connection between testing, content standards, and instruction. If these components work together to deliver a consistent message about what should be taught and assessed, students will have the opportunity to learn and to truly demonstrate what they have achieved (Martone & Sireci, 2009). Systematic alignment studies therefore provide critical priori evidence that the assessment operationalizes its framework as intended, thereby supporting the overall construct-validity argument.

In the context of TIMSS, alignment involves a multistep process where items are reviewed, refined, and approved by subject matter experts, ensuring they adhere to content domains, cognitive processes and intended difficulty levels. While this process is foundational to the psychometric integrity of the assessment, it is also resource-intensive and difficult to scale given growing item pools and evolving frameworks.

To address these challenges, researchers have explored the use of computational methods to support or automate parts of the alignment process. Advances in natural language processing (NLP) have opened new possibilities for supporting alignment through semantic analysis of item texts. Recent studies (e.g., Butterfuss & Doran, 2024; Camilli, 2024; Camili & Suter, 2024) have demonstrated that embedding-based similarity metrics can successfully identify meaningful relationships between standards and item specifications. Such methods have been used in alignment studies involving the Common Core State Standards and NAEP, showing that NLP techniques can reproduce many expert classifications through clustering or regression models. While promising, these approaches often rely on static sentence embeddings and do not fully capture the contextual reasoning that human experts employ when classifying items.

Building on this prior work, the current study investigates the use of a large language model, GPT-4o, to perform classification of TIMSS mathematics items in alignment with the given TIMSS framework. By incorporating the full descriptive language of the framework into the prompt through a structured prompt engineering approach that dynamically loads framework specifications from a framework focused database, this method allows complete content domain descriptions, cognitive skill definitions, and difficulty level characteristics specific to each TIMSS assessment year and grade level. Unlike previous efforts that focus on pairwise similarity, this dynamic framework-informed prompting strategy offers a scalable, interpretable, and multidimensional approach to item classification, potentially streamlining alignment procedures while preserving the integrity of the assessment development process.

3 Methods

3.1 Data Source

This study uses a sample of mathematics items from TIMSS 2019 for Grade 4 and Grade 8 assessments. All selected items were previously reviewed and validated by expert panels convened by TIMSS and PIRLS International Study Center and successfully field tested. Each item includes a final assigned content domain, cognitive domain, and empirical difficulty estimate based on percent correct values from operational test data.

The study includes all newly developed items introduced in the TIMSS 2019 cycle. For items containing images, diagrams, or graphs, the GPT-4o model via the OpenAI API was used to generate descriptive captions, allowing for the full item set to be processed in text-based analyses. In each TIMSS cycle items are selected to ensure coverage across a range of content topics (e.g., number, algebra, life science), cognitive domains (knowing, applying, reasoning), and difficulty levels. The complete dataset initially consisted of 286 items. However, items split into multiple parts (e.g., a, b, c sub-items) were excluded from the classification analysis to avoid duplication and ensure consistency in unit of analysis. After this filtering, the final analytic sample comprised 217 items.

Table A1 shows the item distribution by each category in Appendix A.

3.2 Framework Representation and Prompt Design

To support classification by the language model, we constructed structured prompts embedding full descriptions of TIMSS framework dimensions. TIMSS 2019 Assessment Framework (Mullis & Martin, 2017) served as a primary source for content domain definitions, cognitive domain descriptions, and difficulty-level guidance.

A custom framework database was built utilizing PDF descriptions of the frameworks to dynamically retrieve definitions relevant to the grade level and subject of each item. Prompts followed a template-based structure that presented:

- The item content
- The TIMSS subject, grade level, and year
- Full framework definitions for the content domains
- Full framework definitions for the three cognitive domains
- Empirical guidance for difficulty classification

An example of the prompt is given in Figure A1 in Appendix A.

In addition to aligning with the official TIMSS framework, this study examined how prompt design strategies influence the language model's classification performance across three target dimensions: content domain, cognitive domain, and difficulty level.

Recent advances in natural language prompting have shown that model performance can be improved by structuring reasoning and task representation within the prompt itself. Two key strategies examined in this study are Chain-of-Thought (CoT) prompting and meta-prompting. CoT prompting encourages the model to generate step-by-step reasoning before producing a final answer, supporting tasks that involve multi-step inference or abstract judgment (Wei et al., 2022). This approach is particularly relevant for educational item classification tasks, where judgments such as cognitive demand and difficulty are often nuanced and require the model to simulate student and/or expert thinking.

Building on this, meta-prompting involves instructing the model on how to perform the task itself by embedding structured guidelines directly into the prompt (Reynolds & McDonell, 2021; OpenAI, 2024). In more advanced forms, meta-prompts may enable models to critique or revise their own instructions or those provided by users (Ye et al., 2023). Recent work has further enhanced this approach by labeling individual reasoning steps and implementing step-aware verifiers, which assess each step's contribution to the final decision (Li et al., 2023).

To evaluate the influence of prompt structure on classification performance, the study implemented four prompt conditions shown in Table 1.

Condition	Description	Examples	CoT
Zero-shot (ZS)	Framework definitions + item only	None	No
Zero-shot CoT (ZS-CoT)	Adds "Think step by step" instruction	None	Yes
Few-shot (FS)	Adds one example per cognitive × difficulty cell	9-10	No
Few-shot CoT (FS-CoT)	Adds one example per cognitive × difficulty cell and CoT reasoning	9-10	Yes

Table 1: Prompting Conditions

3.3 Model and Classification Procedure

We used GPT-4o, accessed via OpenAI's API, as the large language model for classification. Each item prompt was submitted independently, and the model's textual response was parsed to extract predicted content domain, cognitive domain, and difficulty level. A post-processing script was applied to standardize terminology and correct minor inconsistencies such as the content domain in grade 4 is 'Measurement and Geometry' but the model specified the items as 'Geometry' or 'Measurement', those were counted as 'Measurement and Geometry'.

The classification process was fully unsupervised; no labeled training data or fine-tuning was used. All responses were generated using temperature = 0 to maximize determinism and reproducibility.

Model performance was evaluated by comparing model's predicted content and cognitive domain classifications to expert-assigned labels. Content and cognitive domain accuracies reflect the proportion of exact matches between the model's predictions and the official domain labels. Difficulty classification was evaluated against empirical difficulty levels derived from operational data. Specifically, items were categorized as Easy, Medium, or Hard based on their percent-correct values, using Easy (>60%), Medium (30-60%), and Hard (<30%). The model's predicted difficulty level was considered correct if it matched the empirically derived category for each item. Cohen's kappa coefficients were also calculated to account for chance agreement. Additionally, misclassifications were analyzed qualitatively to identify systematic patterns of divergence.

4 Results

Classification Performance

Classification performance across prompting conditions is summarized in Table 2. Content domain classification demonstrated consistently high performance, with all prompting conditions exceeding 94% accuracy and kappa values above 0.92, indicating substantial agreement beyond chance. FS-CoT achieved the highest accuracy (94.9%) and kappa (0.933), reflecting the model's strong ability to differentiate TIMSS content domains. In contrast, classification accuracy for the cognitive domain showed more variation, ranging from 60.4% to 64.1% and kappa values between 0.382 and 0.438. The FS-CoT condition yielded the highest accuracy, followed by ZS baseline and FS. Kappa values across these conditions suggest fair to moderate agreement with expert labels, indicating that while the model captures meaningful cognitive distinctions, it does so with less precision than in the content domain.

Difficulty classification, while the most challenging of the three dimensions, showed improvement over previous iterations. Accuracy scores ranged from 44.2% to 49.8%, and all conditions resulted in positive kappa values, indicating better-than-chance agreement. ZS-CoT led in both accuracy and agreement, though overall performance remained modest, highlighting the inherent complexity of predicting empirically derived difficulty levels. Grade level analysis revealed consistently stronger model performance for Grade 4 items across all classification dimensions. For content domain classification, Grade 4 items achieved exceptional accuracy scores ranging from 96.9% to 97.7%. Grade 8 content domain performance, while lower, remained strong with accuracy scores from 91.0% to 92.3%. A similar pattern was also observed in cognitive domain classification. Grade 4 accuracy ranged from 62.6% to 65.0%, while grade 8 performance varied from 57.4% to 62.8%. Notably, the FS-CoT condition achieved the smallest grade-level gap in cognitive domain performance (65.0% vs. 62.8%). For difficulty classification, Grade 4 items consistently outperformed Grade 8 items across all conditions. Grade 4 difficulty accuracy ranged from 50.4% to 57.7%, with ZS-CoT achieving the highest Grade 4 performance (57.7%). Grade 8 difficulty classification proved more challenging, with accuracy scores ranging from 33.0% to 40.4%, with FS-CoT achieving the best Grade 8 performance (40.4%).

Prompt	Content Domain		Cognitive Domain		Difficulty Level	
	Acc	κ	Acc	κ	Acc	κ
ZS	94.1	0.922	62.2	0.410	44.2	0.072
ZS CoT	94.2	0.923	60.4	0.382	**49.8**	**0.134**
FS	94.1	0.923	61.3	0.397	44.7	0.074
FS CoT	**94.4**	**0.930**	64.1	0.438	48.4	0.097

Table 2: Classification Performance

Classification Patterns and Systematic Errors

Given its overall better performance across all three classification dimensions, the FS-CoT condition was selected for detailed confusion matrix analysis to understand specific classification patterns and systematic errors.

For the content domain classification, the model achieved near perfect classifications, but specific patterns emerged when analyzed by grade level (Appendix A Figures A2-A3). For Grade 4 mathematics, the model achieved perfect classification for Data and Number domains but showed some boundary confusion with Measurement and Geometry items. Specifically, 12% of Measurement and Geometry items were misclassified as both Data and Number domains, suggesting overlapping conceptual features in items involving spatial reasoning and numerical computation. For Grade 8, content domain classification revealed different boundary challenges. While Algebra, Data and Probability, and Geometry domains were classified perfectly, Number domain items showed notable confusion (75%). The primary misclassification pattern involved 21% of these items being classified as Algebra, with an additional 4% classified as Geometry. This pattern suggests that model struggles with the increasing integration of algebraic thinking into numerical context in the higher grades.

As shown in Figure A4, the FS-CoT model exhibited a strong bias toward predicting the Applying domain. While Applying items were accurately classified 84% of the time, it also attracted most misclassifications receiving 45% of Knowing and 44% of Reasoning items. Reasoning accuracy was moderate (53%) but showed substantial confusion with Applying. Very few items were confused between Knowing and Reasoning, indicating the model can generally distinguish between higher-order and basic cognitive demands but struggles to differentiate between applying procedures and engaging in mathematical reasoning.

Difficulty classification remained the most challenging task for the model, with a strong tendency toward underestimation (Figure A5 in Appendix A). Easy items were correctly classified 64% of the time and no easy items were misclassified as Hard, indicating a cautious estimation pattern. Medium items had 71% accuracy, with 26% underestimated as Easy and only 3% overestimated as Hard. This suggests the model treats difficulty as a binary decision *Easy versus Not Easy* rather than effectively distinguishing all three levels. If we collapse the difficulty to this more pragmatic *Easy vs. not Easy* decision, the accuracy jumped to 0.78. Hard items were the most frequently misclassified. This reflects a consistent failure to recognize complex mathematical or cognitive demands, particularly when such items are concise or lack surface-level cues of difficulty.

Linguistic Features of Misclassified Items

To better understand the systematic errors in difficulty classification, we examined surface features of misclassified items as shown in Table 3. We focused on textual length, numerical content, and mathematical language.

	Easy	Medium	Hard
Word count	76.8	49.9	66.1
Character count	561.4	338.9	404.5
Reasoning Verb count	0.20	0.09	0.34
Number count	19.10	12.29	10.74
Operations count	0.40	1.12	1.31

Table 3: Average Surface Features of Misclassified Items

Misclassified Easy items had the highest average word count (76.8) and character length (561.4) substantially longer than misclassified Medium (50.0 words, 338.9 characters) and Hard items (66.1 words, 404.5 characters). This suggests the model tends to get confused by textual elaboration with cognitive difficulty, overestimating the challenge of otherwise straightforward tasks. Conversely, Hard items, though shorter, were rich in mathematical content. They contained the highest density of mathematical operations (1.31 per item) and reasoning verbs (0.34 per item) yet were overwhelmingly misclassified as Medium. This indicates that while GPT-4o detects complexity, it fails to properly weight them in difficulty estimation, especially when such cues are embedded in concise text.

5 Conclusion

This study evaluated the potential of GPT-4o to perform automated classification of TIMSS mathematics items. Using a dynamic, framework-aware prompting strategy, we challenged the model to assign Grade 4 and Grade 8 mathematics items to their official content domain, cognitive domain, and difficulty categories without any fine-tuning or labeled training data.

Across all prompting conditions, model consistently provided high agreement with content-domain classifications with about 95% accuracy ($\kappa > 0.92$), and confusion matrices only showed minimal boundary issues. These results suggest that content domain classification is one area where the model can be deployed with confidence.

Model accuracy for cognitive domain classifications clustered around 62% ($\kappa \sim 0.41$). This level of agreement is consistent with prior research, including Nasstrom (2009), who reported moderate inter-rater reliability ($\kappa \sim 0.41$–0.47) among experts classifying items according to Bloom's taxonomy. Similarly, Karpen and Welch (2016) found only 46% agreement among faculty when categorizing exam questions by cognitive demand. While performance improved modestly under the FS-CoT prompting condition, error analysis revealed a systematic tendency to overclassify items into the Applying category, a middle category bias. This highlights a clear opportunity for targeted prompt engineering or probability calibration strategies.

Model performance was weakest for the three-level difficulty classification task, with accuracy around 49%. However, reframing the task as a binary classification, *Easy versus Not Easy*, yielded 78% accuracy. This is particularly notable given that prior research demonstrated limited alignment between expert predictions of item difficulty and examinee performance (e.g., Bejar, 1983; Mansoor, 2024; Wonde, 2024) with accuracy rates hovering around 50-55% even after targeted expert training (Sayin & Bulut, 2024). Moreover, Clauser et al. (2009) demonstrated that physicians involved in Angoff standard setting frequently revised their difficulty estimates to align with whichever performance statistics were presented to them, regardless of their accuracy, highlighting the inherent instability of human judgements. Taken together, these findings show that unsupervised binary screening already matches or in some cases exceeds typical human baselines.

Given this, the model could serve as a first-pass filter content tagging and binary difficulty screening could reduce the number of items requiring full panel review, freeing experts time to focus on distractor quality, fairness checks, and cross-cultural comparability. In addition, because framework definitions are pulled dynamically the same pipeline can be applied to other TIMSS cycles or entirely different frameworks (e.g., NAEP, PISA) with minimal revision.

This study has potential limitations. First, the study focused exclusively on mathematics items from the 2019 TIMSS cycle; generalizability to science items, earlier cycles, or AI-generated content remains to be investigated. Second, all analyses were conducted using text-only representations of items thus visual components such as graphs or diagrams were reduced to captions, which may have affected the model's judgments. Future studies incorporating multimodal inputs may offer a more accurate reflection of the item's full content and complexity. Third, item difficulty levels were defined based on fixed percent-correct thresholds. Future research can consider using IRT-based difficulty estimates or continuous difficulty prediction using fine-tuned LLMs.

Overall, this study shows that GPT-4o, when directed with a targeted prompting strategy, can act as a reliable co-reviewer in the early stages of test development. While current results are strongest for content classification, meaningful performance in cognitive and difficulty domains, with interpretable error patterns, suggests a promising role for AI in supporting expert workflows. Rather than aiming to replace human expertise, these tools are best positioned to augment it by reducing workload and improving the speed and consistency of assessment development.

References

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310.

Butterfuss, R., & Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. Educational Measurement: Issues and Practice.

Camilli, G. (2024). An NLP crosswalk between the common core state standards and NAEP item specifications. arXiv preprint arXiv:2405.17284.

Camilli, G., & Suter, L. (2024). NLP Cluster Analysis of Common Core State Standards and NAEP Item Specifications. arXiv preprint arXiv:2412.04482.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*, *46*(4), 390-407.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. *Currents in Pharmacy Teaching and Learning*, *8*(6), 885-888.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J. G., & Chen, W. (2023, July). Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5315-5333). https://aclanthology.org/2023.acl-long.291

Mansoor, M., Imran, S., Tayyab, A., & Sarfraz, R. (2024). Expert Prediction Versus Difficulty Index Measured by Psychometric Analysis; A Mixed Method Study Interpreted through Diagnostic Judgment by Cognitive Modeling Framework. *Journal of University College of Medicine and Dentistry*, 74-80.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of educational research*, *79*(4), 1332-1361.

Messick, S. (1990). Validity of test interpretation and use.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29.

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Näsström, G. (2009). Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. *International Journal of Research & Method in Education*, *32*(1), 39-51.

OpenAI. (2024). *Prompt generation.* https://platform.openai.com/docs/guides/prompt-generation

Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1-7). https://doi.org/10.1145/3411763.3451760

Sayın, A., & Bulut, O. (2024). The difference between estimated and perceived item difficulty: An empirical study. *International Journal of Assessment Tools in Education*, *11*(2), 368-387.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Wonde, S. G., Tadesse, T., Moges, B., & Schauber, S. K. (2024). Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian Undergraduate Medicine Licensure Examination. *BMC Medical Education*, *24*(1), 1016.

Ye, Q., Axmed, M., Pryzant, R., & Khani, F. (2023). Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*. https://doi.org/10.48550/arXiv.2311.05661

A Appendix

Grade	Category Type	Category	Count
4	Cognitive Domain	Applying	72
		Knowing	53
		Reasoning	39
	Content Domain	Data	60
		Measurement and Geometry	50
		Number	54
	Difficulty	Easy	43
		Medium	87
		Hard	34
8	Cognitive Domain	Applying	54
		Knowing	41
		Reasoning	27
	Content Domain	Algebra	35
		Data and Probability	26
		Geometry	26
		Number	35
	Difficulty	Easy	14
		Medium	56
		Hard	52

Table A1: Item Distribution Across Categories

Act as an expert specializing in the TIMSS assessment framework. Your task is to simulate how students interact with a {subject_name} item, diagnose its cognitive demand, and judge its difficulty level from both an expert and a student perspective.

Analyze the given TIMSS Grade {grade} {subject_name} assessment item.

Classify this item according to the TIMSS {year} {subject_name} Framework. Use these three categories:

1. **Content Domain**: Select the main content domain from this list (use the exact name):

{content_domains_text}

2. **Cognitive Domain**: Identify the main cognitive domain (choose exactly one: Knowing, Applying, Reasoning):

{cognitive_domains_text}

3. **Difficulty Level**: Indicate the item's difficulty (Easy / Medium / Hard), based not only on typical student success rates but also on complexity, required reasoning, potential misconceptions, distractor strength, and student accessibility:

{difficulty_text}

Figure A1: Prompt Structure – Zero Shot

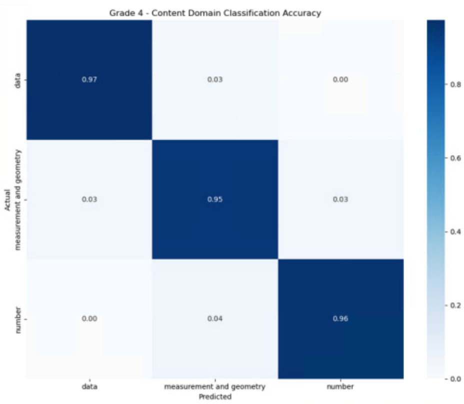


Figure A2: Grade 4 Content Domain Confusion Matrix

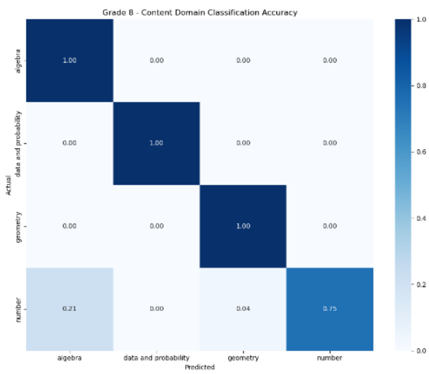


Figure A3: Grade 8 Content Domain Confusion Matrix

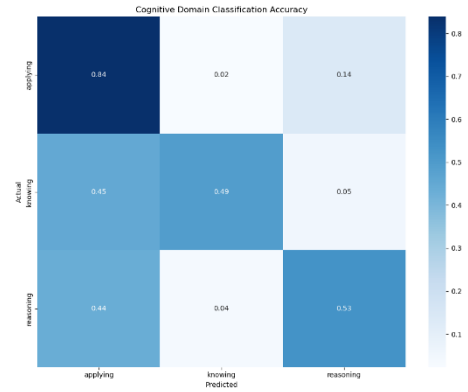


Figure A4: Cognitive Domain Confusion Matrix

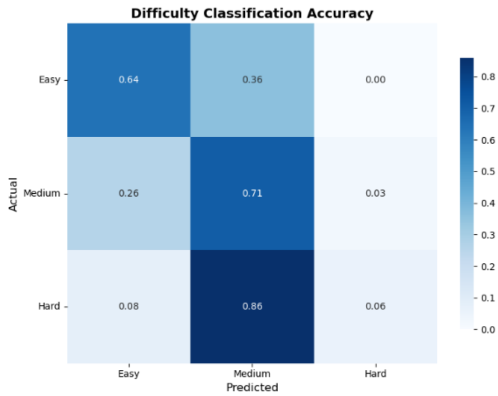


Figure A5: Difficulty Confusion Matrix

Towards Reliable Generation of Clinical Chart Items: A Counterfactual Reasoning Approach with Large Language Models

Jiaxuan Li[1], Saed Rezayi[2], Peter Baldwin[2], Polina Harik[2], Victoria Yaneva[2]

[1]University of California Irvine, [2]NBME

jiaxul19@uci.edu (srezayidemne,pbaldwin,pharik,vyaneva)@nbme.org

Abstract

Educational assessment organizations continuously need new test items. This paper presents an exploratory study on the use of large language models (LLMs) for generating item drafts in medical education, focusing specifically on patient chart items. Using GPT-4, we developed and compared three prompting strategies—Chain-of-Thought, counterfactual reasoning, and information-theoretic sample selection—on the quality of the generated drafts. Our prompts include clinical vignettes from existing multiple-choice questions. Evaluation by two clinical experts showed that at least a quarter of the items were free from major flaws at first assessment, and half were considered useful starting points compared to creating items from scratch. We found our proposed counterfactual framework could generate novel items while maintaining the overall quality and accuracy of generated items. The quality of generated items was sensitive to the information-theoretic properties of examples in few-shot learning settings, where example questions with higher surprisal of the correct answers enhanced the quality of generated items. To the best of our knowledge, this is the first study to explore the potential of LLMs for automatic generation of clinical chart items.

1 Introduction

To ensure the relevance and integrity of examinations, educational assessment organizations must continuously develop new, high-quality test items. This is especially critical in the context of high-stakes assessments[1], where test items must not only cover necessary subject material but also conform to rigorous psychometric standards to ensure fairness, validity, and reliability. The process of crafting such test items is inherently complex and resource-intensive, requiring substantial expertise and time investment from subject matter experts. This is particularly challenging for medical education, where the test items need to accurately capture complex real-world problems and reflect highly specialized and rapidly changing knowledge.

Efforts to automate the full or partial creation of test items have long been explored as a means to address the need for scalable and efficient assessment development. Rule-based approaches and cognitive modeling have been widely applied in automated item generation (AIG) (Gierl and Lai, 2016; Lai et al., 2016a; Falcão et al., 2022; Circi et al., 2023). For instance, rule-based methods have been used to enhance distractor quality in MCQs through the integration of knowledge graphs (Lai et al., 2016b). More recently, LLMs have been profitably used for item generation across a range of domains including STEM education, cognitive assessments, as well as language proficiency testing (Attali et al., 2022; Prasetyo et al., 2020; Laverghetta Jr and Licato, 2023; Lee et al., 2023; Chan et al., 2024; Belzak et al., 2023). For example, LLMs in zero- or few-shot learning settings have successfully generated items that have achieved acceptable validity and reliability for various STEM subjects (Chan et al., 2024).

LLMs have demonstrated impressive performance with various medical tasks (Zhou et al., 2023). These include discriminative tasks like question answering (Jin et al., 2019; Yaneva et al., 2023; Naseem et al., 2021; Romanov and Shivade, 2018) as well as generative tasks such as clinical report generation (Johnson et al., 2016; Zhang et al., 2024b). However, most medical LLMs involve pretraining (Zhang et al., 2024a; Jin et al., 2023; Luo et al., 2022; Gu et al., 2021) or fine-tuning (Christophe et al., 2024; Gururajan et al., 2024; Luo et al., 2023), which may require expensive computation resources. The adoption of pretrained LLMs for AI-assisted item creation in the medical domain remains a challenge (Karabacak et al.,

[1]Examinations with significant consequences for the test-taker, such as professional certification or licensure.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 142–153
October 27-29, 2025

2023). A systematic survey suggests that off-the-shelf generative language models such as ChatGPT struggle to generate high-quality multiple-choice medical questions, even with advanced prompting strategies (Kıyak and Emekli, 2024).

In this paper, we perform an initial investigation of the potential of LLMs to assist with creating comprehensive documents of patient's medical record (clinical charts) and multiple choice questions for medical education exams. We prompt off-the-shelf pretrained language models with clinical vignettes from a publicly available dataset (MedQA; Jin et al., 2021) and develop three different approaches for item generation in a few-shot learning setting, including *Chain-of-Thought Generation*, *Counterfactual Generation*, and *Principled few-shot learning sample selection*. The generated items are evaluated by two licensed medical doctors who are medical school faculty. We found that our proposed counterfactual generation framework produces items with greater lexical and semantic distance from source material while maintaining overall quality, and that information-theoretic properties of samples in few-shot learning settings influence the quality of generated items. To the best of our knowledge, this is the first study to explore the potential of a counterfactual generation framework with principled learning sample selection for generating clinical chart items.

2 Method

2.1 Data

MedQA This study uses data from two distinct sources. The first source is MedQA (Jin et al., 2021), a publicly available dataset containing $\approx$ 60K clinical MCQs in English, simplified Chinese, and traditional Chinese. These MCQs were collected from various test preparation materials available online. In our study, we use the English-language subset, which contains 12,723 items.

Chart items The second source is a dataset of 35 chart items (see Fig. 1 for an example item). These items were developed as part of a research project on assessing clinical reasoning and the specific items used in this study are referred to as SHARP items (SHort Answer, Rationale Provision; see Runyon et al. (2023) for a full description of the item format).

Clinical charts, also known as patient records, are comprehensive documents that typically include a patient's medical and social history, presenting symptoms, chief complaints, physical examination findings, and test results. They may also contain physician notes documenting patient visits, differential diagnoses, and treatment plans. In medical education, clinical charts serve as a structured and effective tool for training future physicians (Deschênes et al., 2025; Goulet et al., 2007), bridging the gap between theoretical knowledge and real-world medical practice (Al-Wassia et al., 2015).

One of the primary benefits of using patient charts in medical education is the enhancement of clinical reasoning and decision-making skills (Daniel et al., 2019). By reviewing and analyzing patient charts, medical students can practice prioritizing information, identifying key features, formulating differential diagnoses, developing treatment plans, and making informed clinical decisions.

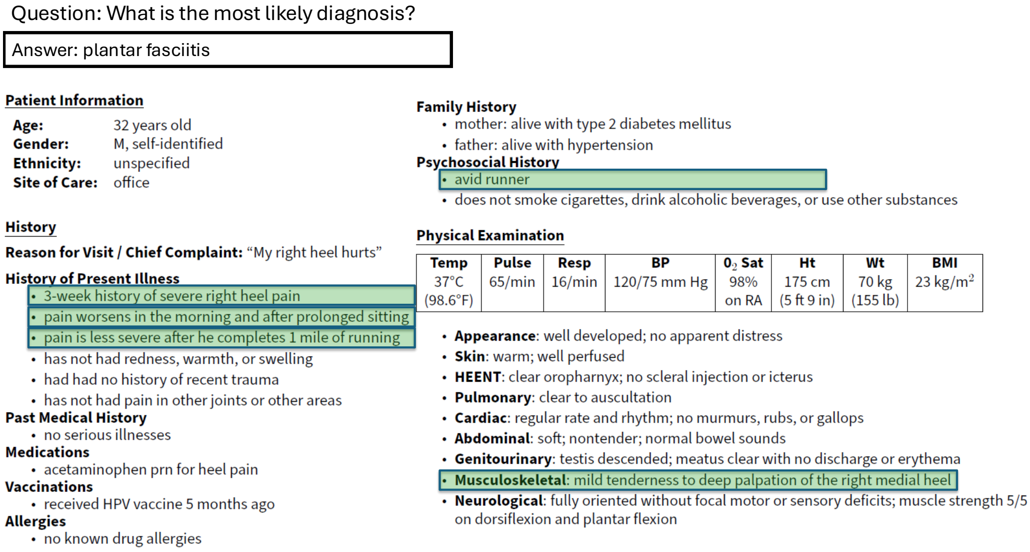
2.2 Setup

Our primary goal is to develop a scalable pipeline to generate chart items by prompting language models with detailed instruction and medical scenarios. Each prompt comprises a medical vignette presented as a multiple-choice medical question from the MedQA dataset along with three examples of chart items from the SHARP dataset.

We implement three generation frameworks using GPT-4: *Chain-of-Thought* generation, which transforms a medical vignette from MedQA into a chart item by creating a medical record for a hypothetical patient (Section 2.3); *Counterfactual Generation*, which incorporates counterfactual reasoning to explore alternative outcomes and generate novel items while leveraging an agent-based self-prompting strategy to create a knowledge base for accuracy (Section 2.4); and an information-theoretic framework where the sample items in few-shot learning settings are selected based on information-theoretic properties, finding that LLMs perform better with "difficult" examples (Section 2.5). For each generation method, we produced 80 items that were evaluated by two licensed medical experts (Section 3).

2.3 Experiment 1: Chain-of-Thought

The first experiment uses Chain-of-Thought (CoT) prompting as a baseline. The approach was designed to be a robust framework for systematically generating high-quality medical assessment items that works by dividing the creation process into a sequence of cognitively manageable steps (Saparov and He, 2022).


Question: What is the most likely diagnosis?

Answer: plantar fasciitis

Patient Information

Age: 32 years old
Gender: M, self-identified
Ethnicity: unspecified
Site of Care: office

History

Reason for Visit / Chief Complaint: "My right heel hurts"

History of Present Illness
- 3-week history of severe right heel pain
- pain worsens in the morning and after prolonged sitting
- pain is less severe after he completes 1 mile of running
- has not had redness, warmth, or swelling
- had had no history of recent trauma
- has not had pain in other joints or other areas

Past Medical History
- no serious illnesses

Medications
- acetaminophen prn for heel pain

Vaccinations
- received HPV vaccine 5 months ago

Allergies
- no known drug allergies

Family History
- mother: alive with type 2 diabetes mellitus
- father: alive with hypertension

Psychosocial History
- avid runner
- does not smoke cigarettes, drink alcoholic beverages, or use other substances

Physical Examination

Temp	Pulse	Resp	BP	O_2 Sat	Ht	Wt	BMI
37°C (98.6°F)	65/min	16/min	120/75 mm Hg	98% on RA	175 cm (5 ft 9 in)	70 kg (155 lb)	23 kg/m^2

- **Appearance**: well developed; no apparent distress
- **Skin**: warm; well perfused
- **HEENT**: clear oropharynx; no scleral injection or icterus
- **Pulmonary**: clear to auscultation
- **Cardiac**: regular rate and rhythm; no murmurs, rubs, or gallops
- **Abdominal**: soft; nontender; normal bowel sounds
- **Genitourinary**: testis descended; meatus clear with no discharge or erythema
- **Musculoskeletal**: mild tenderness to deep palpation of the right medial heel
- **Neurological**: fully oriented without focal motor or sensory deficits; muscle strength 5/5 on dorsiflexion and plantar flexion

Figure 1: An example chart-type item from Runyon et al. (2023). A chart item includes a chart with patient information, medical history, chief complaint, and physical examination findings, as well as an associated question and answer. Not all chart information is equally relevant for correctly diagnosing and test-takers must determine relevancy as part of the task. The green boxes highlight the most relevant information for diagnosis in this example.

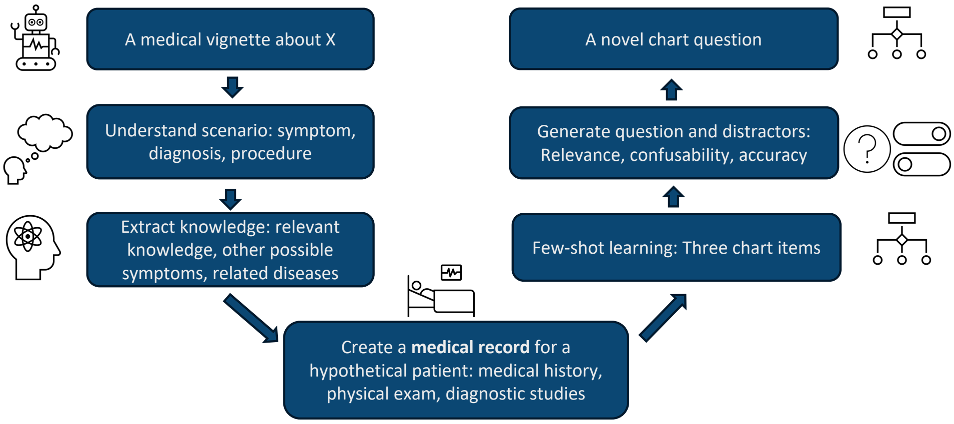


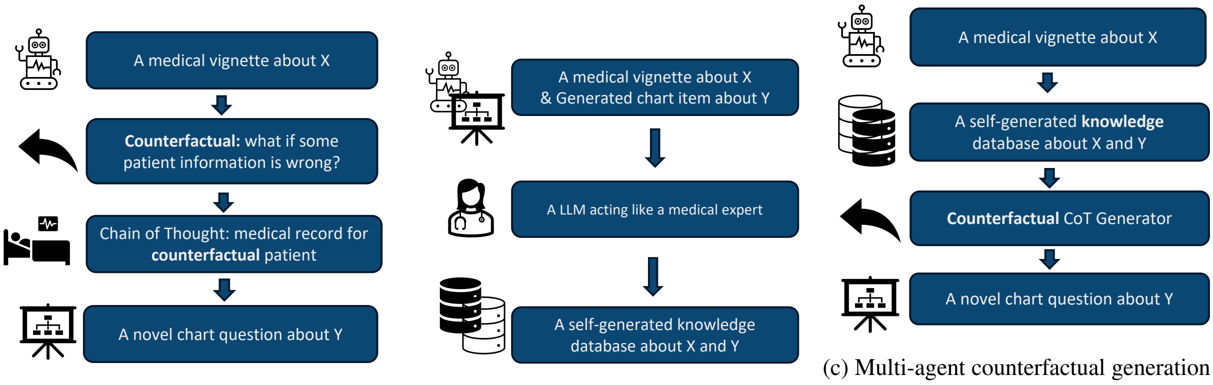
Figure 2: An illustration of Chain-of-Thought generation for chart-type items. The model is instructed to transform a simple medical scenario drawn from the MedQA dataset into a novel chart question step by step.

The CoT generator transforms a medical vignette extracted from the MedQA dataset into a chart question step by step (see Fig. 2). First, the model is instructed to identify the symptoms, diagnosis, and procedures described in the medical vignette to ensure that the model captures the parent medical scenario. Next, the model generates key knowledge relevant to the parent medical scenario, including key symptoms, potential differential diagnoses, and related diseases. The model then creates a detailed medical record for a hypothetical patient incorporating incorporating information from parent medical vignette and relevant information generated by the model. The model is further guided by referencing three sample SHARP chart items as examples of the desired chart-format output. The final output includes a clinical chart, a question with a correct answer and ten distractors. We instruct the model to adhere to some general principles for question and distractor generation (see Appendix A).

2.4 Experiment 2: counterfactual generation

A counterfactual chart item is one whose key diagnostic findings intentionally contradict the parent vignette's findings such that the correct diagnosis changes. It leverages a three-step process that integrates CoT prompting, counterfactual reasoning, and self-generated knowledge infusion (see Fig.3).

The first step focuses on generating content that differs from the source material by transforming



(a) Step 1: Counterfactual generation (b) Step 2: Knowledge generation (c) Multi-agent counterfactual generation with self-generated knowledge

Figure 3: An illustration of three-step knowledge-infused counterfactual generation for chart items. In Step 1, a chart item related to a novel medical scenario is generated by instructing the model to identify misinformation from the parent vignette. In Step 2, a knowledge database is generated for the parent and generated medical scenarios. In Step 3, the database is integrated with the counterfactual generator from Step 1 to regenerate the chart item.

the parent medical vignette into a different medical scenario through counterfactual reasoning. Counterfactual reasoning has been widely applied in various settings to explore alternative scenarios or causal inference in LLM performance (Qin et al., 2019; Zellers et al., 2019; Mostafazadeh et al., 2016; Meng et al., 2022; Rajani et al., 2019; Saparov and He, 2022; Frohberg and Binder, 2022; Elazar et al., 2021; Rudinger et al., 2020; Li et al., 2023). The model is instructed to conduct counterfactual reasoning in a Chain-of-Thought framework. We set up a counterfactual premise where the model is informed that certain elements of the parent vignette are transcribed incorrectly. Based on this counterfactual premise, the model needs to creatively "recover" the clinical chart, leading to a hypothetical patient record based on a new medical scenario. The model reasons based on the generated counterfactual record to develop a clinical assessment. The goal of this process is to generate content that deviates from the parent vignette while maintaining clinical plausibility.

The second step aims to improve the factual grounding of generated items by creating a self-generated knowledge base. This step addresses LLMs' tendency to hallucinate (Xu et al., 2024; Zhang et al., 2023) and is accomplished by initiating a new session in which the language model assumes the role of a medical expert. Agent-based prompting (Wu et al., 2024) enables the model to adapt to this role for generating medical knowledge. The correct answer from the parent vignette (X) and the generated correct answer from the counterfactual scenario (Y) are provided to the model. The task is to synthesize a detailed knowledge base about X and Y, including their symptoms, diagnostic criteria, and distinguishing features. This approach attempts to ground the generated counterfactual scenario in medical knowledge. In the final step, the knowledge base from Step 2 is integrated back into the counterfactual generation process. Combining the content variation from Step 1 with the knowledge grounding from Step 2, the model generates a refined chart item based on the counterfactual scenario.

2.5 Experiment 3: sample selection

We explore whether the performance of the language model is sensitive to the information-theoretic properties of the few-shot learning samples. Language model performance has been shown to depend on the quality of the samples in few-shot learning (Rasheed and Zarkoosh, 2024). Although all chart items used as examples were judged to be of high quality by human medical experts, certain information theoretic properties might make some examples better suited for the item generation task. In this experiment, we evaluate whether the quality of automatic generation is affected by the information content of the few-shot learning examples.

We hypothesize that the information-theoretic properties of example items are directly related to how challenging they are for the LLM to solve. Specifically, we use the surprisal of the correct answer given the question stem as a metric to assess an item's difficulty for the LLM. Surprisal is calculated as the negative logarithm of the probability that the LLM assigns to the correct answer given the question stem ($-\log p(\text{answer} \mid \text{stem})$), thereby quantifying how unexpected the correct answer is. Consequently, an item with relatively low surprisal is considered relatively easy for the LLM to answer correctly.

Sample selection is guided by two complementary hypotheses. The first posits that easier questions lead to better performance because they are straightforward for LLMs to mimic and regenerate (*Easy Sample Hypothesis*). Conversely, the second hypothesis suggests that more challenging examples may compel LLMs to engage in deeper reasoning, improving their ability to generate complex items (*Hard Sample Hypothesis*). By evaluating the effect of the surprisal of selected examples, we can maximize the quality of the generated items.

We calculate the surprisal of the correct answer using GPT-2 (Radford et al., 2019), and select the three sample items with the *lowest* surprisal. These selected samples are used in the multi-agent counterfactual generator with self-generated knowledge (Fig. 3). We then compare the performance with that of the counterfactual generator with randomly selected examples described in Section 2.4. If the items generated in Experiment 3 are considered of higher quality than items in Experiment 2, *Easy Sample Hypothesis* is supported.

3 Evaluation

While there is no consensus on the evaluation protocol of generated items (Circi et al., 2023), we aim to evaluate various aspects related to their practical use in assessment. The generated items may contain various flaws that affect their suitability for assessment. These flaws include, but are not limited to, clinical inaccuracies, contradictions, or hallucinations; incorrect designation of the correct answer; distractors (incorrect answers) that may actually be correct; or content that is unsuitable for assessment due to overly high or low complexity. Evaluating these issues requires review by human experts, as they cannot currently be assessed automatically.

Since an exhaustive list of all potential flaws could not be constructed a priori due to the unknown nature of AI-generated items, we focused our evaluation on the general suitability of these items for use in high-stakes medical education assessment as perceived by experts with both clinical and educational backgrounds. We designed a rubric that covered the following questions, with the full list provided in Appendix B:

(1) Can the chart stem be used on a high-stakes assessment?

(2) Please select up to 5 distractors that would, as a group, constitute a partial or full option set. Do not select any that would not be suitable for this chart, or that are too similar to others that have been selected as suitable.

(3) Can the chart item as a whole be used on a high-stakes assessment as currently written?

(4) Is this draft a usable starting point for writing or updating a chart item?

Two licensed medical doctors who also served as faculty at accredited medical schools in the United States were recruited. Each expert was assigned the same set of 100 automatically generated items, of which 33-34 were generated using each of the three methods (see Appendix C).

The results from the expert evaluation are presented in Table 1. Responses to each of the four rubric questions were dichotomized: (1) *stem quality*: minor changes / substantive changes; (2) *distractor quality*: substantive changes *not* required / substantive changes required; (3) *chart quality*: minor changes / substantive changes; and (4) *helpfulness*: helpful / *not* helpful. For each question, we calculate two success metrics: *strict*, which requires two favorable expert judgments and *loose*, which only requires one favorable judgment.

Across the three methods, both experts agreed that over 24% of generated stems required only "minor changes." In addition, the quality of the generated disractors was perceived to be high, with both raters agreeing that the distractors for at least 79% of the items required only minor changes. Across three methods, at least 85% generated chart items were considered usable with minor changes by at least one annotator. Although the CoT framework's items were deemed usable most often, the counterfactual framework performed similarly. The information-theory-based framework using sample items with lowest surprisal has a reduced performance compared to other methods. This suggests that language models' generation performance benefits more from examples with higher item surprisal, supporting the *Hard Sample Hypothesis*. Moreover, both experts agreed that over half of the items (52%) were helpful starting points for writing a new item. Here, items generated using

	Stem	Suggested distractors	Whole item	Helpfulness
CoT	**91% (35%)**	**100% (82%)**	**94% (26%)**	97% (79%)
Counterfact	**91% (32%)**	**100% (82%)**	91% (24%)	**100% (53%)**
Info theory	85% (24%)	97% (79%)	85% (15%)	100% (52%)

Table 1: This table displays the proportion of items that were favorably judged for each of the four questions in the annotation rubric. Two evaluation criteria were used: the *loose* criterion, where an item is considered favorably judged if at least one of the two participating physicians judged it favorably; and the *strict* criterion, where an item is favorably judged only if both physicians agreed. Proportions are presented in the format: loose (strict).

the CoT method significantly outperformed other items on that criterion with 79% of the CoT items judged to be helpful starting points by both experts. Overall, the expert evaluation suggests that approximately a quarter of the generated items were free from major flaws *at first assessment*, and half were regarded as useful starting points for item development compared to creating items from scratch.

The variability in expert agreement underscores the subjective nature of evaluating item quality, particularly for stems and charts, where the experts exhibited the most disagreement. The two experts had inter-annotator agreement of $\kappa = 0.1$ on chart stem quality. A qualitative inspection of the annotators' comments suggests that the low inter-rater agreement might be due to different conceptual understanding of the rubric. For example, both annotators commented that one question "needs mother's prenatal history", but one annotator considered this critical and suggested substantial changes needed, whereas the other considered it a minor modification (see Appendix E and F for more discussion on limitations and ethical considerations).

We also evaluated whether counterfactual generation produces items with greater semantic distance from their source material. To quantify semantic distance, we computed cosine similarity between word embeddings of each generated item and its parent vignette, with lower similarity indicating greater lexical/semantic divergence. Results showed that methods based on counterfactual generation (Exp 2 & 3) produced items with significantly lower cosine similarity to their parent vignettes than CoT generation (Exp 1), suggesting greater variation from the source material (see Appendix D).

4 Discussion

This study demonstrated the potential of LLMs to be used as automated assistive tools when writing items for medical assessments. The findings highlight key insights into the quality of the items generated across the three methods. Notably, over 24% of the generated stems were rated as requiring "minor changes" by both experts, with 85% of the items judged to require minor changes by at least one expert. This suggests that a significant portion of the generated items lack what could initially be considered irreparable flaws, inaccuracies, or contradictions. While this cannot yet be considered evidence that the items can be profitably used on an assessment without significant review and modifications, it is an encouraging initial assessment.

The integration of counterfactual reasoning and agent-based knowledge infusion showed effectiveness in producing content that differs more from source material. This suggests that tasking the model with identifying misinformation and generating counterfactual scenarios helps prevent the model from simply replicating existing data.

Of particular interest are the findings on information-theoretic sample selection, which highlight the nuanced role of item surprisal in few-shot learning. The observed differences in item generation when challenging examples were used suggest that example difficulty may influence LLM generation patterns. This insight underscores the importance of principled sample selection in optimizing LLM performance for automated item generation.

Future research should focus on automating the evaluation process, expanding applicability to other domains, and reducing the computational overhead of LLM-based pipelines. Integrating external knowledge sources, such as medical databases, could potentially improve the factual grounding of generated chart items. Retrieval-Augmented Generation techniques could be explored to access and incorporate external data during the item generation process. This approach might allow the model to generate more contextually informed items and better adapt to specialized knowledge domains.

References

Heidi Al-Wassia, Rolina Al-Wassia, Shadi Shihata, Yoon Soo Park, and Ara Tekian. 2015. Using patients' charts to assess medical trainees in the workplace: a systematic review. *Medical teacher*, 37(sup1):S82–S87.

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.

William CM Belzak, Ben Naismith, and Jill Burstein. 2023. Ensuring fairness of human-and ai-generated test items. In *International Conference on Artificial Intelligence in Education*, pages 701–707. Springer.

Susan M Case and David B Swanson. 1998. *Constructing written test questions for the basic and clinical sciences*. National Board of Medical Examiners Philadelphia.

Kuang Wen Chan, Farhan Ali, Joonhyeong Park, Kah Shen Brandon Sham, Erdalyn Yeh Thong Tan, Francis Woon Chien Chong, Kun Qian, and Guan Kheng Sze. 2024. Automatic item generation in various stem subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, page 100344.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in Education*, volume 8, page 858273. Frontiers Media SA.

Michelle Daniel, Joseph Rencic, Steven J Durning, Eric Holmboe, Sally A Santen, Valerie Lang, Temple Ratcliffe, David Gordon, Brian Heist, Stuart Lubarsky, et al. 2019. Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine*, 94(6):902–912.

Marie-France Deschênes, Nicolas Fernandez, Kathleen Lechasseur, Marie-Ève Caty, Busra Meryem Uctu, Yasmine Bouzeghrane, and Patrick Lavoie. 2025. Transformation and articulation of clinical data to understand students' clinical reasoning: a scoping review. *BMC Medical Education*, 25(1):52.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Filipe Falcão, Patrício Costa, and José M Pêgo. 2022. Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 27(2):405–425.

Jörg Frohberg and Frank Binder. 2022. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140.

Mark J Gierl and Hollis Lai. 2016. The role of cognitive models in automatic item generation. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, pages 124–145.

François Goulet, André Jacques, Robert Gagnon, Pierre Racette, and William Sieber. 2007. Assessment of family physicians' performance using patient charts: interrater reliability and concordance with chart-stimulated recall interview. *Evaluation & the health professions*, 30(4):376–392.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

B Guimarães, J Pais, E Coelho, A Silva, A Povo, I Lourinho, M Severo, and MA Ferreira. 2013. Assessing inter-rater agreement about item-writing flaws in multiple-choice questions of clinical anatomy. In *EDULEARN13 Proceedings*, pages 5921–5924. IATED.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, et al. 2024. Aloe: A family of fine-tuned open healthcare llms. *CoRR*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, and Sotirios Bisdas. 2023. The advent of generative language models in medical education. *JMIR Medical Education*, 9:e48163.

Yavuz Selim Kıyak and Emre Emekli. 2024. Chatgpt prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgraduate medical journal*, page qgae065.

Hollis Lai, Mark J Gierl, B Ellen Byrne, Andrew I Spielman, and David M Waldschmidt. 2016a. Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of dental education*, 80(3):339–347.

Hollis Lai, Mark J Gierl, Claire Touchie, Debra Pugh, André-Philippe Boulais, and André De Champlain. 2016b. Using automatic item generation to improve the quality of mcq distractors. *Teaching and learning in medicine*, 28(2):166–173.

Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 414–428.

Philseok Lee, Shea Fyffe, Mina Son, Zihao Jia, and Ziyu Yao. 2023. A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, 38(1):163–190.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Septian Eko Prasetyo, Teguh Bharata Adji, and Indriana Hidayah. 2020. Automated item generation: Model and development technique. In *2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 64–69. IEEE.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Areeg Fahad Rasheed and M Zarkoosh. 2024. Mashee at semeval-2024 task 8: The impact of samples quality on the performance of in-context learning for machine text classification. *Authorea Preprints*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Christopher R Runyon, Miguel A Paniagua, Francine A Rosenthal, Andrea L Veneziano, Lauren McNaughton, Constance T Murray, and Polina Harik. 2023. Sharp (short answer, rationale provision): A new item format to assess clinical reasoning. *Academic Medicine*, pages 10–1097.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *CoRR*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.

Ziwei Xu, Sanjay Jain, and Mohan S Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *CoRR*.

Victoria Yaneva, Peter Baldwin, Daniel P Jurich, Kimberly Swygert, and Brian E Clauser. 2023. Examining chatgpt performance on usmle sample items and implications for assessment. *Academic Medicine*, pages 10–1097.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. 2024a. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2024b. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A General Principles

We instruct the model to adhere to the following principles during generation: (a) *Informativity*: the new item should contain all the necessary information for a chart item; (b) *Accuracy*: the generated answer should be consistent with all of the information from the generated chart; (c) *Novelty*: the generated chart item should be sufficiently different from the parent item; and (d) *Validity*: the generated chart must include sufficient information to unambiguously identify the correct answer.

According to the chart question stem and correct answer, the model then crafts ten distractors—plausible but incorrect answer choices that are meant to be attractive to examinees who do not know the correct answer. We instruct the model to focus on the following properties during the distractor generation process: (a) *Relevance*: the distractors should be relevant to the chart question stem; (b) *Dissimilarity*: the distractors should not be synonyms or very similar to the correct answer; (c) *Incorrectness*: the distractors cannot be plausible correct answers for the generated chart question. Distractors with these characteristics enhance items' discriminative power.

The model is instructed to use descriptive language about any physical exam findings that follows patient chart documentation standards, such as specifying *warm*, *dry*, or *no rashes or lesions* instead of vague terms like *normal*.

B Evaluation Protocol

1) Evaluation of the Chart: Evaluate the Chart's suitability for use on a high stakes assessment. Minor changes are defined as the necessity to make minor changes to the chart including but not limited to: the addition, modification, or deletion of three or fewer minor history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level. Substantive changes entail an extensive rewrite of the chart and include but are not limited to: the addition, modification, or deletion of four or more substantive history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level.

Question: Can the chart be used on a high stakes assessment?

i. Yes, with some minor changes

ii. Substantive changes required, or the chart is too flawed to be useful

Note that if the expert selected "Substantive changes required", they would skip the next two questions and go directly to the fourth question on Helpfulness.

2) Selection of Appropriate Option Set: Please select up to 5 distractors that would, as a group, constitute a partial or full option set. Do not select any that would not be suitable for this chart, or that are too similar to others that have been selected as suitable. The N/A option should be used if you selected "Substantive changes required, or the chart is too flawed to be useful" in response to the above question about the associated chart.

i. N/A – Substantive changes required, or the chart is too flawed to be useful

ii. Distractor Suggestion 1

...

xi. Distractor Suggestion 10

3) Evaluation of the Chart Item as a whole (chart plus the option Set): To what extent can the chart item as a whole (i.e., chart plus the options set) be used on a high stakes assessment? Minor changes to the chart item as a whole is defined as the necessity to make minor changes to EITHER the chart (i.e., requires a minor rewrite of the chart including but not limited to: the addition, modification, or deletion of three or fewer minor history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level OR minor changes to the option set (i.e., the need to create one additional option to complete a sufficient option set of at least 4 options (preferably 5) with appropriate difficulty for a high stakes assessment). Substantive changes to the chart item as a whole is defined as the necessity to make substantive changes to EITHER the chart (i.e., requires an extensive rewrite of the chart including but not limited to: the addition, modification, or deletion of four or more substantive history/physical exam details to make the item more correct, realistic, or at a more appropriate difficulty level (i.e., suitable for high stakes assessment) OR substantive changes to the option set (i.e., the need to create three or more options to complete a sufficient option set of at least 4 options (preferably 5) with appropriate difficulty for a high stakes assessment). If EITHER the chart OR the option set need substantive changes, then this is considered as the need for substantive changes to the chart item as a whole. If BOTH the chart and the option set require minor changes, this is considered as the need for minor changes to the chart item as a whole.

Question: Can the chart item as a whole be used on a high stakes assessment as currently written?

i. Yes, with some minor changes

ii. Substantive changes required, or the chart is too flawed to be useful

4) Evaluation of helpfulness: Is this draft a usable starting point for writing or updating a chart item?

i. Yes, this draft would be helpful

ii. No, it would be easier for me to write an item from scratch

It is important to clarify that *we do not consider the "minor changes" category as suggesting an item is ready for assessment without significant additional work* (see Section F for discussion on ethical considerations). Instead, the distinction between minor and substantive changes serves as a simple way to differentiate items with major flaws from those with flaws that may be fixable.

C Recruitment

To perform this evaluation, two licensed medical doctors who also served as faculty at accredited medical schools in the United States were recruited. The recruitment was performed by ANONYMIZED INSTITUTION's Assessment Alliance, which engages with educators, learners, and other members of the health profession's education community to identify how to best prepare medical professionals to safely care for a diverse patient population.

Once recruited, the human experts were invited to a kickoff meeting, where they were briefed on the purpose of the experiment and the evaluation rubric, instructed on the use of the annotation platform (items were displayed using the John Snow Labs annotation system), and given an opportunity to ask questions. Following this meeting, the experts were given two weeks to complete their annotations. Each expert was assigned the same set of 100 automatically generated items, of which 33-34 were generated using each of the three methods described in Section 2.

D Automated evaluation of item variation

An important consideration for newly generated items is the extent to which they differ from their source material. Understanding these differences can help identify which generation methods produce more varied content and potentially guide selection of items for further development by human item writers. To this end, we explore the use of cosine similarity between word embeddings of generated items and their parent medical vignettes as one measure of content variation.

Cosine similarity between word embeddings quantifies lexical and semantic overlap between generated and parent items, with lower values indicating less overlap—i.e., greater textual divergence. We define an experimental group where cosine similarity is calculated between each generated item and its corresponding parent vignette. This is compared against a baseline group, where cosine similarity is computed between each generated item and a random non-parent vignette from the same set of parent vignettes. Figure 4 shows the average cosine similarity in experimental and baseline groups across the three generation methods.

Paired t-tests between experimental and baseline groups within each method did not reveal statistically significant results (all $p > 0.25$), possibly due to high variability in similarity values or limited sample size. Nevertheless, we observed consistent trends across all conditions, where similarities between generated items and their parent vignettes were not significantly different from similarities with unrelated vignettes. To quantify relative content variation across methods, we used the difference in cosine similarity between experimental and baseline groups as an index of textual divergence. A second set of paired t-tests with Bonferroni correction was conducted to compare this divergence index across generation methods. The results revealed that CoT generation produced significantly smaller divergence from source material than both counterfactual ($t = 4.64$, $p < 0.001$) and information-theory-based generation ($t = 4.41$, $p < 0.001$). No significant difference was found between counterfactual and information-theory-based methods ($t = -0.1$, $p = 0.91$).

These findings suggest that counterfactual and information-theoretic approaches produce content with greater lexical and semantic distance from their source vignettes compared to CoT generation. However, it is important to note that cosine similarity captures only surface-level textual differences and does not necessarily reflect clinically meaningful variation or educational value of the generated items.

E Limitations

A key limitation for this research is the fact that the evaluation relied on only two human raters. These raters had not undergone specific training in item writing for high-stakes clinical exams, and this was their first time evaluating AI-generated items. These factors may have contributed to the observed variability in their judgments while limiting their generalizability. Additionally, given the well-documented variability in how human experts write clinical MCQs (e.g., Guimarães et al., 2013), judgments about the need for "minor" vs "substantive" changes may reflect subjective differences in opinion rather than a definitive standard of quality.

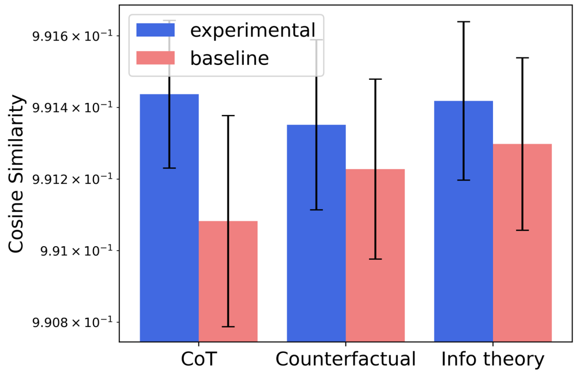


Figure 4: Average cosine similarity in experimental and baseline groups across three generation methods. Blue bars represent cosine similarity between generated item and its corresponding parent vignette. Red bars represent cosine similarity between generated item and a random vignette from a set of parent vignette.

Rater performance may have been further influenced by biases such as social desirability or confirmation bias. Social desirability bias could lead raters to align their evaluations with perceived research goals or provide overly favorable feedback due to the novelty of AI in clinical item generation. Confirmation bias might cause raters to focus on strengths or weaknesses based on their pre-existing beliefs about AI's capabilities. Measuring attitudes toward AI as part of the recruitment process is an area for improvement in future research.

In terms of evaluation design, the rubric was purposefully broad given the stage of this research and did not account for specific flaws that might arise in clinical MCQs. Examples of such flaws include susceptibility to "testwiseness," which refers to an examinee's familiarity with general test-taking strategies, and "construct-irrelevant difficulty," which refers to item features that increase an item's difficulty for reasons unrelated to the trait that is the intended target of the assessment (Case and Swanson, 1998). Future research should endeavor to better understand and identify specific flaws that may be prevalent within AI generated items, and facilitate their evaluation through more granular rubrics.

Similar to the human evaluation, the automated evaluation also suffered limitations stemming from the preliminary nature of this study. While a useful approximation of the differences that exist between items, cosine similarity focus only on relative item variation and do not guarantee that items are sufficiently novel for a given application.

Last but not least, the performance of the generated items in practical settings is currently unknown. Key metrics such as the extent to which examinees find an item difficult, the power of an item to discriminate between examinees of different proficiency levels, and examinee perceptions of clarity require pretesting with an examinee sample and remain untested at this stage.

In summary, future research should not only focus on improving the technical components of item generation, but also include larger-scale evaluations, enhanced rubrics, qualitative analyses, the utilization of raters trained in item writing, and the collection of examinee response data in real-world assessment settings.

F Ethical considerations

As AI continues to evolve and its application is extended to more domains, its integration into item development raises important ethical considerations. A key concern is ensuring that AI-generated items meet the necessary quality standards for a given type of assessment. While AI can generate item drafts, these items must be thoroughly reviewed by expert item writers to ensure that they are appropriate, clinically accurate, and meet the intended learning or assessment objectives. Human oversight remains essential to finalize each item, and AI-generated content should undergo the same rigorous review processes as items that are written without AI assistance.

The use of AI also requires clear accountability and transparency in the development process and avoidance of over-reliance on technology. While AI can assist in generating drafts, the final responsibility for ensuring the quality, fairness, and ethical use of any test item remains with human experts. It is crucial to maintain transparency about how AI is used and to ensure that stakeholders are aware of both the capabilities and limitations of AI in this context.

By ensuring that human expertise remains central to the item development process, establishing rigorous review procedures, and maintaining transparency and accountability, AI can be used ethically and responsibly to support the creation of high-quality assessment items.

Using Whisper Embeddings for Audio-Only Latent Token Classification of Classroom Management Practices

Wesley Morris, Jessica Vitale, and Isabel Arvelo
Vanderbilt University

Abstract

In this study, we developed a textless NLP system using a fine-tuned Whisper encoder to identify classroom management practices from noisy classroom recordings. The model segments teacher speech from non-teacher speech and performs multi-label classification of classroom practices, achieving acceptable accuracy without requiring transcript generation.

1 Introduction

Positive and proactive classroom management establishes a foundation for equitable and inclusive environments where all students can learn. Research demonstrates that effective classroom management increases student engagement and academic achievement, particularly for students with learning and behavioral differences [1]. Despite identifying evidence-based classroom management practices, a significant implementation gap exists in their consistent classroom application [25]. Teachers often report feeling underprepared to support student behavior and express a need for ongoing, job-embedded professional development to implement practices effectively. Coaching and observational feedback improve teachers' classroom management practices and enhance their self-efficacy, reducing stress and mitigating burnout [19, 30]. However, these traditional approaches are resource-intensive and difficult to scale, particularly in historically marginalized communities. Advances in natural language processing (NLP) and machine learning present an innovative opportunity to address these challenges. Automated feedback tools can deliver frequent, timely, and actionable insights to teacher practice, bridging the gap between evidence-based practices and their real-world implementation, providing accessible professional development at scale.

Current automated feedback tools for teacher classroom practices rely solely on transcripts generated by Automatic Speech Recognition (ASR) tools. However, teacher affect, including tone and delivery, is critical in shaping positive student-teacher interactions, fostering social-emotional learning, and reinforcing classroom expectations [15]. Research indicates that transcription alone often fails to capture these suprasegmental speech features, resulting in losing vital information about prosody and intonation [26]. To address this limitation, we are developing a Multimodal Automatic System for the Classification of Teacher Classroom Practices (MASCoT-CP) to automatically detect classroom management practices using both audio and text-based data. This system aims to provide teachers with actionable insights into their practices, leveraging multi-modal inputs to enhance the feedback they receive. Unlike current automated feedback tools that rely exclusively on text-based transcript analysis, MASCoT-CP incorporates prosody, intonation, and affect, key elements of spoken language essential for understanding the nuances of classroom culture and teacher-student interactions.

This study presents findings from the audio-only component of the MASCoT-CP system. This component, designed as part of a larger, multi-modal system that will integrate audio and text transcripts, serves two purposes: diarizing classroom audio into teacher speech and non-teacher speech segments, and generating predictions about classroom management practices present within those segments. Future research will

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 154–162
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

integrate the output of the audio-only model with a text classification model to create an ensemble system that enhances classification accuracy. This comprehensive approach will provide teachers with fine-grained feedback on their classroom practices, allowing them to focus on refining specific elements of their practices, thereby enhancing their students' learning experiences.

2 Background

2.1 Classroom Management Practices

Classroom management includes the strategies and practices teachers implement to establish and maintain structured, supportive learning environments. Research consistently demonstrates that effective classroom management is fundamental to maximizing instructional time, sustaining student engagement, and building positive student-teacher relationships. Systematic reviews identify several evidence-based practices that contribute to successful classroom management, particularly frequent opportunities for active student engagement and feedback for student behaviors [5, 8].

Central to effective classroom management are opportunities to respond (OTRs), questions or prompts that elicit student participation. Research shows that high rates of OTRs help sustain student engagement, increase on-task behaviors, and improve accuracy in student responses [7]. Complementing these engagement strategies, teacher feedback further shapes student behavior. Feedback typically falls into two categories within classroom management: reinforcing appropriate behavior through positive feedback (such as specific praise) and addressing inappropriate behavior through redirections or corrective responses. Evidence indicates that delivering specific praise and maintaining a positive ratio of positive to corrective interactions strengthens student-teacher relationships and increases students' on-task behaviors [2, 9]. Together, these practices create positive classroom environments that establish a foundation necessary for effective academic instruction.

Despite strong evidence supporting classroom management's impact on student outcomes, many teachers face challenges in consistently implementing these practices. Pre-service teacher preparation programs often provide limited training in classroom management [12], leading teachers to identify it as one of the most challenging aspects of their job and a primary factor contributing to teacher attrition [13, 28]. These implementation challenges underscore the need for effective professional development. Traditional approaches to supporting teacher development, such as coaching and observational feedback, have effectively improved practice implementation. However, scaling these support presents logistical and financial barriers due to time and resource constraints. Recent advances in NLP technologies offer promising solutions for addressing these scalability challenges. NLP tools capable of analyzing classroom discourse and generating automated feedback represent an emerging approach to supporting teaching practices at scale [10, 15].

Multiple research teams have developed text-based classification models using transformer architectures to analyze classroom transcripts. These studies demonstrate the feasibility of automated classroom discourse analysis across different instructional contexts and pedagogical practices. Alic et al. [1] fine-tuned a RoBERTa-based model with paired teacher-student utterances for binary classification of focusing questions, achieving an F1 score of 0.501. Suresh et al. [24, 25] trained a RoBERTa-base model to classify teacher utterances into one of ten math talk moves, incorporating surrounding transcript lines as context, and achieved an average F1 score of 0.79. Similarly, Jensen et al. [17] fine-tuned BERT to classify seven discourse-related teaching practices, obtaining an average area under the curve (AUC) of 0.84 across classifications.

2.2 Audio Classification

The studies mentioned above analyzed transcripts of teacher speech, rather than classifying directly from audio. Unlike text data, which consists of discrete words and subwords easily tokenized through dictionary lookup, audio data presents as a continuous information stream. While previous research has used feature engineering approaches to extract information from classroom audio [11, 16, 23], the current study uses a modified form of token classification approach that converts raw audio into latent token embeddings. Whisper [20], developed by OpenAI, is a sequence-to-sequence transformer model for automatic speech recognition (ASR). In the original architecture, the encoder's final hidden state feeds into a decoder block that recursively generates text conditioned on both the encoder's final hidden state and previously generated tokens. The model was trained on 680,000 hours of speech with transcripts, including 117,000 hours in 96 non-English languages. As a result, Whisper achieves strong results in ASR and translation tasks [20].

Recent interest in textless NLP has focused on directly extracting semantic information from the audio without intermediate transcription [14]. Although designed for ASR and translation, multimodal sequence-to-sequence models show promise for audio classification tasks. Ma et al. [19] fine-tuned Whisper to generate label tokens, effectively performing zero-shot audio sound event classification. Classification can also be performed by separating the Whisper encoder block and using the final hidden state embeddings directly, as demonstrated in predicting speech disorders such as dysarthria [21] and stuttering [3]. In this audio classification approach, the encoder's final hidden states pass through a projection layer into a classification head that generates sequence predictions.

2.3 Current Study

In this study, we develop an audio-only tool that identifies classroom management practices in teacher speech segments. Our approach uses a three-state process using a modified Whisper architecture. First, we detach the Whisper encoder from the decoder and fine-tune it for latent token classification, similar to text-based NLP token classification, to predict the most probable teaching practice in each 0.02-second audio window. Second, we use these predictions to differentiate segments containing teacher speech from non-teacher speech segments. Finally, we use the predictions from the Whisper encoder to perform multi-label classification on teacher speech segments to identify which specific classroom management practices are present. The study addresses two primary research questions:

RQ1: Can an audio-only model accurately distinguish between teacher and non-teacher speech in elementary classroom recordings?

RQ2: Can an audio-only model accurately identify classroom management practices present within teacher speech segments from elementary classroom recordings?

3 Methods

3.1 Dataset

The dataset used to train the classification model included 29.91 hours of audio recordings from 131 classroom sessions. The recordings were collected from 28 teachers (15 general education, 13 special education) across kindergarten through 4th-grade classrooms. The sample included 6 male and 22 were female teachers. Teachers self-identified as White (n=16), Black (n=7), Latinx (n=4), and Biracial (n=1). Their average teaching experience was 11 years (range = 1-30). Each teacher contributed 4 to 5 recordings to the dataset. The recordings from special education teachers primarily consisted of small-group interventions, while general education teachers recorded themselves conducting whole-group instruction with an average of 21 students per class.

The audio recordings were annotated for 10 specific teaching practices and two non-teacher talk labels, organized into 6 broader categories related to classroom management. The six categories include instructional talk, social talk, positive teacher-student interactions (i.e., specific praise, general praise, and affirming correct student responses), negative teacher-student interactions (i.e., reprimands, redirections, and correcting incorrect student responses), opportunities to respond (OTRs) (i.e., academic and social demands and questions) and non-teacher speech (e.g., student talk and prolonged instances of silence).

Each audio file was annotated by trained labelers using Audacity [4], where labelers listened to the complete recording and noted each segment's start and end times. This approach allowed us to establish ground-truth boundaries for each segment, enabling us to compare multiple diarization tools and align with methods used in systematic directional observation of classrooms [18, 29]. Since spoken language in classrooms does not follow traditional written sentence structures, annotators applied two stop rules to determine segment boundaries: a shift to a new practice category (e.g., a teacher transitioning from providing instructional talk to asking a question, signaling an opportunity to respond) or silence lasting at least two seconds (e.g., a teacher pausing mid-instructional talk to think). Table 1 displays the count of each classroom practice in the full dataset as well as aggregate statistics about their durations. To ensure reliability, each recording was annotated by two independent labelers, followed by consensus coding meetings to resolve discrepancies. Inter-rater agreement (IRA) was calculated using the Multi-Option Observation System for Experimental Studies (MOOSES) [26], with agreement defined as both labelers identifying the same practice within a two-second window. The average IRA across the 131 recordings was 74%, with most disagreements occurring around segment start and end times rather than label assignment.

		duration (s.)	
	count	mean	std
Instructional Talk	5318	6.65	9.2
Social Talk	2477	3.52	3.8
Positive Interactions	2270	2.76	2.2
Corrective Interactions	981	3.25	2.7
Opportunity to Respond	5749	2.99	2.1
Non-teacher	9102	2.78	3.4

Table 1: Counts and durations of each classroom management practice category

3.2 Training

We used the encoder stack of Whisper base [20], as the foundation for a custom audio latent token classification model. Figure 1 illustrates the architecture of our modified version of the Whisper encoder stack with the shapes of embedding matrices listed on the bottom. The Whisper preprocessor first uses fast Fourier transforms that generate 80-channel log-mel spectrograms from 30 second segments of raw audio using 16 kHz sampling, 25ms window length, and 10ms stride. These spectrograms serve as input to two convolutional layers with a filter width of 3 and GELU activation function. The first layer maps the 80 spectrogram channels to embedding dimension d = 768. The second layer uses a stride of 2 to reduce the 3,000 windows to T = 1,500 latent token embeddings, each spanning 0.02 seconds. Sinusoidal position embeddings are then added to produce the final T×d dimensional hidden states $\mathbf{h}_{1,2\ldots L} \in \mathbb{R}^{T\times d}$ that define each layer's embedding dimensionality in the encoder.

We first removed all audio files from nine (31%) of the teachers as a hold-out test set to ensure that the model generalizes to speakers outside of its training set. We then split each audio file into thirty-second clips with a fifteen-second overlap so that the model would be exposed to all audio twice

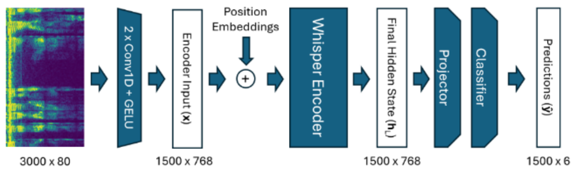


Figure 1: Architecture of the latent token classification model with dimensionality of matrices

per epoch except for the first and last fifteen seconds of the audio file. We included this overlap to ensure that each timestep had at least 15 seconds of previous context to inform the classification. Thus, the final hidden state of the Whisper encoder had a dimensionality of $\mathbf{h}_L \in \mathbb{R}^{T\times d}$ where T is the total number of time steps (i.e. 1,500) and d is the embedding dimensionality (i.e. 768). On top of the Whisper encoder block, we applied linear layers for token classification. The first, a projector, reduced the dimensionality from 768 to 256 and applied a ReLU activation function. Finally, our classification head further reduced the dimensionality to six, our number of labels k, with a sigmoid activation function. Therefore, the output of the model had a dimensionality of $\hat{\boldsymbol{y}} \in \mathbb{R}^{T\times k}$.

We used the Whisper encoder's output to create target labels for training. For each 30-second audio clip, we generated 1,500 target labels by mapping the original hand-annotated labels to each of the 1,500 timesteps t. At each timestep, we identified the predominant label from the annotations. The model's predictions were then compared against these labels using cross-entropy loss. We fine-tuned the model for six epochs using the AdamW optimizer. Following the specifications from the original Whisper training [20], we used a learning rate of 3.75e-05 and a weight decay of 0.1.

3.3 Diarization

Our first goal was to correctly distinguish segments of audio where the teacher was speaking from segments of audio where the teacher was not speaking (e.g. student speech, silence). For inference, we first split the audio in the test set into thirty-second clips, overlapping with a step of fifteen seconds, as during training. We then used our model to generate logits for each 0.02-second window. Because of our method of splitting the audio files into overlapping clips, all audio in a file aside from the first and last 15 seconds is analyzed twice. We therefore calculate final logits for each

0.02-second window as the mean of the two predictions. Finally, we take the maximum logit for each window to determine the predicted class. If the class is predicted to be anything other than one of the classroom management practices, then we classify it as non-teacher speech. Any classroom management practice was classified as teacher speech. We evaluated our success using a modified diarization error rate (mDER), defined as:

$$mDER = \frac{Seconds\ of\ misclassified\ audio}{Total\ seconds}$$

Due to the "noisy" nature of elementary classroom environments, we noticed occasional very short segments. To address this, we implemented a minimum speaker turn length, merging segments shorter than a certain threshold with adjacent speech. We empirically determined the optimal threshold length, by assessing mDER at minimum length thresholds between 0.1 and 0.8. This threshold optimization was conducted exclusively on the training set to prevent information leak into the test set.

Finally, we tested our diarization method on the withheld test set, comparing our results against other open-source diarization tools including Pyannote [6] and SpeechBrain [22]. Unlike traditional diarization models that precisely mark the start and stop boundaries of speech and silence, our labeling scheme captures higher level speaker turns. For example, if a teacher pauses briefly during a classroom practice and then continues, our label extends across the entire segment rather than breaking it at the silence. This distinction is particularly important when comparing our approach to diarization tools designed to detect precise speech boundaries. These models segment speech with frequent breaks and allow for speaker overlap, which is not possible in our framework. Because our evaluation metric is based on non-overlapping, high-level speaker turns, other diarization models may be penalized under our modified DER, even when they have correctly identified what occurred in the audio. For our use case, where we aim to broadly classify whether a given segment of audio represents teacher speech or non-teacher speech, diarization serves primarily as a necessary preprocessing step rather than an end goal. Our segmentation approach is well-suited for our application because it reduces noise from minor pauses, interruptions, or overlapping speech that are not critical to our analysis.

3.4 Classification

After diarizing the audio into teacher speech and non-teacher speech, we used the logits computed by the classification tool to identify all teacher classroom practices present in each segment of audio. Each segment was assigned a vector $\hat{\mathbf{y}} \in \mathbb{R}^{K\times 1}$ where K is equal to the number of classes. If the model predicted the label for any of the 0.02-second windows within that segment, its value was predicted as 1, otherwise it was predicted as 0. Similarly, if a label k was present in a segment of the target dataset, $\mathbf{y}_k = 1$ otherwise 0. We evaluated success by calculating precision, recall, and f1 scores for each of the classes across all the segments.

4 Results

4.1 Diarization Results

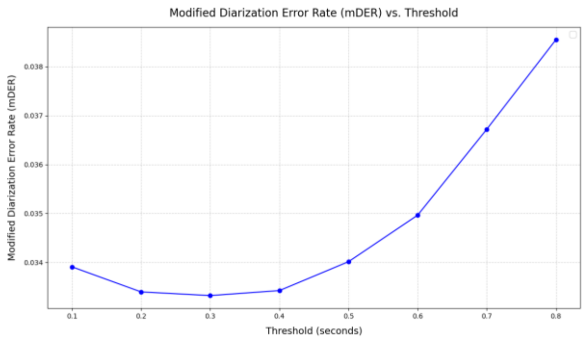


Figure 2: Identifying optimal maximum segment length for audio segmentation

Tool	Total mDER	Teacher mDER	Non-Teacher mDER
MASCoT-CP	0.086	0.06	0.149
Pyannote	0.264	0.29	0.196
SpeechBrain	0.324	0.28	0.438

Table 2: Modified Diarization Error Rate for MASCoT-CP vs. other diarization systems

We first attempted to determine the optimal minimum segment size. As Figure 2 shows, we found 0.3 seconds to be the optimal minimum segment length, and used this parameter for all further experiments. Using the minimum segment length on our test set, we found an mDER of 0.086, indicating that 8.6% of all audio segments were misclassified. This outperformed other open source diarization tools such as Pyannote (mDER = 0.264) and SpeechBrain (mDER = 0.324). However, it should be noted that our study does not take other elements of diarization into account, such as voice overlap and speaker identification. Table 2 shows results from the three diarization tools.

4.2 Classification Results

Practice Category	n	prec.	recall	F1
Instructional Talk	2,011	0.627	0.509	0.562
Social Talk	1,222	0.334	0.548	0.415
Positive Interactions	757	0.556	0.528	0.542
Corrective Interactions	512	0.179	0.221	0.198
Opportunity to Respond	2,194	0.637	0.528	0.577
Non-teacher	4,360	0.875	0.55	0.675
mean	**1,843.7**	**0.535**	**0.481**	**0.495**

Table 3: Counts in test set and metrics for each classroom practice

Once we had separated each audio file into teacher-speech and non-teacher-speech segments, we generated labels for each segment according to whether a classroom practice was predicted in each segment. Table 3 shows precision, recall, and f1 scores for each classroom practice, as well as the number of occurrences of each classroom practice in the test set. The model's classification F1 scores were above 0.4 for all classroom practices aside from corrective, which may be a result of the low prevalence of this practice. However, while praise had a similarly low prevalence, the model was much more likely to identify this classroom practice correctly (F1 = 0.542).

5 Discussion

In this study, we developed an audio-only tool which uses a fine-tuned version of the Whisper base model's encoder stack to segment and classify teacher speech for the classroom management practices. We fine-tuned the model on a dataset of almost 30 hours of classroom audio annotated by expert raters for the start and end times of classroom management practices. Finally, we process the output of the model to identify segments of teacher speech and classify the classroom management practices in those segments. This study demonstrates that models can be trained to identify classroom practices with reasonably performance levels without access to text transcripts.

Our model effectively distinguishes between teacher speech and non-teacher speech, achieving a low misclassification rate of 8.6% - a significant improvement over other open-source diarization models. However, it is important to note that other diarization models are not specifically tuned for this task or classroom contexts. Our approach differs from traditional diarization methods, which precisely segment speech boundaries and capture overlapping speakers. Regardless, our results suggest that our model is well-suited for automatic identification of teacher speech in classroom recordings without requiring prior training on individual teacher voices, making it a practical alternative to traditional diarization tools when the goal is classification of classroom discourse rather than precise speaker diarization.

For classification performance, our tool attained F1 values between 0.4 and 0.7 for all but one teaching practice. The lowest F1 score of 0.2 occurred for corrective interactions, likely due to the limited representation of this class in the training dataset. With only 981 instances (3.8% of the training dataset), correctives were the least frequent classroom practice we labeled, potentially limiting the model's ability to learn robust patterns for this category. While our classification accuracy was lower than that of previous studies, reporting F1 scores between 0.79 and 0.84 for multi-class classification of teacher discourse moves [17, 25], it is important to note that prior work relied on hand-transcribed textual data. In contrast, our study uses raw, noisy, audio-only data.

Our study was principally limited by the relatively small sample size of only 30 hours from 28 teachers. We need to train our model on a larger and more diverse labeled dataset to develop a tool that generalizes effectively across diverse linguistic environments. Additionally, while our results are promising, given that they are derived directly from audio in naturally noisy classroom recordings, they lag behind studies using clean text transcripts. One potential solution is to integrate this audio model into a larger multi-modal ensemble model that leverages audio and transcripts to achieve higher accuracy in identifying classroom practices.

6 Conclusion

In this study, we trained the encoder block of the Whisper to predict classroom management practices in small time windows of teacher speech. We then used these predictions for two purposes: segmenting audio into teacher speech and non-teacher speech segments with high accuracy and predicting which classroom management practices were present in the segments with reasonably high performance. These results demonstrate that it is possible to classify classroom management practices using textless NLP methods, even in noisy classroom recordings.

While observation and feedback are established methods for supporting teacher development, their implementation is resource-constrained, particularly in under-resourced educational settings. Automatically identifying teaching practices from authentically noisy audio recordings can allow teachers to reflect and improve their use of effective classroom management practices. This can have significant downstream effects on students' educational experiences, particularly those with learning differences and those in under-resourced settings.

This study contributes to advancements in textless NLP and automated measurement of classroom practices. Future research will build on the audio-only model by integrating it with a text-based classification approach using ASR-derived transcripts, forming a multi-modal automatic system for classifying classroom management practices (MASCoT-CP). By combining transcript analysis with prosodic and intonational features from the audio-only model, we anticipate improved accuracy in predicting teaching practices. This potentially enhanced measurement capability could be a foundation for developing automated feedback tools that provide teachers with data-driven insights into their classroom management strengths and areas for reflection and growth.

References

[1]Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H. and Jurafsky, D. 2022. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. arXiv.

[2]Allday, R.A., Hinkson-Lee, K., Hudson, T., Neilsen-Gatti, S., Kleinke, A. and Russel, C.S. 2012. Training General Educators to Increase Behavior-Specific Praise: Effects on Students with EBD. *Behavioral Disorders*. 37, 2 (Feb. 2012), 87–98. DOI:https://doi.org/10.1177/019874291203700203.

[3]Ameer, H., Latif, S., Latif, R. and Mukhtar, S. 2023. Whisper in Focus: Enhancing Stuttered Speech Classification with Encoder Layer Optimization. arXiv.

[4]Audacity Team 2014. Audacity(R): Free Audio Editor and Recorder.

[5]Brandi Simonsen, Sarah Fairbanks, Amy Briesch, Diane Myers, and George Sugai 2008. Evidence-based Practices in Classroom Management: Considerations for Research to Practice. *Education and Treatment of Children*. 31, 1 (2008), 351–380. DOI:https://doi.org/10.1353/etc.0.0007.

[6]Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W. and Gill, M.-P. 2020. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona, Spain, May 2020), 7124–7128.

[7]Cavanaugh, B. 2013. Performance Feedback and Teachers' Use of Praise and Opportunities to Respond: A Review of the Literature. *Education and Treatment of Children*. 36, 1 (2013), 111–136. DOI:https://doi.org/10.1353/etc.2013.0001.

[8]Conroy, J., Hulme, M. and Menter, I. 2013. Developing a 'clinical' model for teacher education. *Journal of Education for Teaching*. 39, 5 (Dec. 2013), 557–573. DOI:https://doi.org/10.1080/02607476.2013.836339.

[9]Conroy, M., Sutherland, K., Snyder, A., Al-Hendawi, M. and Vo, A. 2009. Creating a Positive Classroom Atmosphere: Teachers' Use of Effective Praise and Feedback. *Beyond Behavior*. 18, 2 (2009), 18–26.

[10]Demszky, D., Liu, J., Hill, H.C., Sanghi, S. and Chung, A. 2025. Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*. 227, (Apr. 2025), 105183. DOI:https://doi.org/10.1016/j.compedu.2024.105183.

[11]Donnelly, P.J., Blanchard, N., Olney, A.M., Kelly, S., Nystrand, M. and D'Mello, S.K. 2017. Words

matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (Vancouver British Columbia Canada, Mar. 2017), 218–227.

[12]Freeman, J., Simonsen, B., Briere, D.E. and MacSuga-Gage, A.S. 2014. Pre-Service Teacher Training in Classroom Management: A Review of State Accreditation Policy and Teacher Preparation Programs. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*. 37, 2 (May 2014), 106–120. DOI:https://doi.org/10.1177/0888406413507002.

[13]Gable, R.A., Tonelson, S.W., Sheth, M., Wilson, C. and Park, K.L. 2012. Importance, Usage, and Preparedness to Implement Evidence-based Practices for Students with Emotional Disabilities: A Comparison of Knowledge and Skills of Special Education and General Education Teachers. *Education and Treatment of Children*. 35, 4 (2012), 499–520. DOI:https://doi.org/10.1353/etc.2012.0030.

[14]Haghani, P., Narayanan, A., Bacchiani, M., Chuang, G., Gaur, N., Moreno, P., Prabhavalkar, R., Qu, Z. and Waters, A. 2018. From Audio to Semantics: Approaches to End-to-End Spoken Language Understanding. *2018 IEEE Spoken Language Technology Workshop (SLT)* (Athens, Greece, Dec. 2018), 720–726.

[15]Jacobs, J., Scornavacco, K., Clevenger, C., Suresh, A. and Sumner, T. 2024. Automated feedback on discourse moves: teachers' perceived utility of a professional learning tool. *Educational technology research and development*. 72, 3 (Jun. 2024), 1307–1329. DOI:https://doi.org/10.1007/s11423-023-10338-6.

[16]James, A., Kashyap, M., Victoria Chua, Y.H., Maszczyk, T., Nunez, A.M., Bull, R. and Dauwels, J. 2018. Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (Wollongong, NSW, Dec. 2018), 983–988.

[17]Jensen, E., L. Pugh, S. and K. D'Mello, S. 2021. A Deep Transfer Learning Approach to Modeling Teacher Discourse in the Classroom. *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine CA USA, Apr. 2021), 302–312.

[18]Ledford, J.R. and Gast, D.L. 2024. *Single Case Research Methodology: Applications in Special Education and Behavioral Sciences*. Routledge.

[19]Ma, R., Liusie, A., Gales, M.J.F. and Knill, K.M. 2023. Investigating the Emergent Audio Classification Ability of ASR Foundation Models. arXiv.

[20]Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*. (2022), 28.

[21]Rathod, S., Charola, M. and Patil, H.A. 2023. Noise Robust Whisper Features for Dysarthric Severity-Level Classification. *Pattern Recognition and Machine Intelligence*. P. Maji, T. Huang, N.R. Pal, S. Chaudhury, and R.K. De, eds. Springer Nature Switzerland. 708–715.

[22]Ravanelli, M. et al. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv.

[23]Schlotterbeck, D., Uribe, P., Araya, R., Jimenez, A. and Caballero, D. 2021. What Classroom Audio Tells About Teaching: A Cost-effective Approach for Detection of Teaching Practices Using Spectral Audio Features. *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine CA USA, Apr. 2021), 132–140.

[24]Suresh, A., Jacobs, J., Clevenger, C., Lai, V., Tan, C., Martin, J.H. and Sumner, T. 2021. Using AI to Promote Equitable Classroom Discussions: The TalkMoves Application. *Artificial Intelligence in Education*. I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, eds. Springer International Publishing. 344–348.

[25]Suresh, A., Jacobs, J., Perkoff, M., Martin, J.H. and Sumner, T. 2022. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (Seattle, Washington, 2022), 71–81.

[26]Tapp, J., Wehby, J. and Ellis, D. 1995. A multiple option observation system for experimental studies: MOOSES. *Behavior Research Methods, Instruments, & Computers*. 27, 1 (Mar. 1995), 25–31. DOI:https://doi.org/10.3758/BF03203616.

[27]Wallace, T., Anderson, A.R., Bartholomay, T. and Hupp, S. 2002. An Ecobehavioral Examination of High School Classrooms That Include Students with Disabilities. *Exceptional Children*. 68, 3 (Apr. 2002), 345–359. DOI:https://doi.org/10.1177/001440290206800304.

[28]Wei, R.C., Darling-Hammond, L. and Adamson, F. 2010. *Professional Development in the United States: Trends and Challenges. Phase II of a Three-Phase Study. Executive Summary*. National Staff Development Council.

[29]Yoder, P.J. and Symons, F.J. 2010. *Observational measurement of behavior*. Springer Pub. Co.

Comparative Study of Double Scoring Design for Measuring Mathematical Quality of Instruction

Jonathan K. Foster[1], James Drimalla[2], Nursultan Japashov[1]
[1]University at Albany
[2]Gordon College
jkfoster@albany.edu, james.drimalla@gordon.edu, njapashov@albany.edu

Abstract

The integration of automated scoring and addressing whether it might meet the extensive need for double scoring in classroom observation systems is the focus of this study. We outline an accessible approach for determining the interchangeability of automated systems within comparative scoring design studies.

1 Introduction

Classroom observation instruments may be deployed in different *classroom observation systems*, i.e., the collection of elements that work together to produce instructional quality ratings such as the observation instrument, raters, and scoring design (Hill et al., 2012). Classroom observation systems operating within education research or large-scale operational use have different goals and constraints than those operating for practical judgements on instructional quality (Liu et al., 2019). For instance, some classroom observation systems embedded in educational research may need calibration and monitor ratings, double scoring of observations, and complete multiple observations of teachers whereas classroom observation systems embedded in a large school district may not match all of these elements. Recent research highlights the need for extensive double scoring to determine whether raters are scoring accurately and consistently (White and Ronfeldt, 2024).

One potential approach to address the extensive need for double scoring is to pair human raters with an automated scoring system (Rotou and Rupp, 2020; Rupp, 2018). In recent years, a growing number of machine learning techniques have been used to identify features of instructional quality in classrooms from videos and audio recordings, or classroom transcripts. In one such study, researchers explored the zero-shot performance of ChatGPT (`gpt-3.5-turbo`) in scoring transcript segments from 4th- and 5th-grade mathematics instruction by applying the Mathematical Quality of Instruction (MQI) tool, a classroom observation instrument (for more information about MQI, see Hill et al., 2008). Results indicated the Spearman correlation between human and machine ratings for dimensions of MQI were low (Wang and Demszky, 2023). In another study, researchers applied a multimodal model and ChatGPT (`gpt-3.5-turbo-1106` and `gpt-4-1106-preview`) to video, audio, and transcripts to score encouragement and warmth in classrooms, a key component of the Global Teaching Insights (GTI) study's observation protocol (Hou et al., 2024). They found pairing the multimodal model with ChatGPT-4 yielded a moderate Pearson correlation ($r = 0.513$). Studies such as these illustrate the opportunities for automated scoring systems in classroom observation.

Current research investigating these automated scoring systems for classroom observation have primarily compared the performance of the automated system to that of human ratings. In terms of automated scoring systems, this focus is one of several components in an argument-validity framework (Rotou and Rupp, 2020; Williamson et al., 2012). These systems depend on human scoring for development. Yet, some scholars critique the lack of theoretical attention to measurement and reporting of inter-rater reliability for classroom observations and question whether classroom

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 163–171
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

observation systems that rely only on human raters can even consistently and accurately measure instructional quality (Liu et al., 2019; White and Ronfeldt, 2024; Wilhelm et al., 2018). Rather than shy away from these complexities with classroom observation systems or call into question the conclusions of some of the recent research on automated scoring systems for classroom observation, we propose an approach to guide others in this area in reporting their results.

The purpose of this paper is to examine an approach for illustrating the implications for double scoring in classroom observation systems when one of the raters is an automated scoring system and the other is a human, especially in the context of smaller datasets for initial system development. We make use of a dataset from a longitudinal study investigating the mathematics instructional quality of early-career elementary teachers in the United States. The automated scoring system includes a random forest classifier using the outputs of a deep neural network capable of detecting instructional activities in videos to score the mathematics instructional quality. Within this context, we present an approach to reporting the accuracy and consistency of double scoring within a *classroom observation system* when one set of scores was automated and the degree of degradation observed. This study seeks to answer the following key research questions:

1. What is the agreement between human and machine scoring? Is there a relative bias between the mean differences of human and machine scores?

2. How reliable is the machine scoring in relation to the human scoring?

3. Is the double scoring method by human and machine interchangeable to that of "gold standard" double scoring by human raters?

2 Background

2.1 Activity Detection with Deep Learning Neural Networks and Random Forests

Deep learning has become the state-of-the-art choice for various challenges including recognizing human activities in video content (Beddiar et al., 2020). A deep neural network is a hierarchical learning structure that can learn complex and abstract features of a given set of data. It is feasible to train neural networks to classify activities in videos of instruction such as the activity structure (i.e., whole group instruction, small group instruction, individual work, and transitions; Ahuja et al., 2019; Foster et al., 2024a), student and teacher behaviors (Foster et al., 2024a; Patidar et al., 2024; Sharma et al., 2021; Sun et al., 2021), and their location (Foster et al., 2024a; Patidar et al., 2024).

In this study, a deep neural network was used to detect instructional activities within video content of elementary mathematics instruction. From the output of the neural network, a random forest classifier was then used to predict the mathematics instructional quality. Random forests are a supervised machine learning algorithm that use many tree-like structures (i.e., decision trees) to make predictions or classifications (James et al., 2021). In the case of classification, a random forest selects the majority vote from decision trees.

2.2 Classroom Observation Measures for Ambitious Mathematics Instruction

There is no single conceptualization of quality mathematics instructional, although there is a fair amount of overlap in what should be regarded as high-quality instruction in mathematics (Praetorius and Charalambous, 2018; Schlesinger and Jentsch, 2016). We conceptualize high-quality mathematics instruction as teaching practices aligned with ambitious mathematics teaching (Franke et al., 2007; Lampert et al., 2013; Newmann and Associates, 1996; Thompson et al., 2013). The Mathematics Scan (M-Scan) is a classroom observation protocol for mathematics teaching aligned with ambitious mathematics instruction (Berry et al., 2013; Walkowiak et al., 2018). It is operationalized at the lesson level and has been empirically validated (Walkowiak et al., 2014). M-Scan has nine dimensions organized under four domains. For each dimension, there are indicators with descriptions for low (1-2), medium (3-5), and high (6-7) ratings.

2.3 Interrater Agreement and Reliability in Classroom Observations

A concern within classroom observation systems is whether raters can accurately and consistently apply an observational instrument. There are several approaches to reporting interrater agreement and reliability and some literature lists these terms interchangeably (Tinsley and Weiss, 2000; White, 2018; White and Ronfeldt, 2024; Wilhelm et al., 2018). However, interrater agreement and interrater reliability are different measures. Interrater agreement indicates the extent to which different raters assign the exact same rating to each observation. Interrater reliability indicates the consistency of raters when scoring a collection. It is possible to have high interrater agreement and low interrater reliability and vice versa (Cicchetti and Feinstein, 1990). Therefore, it is important to consider the implications of both measures and their magnitude (White, 2018; Wilhelm et al., 2018).

Recent research has also brought attention to monitoring rater quality in classroom observation research (White and Ronfeldt, 2024). Typically, only a small subset of videos (i.e., reliability set) is ever double scored to monitor rater quality. However, current recommendations advise more than 20 observations to estimate rater error with 95% confidence (White and Ronfeldt, 2024). One suggestion to help guide researchers' decision making about rater errors is to use simulations. These simulations assume a level of rater error (e.g., ± 1 or 2 points) and hypothesized distribution of "true" or master scores and then examine the implications of the number of observations needed for suitable rater error rates.

2.4 Methods Comparison

This paper makes use of a technique for comparing two methods for measurement that arose out of clinical medical research (Altman and Bland, 1983, Bland and Altman, 1999, 2003, 2010). The technique is not widely used in education research, but there have been recent calls for its use (Wilhelm et al., 2018). The technique can compare one established method to another indirect or less costly alternative method. It assumes that neither method can provide a true measure and so seeks to determine how much the two methods agree and whether they are interchangeable in practice. Both methods are applied to the same observations. Then, the difference between the measures for each observation is calculated to compute the mean difference ($\bar{d}$). If the mean difference is non-zero, then this indicates there is a relative bias.

A range, in which most differences between measurements by the two methods will lie, is called the limit of agreement (LOA). This LOA can be determined using parametric and non-parametric approaches (Bland and Altman, 1999). In this study, we take a non-parametric approach as the distribution of between-method differences (i.e., difference in human ratings) is not well-known. First, we order the differences observed from least to greatest. Then, we remove 5% of the observed

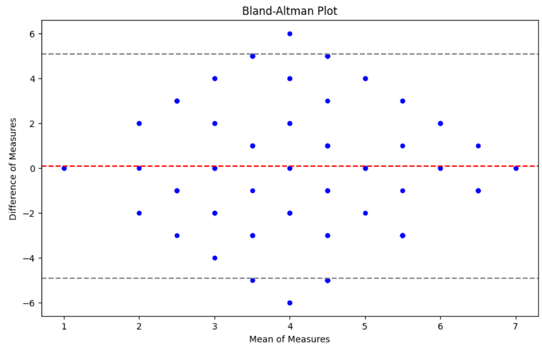


Figure 1: A Bland-Altman plot of two simulated measurements.

differences beginning with the most extreme from either end of the distribution. After removing those extreme differences, we report the LOA by finding the difference of the remaining two endpoints of the observed differences.

There is a related graphical representation, referred to as a Bland-Altman plot. It plots the average of the two methods against the difference of the two methods for each observation. Figure 1 shows a Bland-Altman plot of two simulated measurements of 100 lesson observations. The first and second measure observations range between 1 and 7. In Figure 1, we see that $\bar{d} = 0.08$ and $LOA = 10$. From this Bland-Altman plot, we interpret there is little to no relative bias when $\bar{d}$ is close to 0 and we could say that for 95% of observations, a measurement by the first approach would differ no more than ± 5 units from the second approach. If $LOA \leq 10$ is negligible in practice, then we may conclude the two methods for measuring are interchangeable.

3 Current Study

3.1 Video Data

Videos of elementary instruction used in this study were collected as part of a previous research study known as the Developing Ambitious Instruction (DAI, Youngs et al., 2022). The DAI focused on 83 beginning elementary teachers who graduated from teacher preparation programs at five universities in the United States either in 2015-16 or 2016-17. After graduation, these individuals started teaching young children (ages 5 to 11) full-time in grades K-5 in general education settings. The DAI team observed each teacher as they taught mathematics and English language arts (ELA) at least six times each during their first two years of full-time teaching (i.e., three times each for mathematics and ELA per school year). Each video-recorded lesson was about 45 minutes in length and the current study used a total of 360 hours of video from over 400 lessons.

3.2 Scoring of Lesson Videos with M-Scan

As part of the DAI study, videos of mathematics lessons were assigned scores with M-Scan by at least one human rater. The following steps were taken to train M-Scan raters and ensure high levels of reliability. First, each rater watched three videos of elementary mathematics lessons, assigned scores for each M-Scan domain, and reviewed the master ratings and justifications. Raters then met with the master rater and watched video clips that exemplified different scores on each of the M-Scan dimensions and practiced rating two additional lessons. To determine if raters met certification requirements, raters independently coded a series of lessons without conferring with the master rater; then they met with the master rater to confer on scores. The master rater computed agreement scores (at least 80% exact or adjacent matches were required), identified items that were sources of systematic error, and looked at convergence of ratings. If a rater did not meet the 80% threshold, they were required to rate an additional two lessons. On a regular basis, the master rater conducted a meeting in which raters viewed, coded, and discussed one or two lessons from the reliability set. These meetings were used to monitor raters' ongoing performance.

3.3 Annotations of Videos

For the purpose of training a neural network, the team developed a list of 24 instructional activities for annotating the video dataset. For example, the annotation label of *Using or holding an instructional tool* was developed in reference to the M-Scan dimension Students' Use of Math Tools. Prior to annotating the video dataset, annotators went through training on how to apply the classroom-based activity labels. At the end of the training, annotators' performance was periodically monitored (Foster et al., 2024c).

3.4 Neural Network Model for Instructional Activity Detection

From our prior investigation, we found The Background Suppression network (BaS-Net, Lee et al., 2020) was advantageous for detecting activities in classroom videos (Foster et al., 2024b). The 268 hours of video recordings were used to train and test a modified BaS-Net to detect the 24 instructional activity labels, which we call BaS-Net+. In our experimental setup, training and testing sets were split 80 and 20 percent respectively. In comparison to previous reported results (Foster et al., 2024a), we restructured the testing set so that it did not feature any of the teachers from the training set. Once the neural network was trained and tested on 268 hours from DAI dataset, we then used it to detect the 24 instructional activities in an evaluation set featuring 92 additional math lessons.

3.5 Random Forest Classifiers for M-Scan Scoring

Random forest classifiers were used to predict scores for each M-Scan dimension. We developed the random forest classifiers with the package `randomForest` in R (Breiman, 2001). All 24 instructional activity labels generated by human annotations were used as initial predictors for each of the nine M-Scan dimensions scores. Each random forest included 41 decision trees with the mtry hyper-parameter set between 3 and 5 features at each step of branching.

After building the random forest classifiers, we applied them to the aggregated data in the evaluation set that was generated by BaS-Net+. We then compared the predicted score by the random forest classifiers to human scores.

3.6 Measuring Interrater Agreement and Reliability

We report agreements as ratings that agree exactly or differ by no more than 1 point (Lawlis and Lu, 1972), which we denote as p_0 and p_1. These levels of agreement are often used in practice with human raters for M-Scan (Walkowiak et al., 2018). For interrater agreement, we also report a descriptive index of agreement, developed by Tinsley and Weiss (1975), called the T-index:

$$T = \frac{N_a - Np_c}{N - Np_c} \quad (1)$$

where N_a is the number of agreements, N is the number of ratings, and p_c is the probability of chance agreement of an individual. If T is positive, then the observed agreement is greater than agreement that would occur by chance. If T is negative, then the observed agreement is less than chance agreement. When T is zero, then the observed agreement is equal to the expected chance agreement. There is also a nonparametric chi-square test of significance for the T-index (Lawlis and Lu, 1972; Tinsley and Weiss, 2000). We use T_0 and T_1 to indicate the index that corresponds to exact agreement and within one point agreement, respectively.

To report any relative bias between machine and human ratings, we report the mean difference $(\bar{d})$. A positive mean difference indicates on average the machine scored higher than the human raters while a negative value indicates on average the human scored relatively higher for a given dimension of M-Scan. We use a threshold of $\bar{d} > 0.20$ to conclude a possible relative bias between machine and human ratings.

When reporting interrater reliability for ordinal-scaled ratings, we use Finn's coefficient (r_f) and Gwet's AC_2 for their respective advantages. Both indices range between 0 and 1 with higher values indicating higher levels of consistency between raters. A nonparametric chi-square test for significance is available for r_f and AC_2. Finn's coefficient is recommended for use when the within-raters variance is highly constrained and to use a $p < 0.01$ for applied research (Tinsley and Weiss, 2000). It does not require independent subjects, which in our use case is important as some of the lessons were taught by the same teacher. Gwet's AC_2 is a generalization of Gwet's AC_1 (Gwet, 2008) and it does not assume all raters will be paired randomly for each observation.

3.7 Comparing Methods of Double Scoring

To determine two methods are interchangeable, the Bland-Altman method requires specificity beforehand as to how small the LOA should be to conclude that either method is sufficient in practice. This decision is a practical one, not a statistical decision (Bland and Altman, 1999). Practitioners should provide a strong rationale for this decision. We decided to use what has been observed in prior research with human raters as the "gold standard," although some could argue this may not be sufficient evidence (White and Ronfeldt, 2024). With expert human raters scoring with M-Scan, it was found that they agreed exactly 66.7% and within one point 97.6% of the time (Walkowiak et al., 2018). Thus, we decided to use 65% exact agreement and 95% agreement within one point. We may conclude the two methods are interchangeable if they met or exceeded each of these levels of agreement if the $LOA \leq 1$. However, before we may make such a conclusion, we must check the assumption that there is no relation between the difference between the ratings and average ratings. We use Spearman's rank correlation coefficient (ρ) to examine for any monotonic relations. We use the criteria $|\rho| > 0.30$ to conclude the possibility of any monotonic relationship.

4 Results

4.1 Research Question 1: Agreement

The exact agreement between human and machine scoring ranged between 10.9% to 58.7%. The corresponding T_0-index values are listed in Table 2. Most T_0-index values indicated little to no agreement between the scoring except for Mathematical Accuracy, which indicated moderate agreement ($0.41 \leq T_0 < 0.60$). Allowing for one point difference, we found the extended percent agreements of human and machine ratings between 57.6% and 89.1% agreement. The corresponding T_1-index values range between 0.31 and 0.82. These are moderate to substantial ($T_1 \geq 0.61$) levels of agreement between human and machine ratings. Almost all agreements between human and machine ratings for each dimension of M-Scan were found to be statistically significant; thus, it is highly unlikely these levels of agreement were the result of chance.

M-Scan Dimension	Interrater Agreement			
	T_0	p_0	T_1	p_1
Structure of the Lesson	0.18***	29.3%	0.70***	81.5%
Use of Representations	0.09	21.7%	0.50***	69.6%
Students' Use of Math Tools	-0.04	10.9%	0.49***	68.5%
Cognitive Demand	0.11*	23.9%	0.40***	63.0%
Math Discourse Community	0.22***	33.7%	0.50***	69.6%
Explanation and Justification	0.14**	26.1%	0.52***	70.7%
Problem Solving	0.11*	23.9%	0.31***	57.6%
Connections and Applications	0.18***	29.3%	0.57***	73.9%
Mathematical Accuracy	0.52***	58.7%	0.82***	89.1%

$*p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$

Table 2: Interrater agreements.

Next, we report if there were any relative bias between the human and machine ratings. Examining the mean differences between the paired human and machine scores (see Table 3), we found a relative bias for nearly all the M-Scan dimensions. The human raters were, on average, rating higher scores in comparison to the random forest classifiers for some of the M-Scan dimensions (e.g., Problem Solving). On other dimensions, the random forest classifiers were, on average, scoring higher than the human raters (e.g., Explanation and Justification). No systematic bias was found for the dimensions of Use of Representation when observing the mean differences between the human and machine ratings.

M-Scan Dimension	Mean Differences
	$\bar{d}$
Structure of the Lesson	0.48
Use of Representations	-0.03
Students' Use of Math Tools	-0.23
Cognitive Demand	-0.82
Math Discourse Community	-0.64
Explanation and Justification	0.73
Problem Solving	-0.99
Connections and Applications	-0.33
Mathematical Accuracy	0.30

Table 3: Mean differences between machine and human ratings.

4.2 Research Question 2: Reliability

For all dimensions of M-Scan, we found the interrater reliability between human and machine ratings to be more than substantial (> 0.600) and statistically significant ($p < 0.001$) according to Finn's reliability coefficient (r_F) and Gwet's AC_2. These interrater reliability statistics for each dimension of M-Scan are listed in Table 4.

M-Scan Dimension	Interrater Reliability	
	r_F	AC_2
Structure of the Lesson	0.812***	0.855***
Use of Representations	0.765***	0.851***
Students' Use of Math Tools	0.621***	0.757***
Cognitive Demand	0.800***	0.814***
Math Discourse Community	0.812***	0.860***
Explanation and Justification	0.722***	0.838***
Problem Solving	0.702***	0.706***
Connections and Applications	0.802***	0.877***
Mathematical Accuracy	0.891***	0.946***

$*p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$

Table 4: Interrater reliability

4.3 Research Question 3: Interchangeability

In this section, we report whether the double scoring done by human and machine is interchangeable to the "gold standard" between human raters. For our purpose, we decided if we observed at least 95% agreement within the $LOA \leq 1$ and at least 65% exact agreement between human and machine ratings, then the double scoring done by human and machine would be interchangeable with the method of two human raters. Meeting this condition would indicate that the method of rating a lesson by a human rater and machine rater agrees sufficiently in practice. Table 5 lists all the LOA for each dimension of M-Scan. Before finding the LOAs using the Bland-Altman method, we checked the assumption needed that there is no relation between the difference between the ratings and average ratings using Spearman's ρ-statistic. As shown in Table 5, all dimensions except Problem Solving did not satisfy the needed assumption; thus, these LOA should be interpreted with caution. Nevertheless, we found no evidence to suggest the method of pairing human raters with any of the random forest classifiers is interchangeable with the double scoring with two human raters. This conclusion came from the two necessary criteria: the exact agreement was ≥65% and at least 95% agreement for a $LOA \leq 2$.

M-Scan Dimension	Limit of Agreement	Spearman's Coefficient
	LOA	ρ
Structure of the Lesson	5*	-0.37
Use of Representations	4*	-0.99
Students' Use of Math Tools	6*	-0.80
Cognitive Demand	5*	-0.81
Math Discourse Community	4*	-0.77
Explanation and Justification	5*	-0.82
Problem Solving	6	-0.24
Connections and Applications	4*	-1.0
Mathematical Accuracy	3*	-0.88

Note: (*) indicates these LOA interpretations should be interpreted with caution as there is an association between the mean score and scoring difference, as evidenced by corresponding value of ρ, which does not satisfy one of the criteria for use of the Bland-Altman method.

5 Discussion

Rater error is highly complex and so it is difficult to claim that raters are not significantly altering a measure such as instructional quality. Although interrater agreement and reliability provide some estimates of rater error, recent research suggests a precise measure of rater error requires more scoring occasions than what is typical (White and Ronfeldt, 2024). As a result, this means there is a significant need to double score a sizeable collection to capture a robust measure of rater error.

One potential solution to meeting this size of double scoring is to develop an automated rater. We used our study as a context to illustrate an approach for determining whether double scoring when one of the raters is an automated scoring system is interchangeable with the "gold-standard" of two human raters. We drew on classroom observation systems research and methods comparison studies.

In the context of this study, we found insufficient evidence that the method of double scoring the video by a human and machine was interchangeable with the "gold-standard" method of double scoring by two human raters. Although we found some agreement and reliability between the human and machine ratings, the current level of performance did not provide evidence for the ability to interchange the two methods as set by our outset criteria from what had previously been observed. We acknowledge decisions that we made may not be appropriate for every scoring design.

However, this study goes beyond what is typically reported in findings about the performance of automated classroom observation systems, which typically detail the association between human and machine scores. This study also examined potential impacts on scoring design decisions as they relate to automated scoring such as double scoring when one rater is an automated system. This decision could have several consequences for rater monitoring and associated time and financial costs. There is a need for evaluators of these automated systems to consider methods and frameworks for addressing this issue and others that are beyond calibration between human and machine raters (c.f., Doewes et al., 2023; Johnson et al., 2022; Rotou and Rupp, 2020; Williamson et al., 2012).

Acknowledgments

We would like to thank the members of the Artificial Intelligence for Advancing Instruction team and the Development of Ambitious Instruction team at the University of Virginia for their contributions. This work was supported by the National Science Foundation under Grant No. 2000487 and The Robertson Foundation. Opinions, findings, and conclusions in this presentation are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26.

D. G. Altman and J. M. Bland. 1983. Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3):307–317.

Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. Vision-based human activity recognition: A survey. *Multimedia Tools and Applications*, 79(41):30509–30555.

Robert Q Berry, Sara E Rimm-Kaufman, Erin M Ottmar, Temple A Walkowiak, Eileen G Merritt, and Holly H Pinter. 2013. The Mathematics Scan (M-Scan): A measure of standards-based mathematics teaching practices.

J. M. Bland and D. G. Altman. 2003. Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology*, 22(1):85–93.

J Martin Bland and Douglas G Altman. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160.

J. Martin Bland and Douglas G. Altman. 2010. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 47(8):931–936.

Jonathan Bostic, Kristin Lesseig, Milan Sherman, and Melissa Boston. 2021. Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, 24(1):5–31.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

D. V. Cicchetti and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In pages 103–113.

Jonathan K. Foster, Matthew Korban, Peter Youngs, Ginger S. Watson, and Scott T. Acton. 2024a. Automatic classification of activities in classroom videos. *Computers and Education: Artificial Intelligence*, 6:100207.

Jonathan K. Foster, Matthew Korban, Peter Youngs, Ginger S. Watson, and Scott T. Acton. 2024b. Classification of instructional activities in classroom videos using neural networks. In Xiaoming Zhai and Joseph Krajcik, editors, *Uses of Artificial Intelligence in STEM Education*, pages 439–464. Oxford University Press.

Jonathan K. Foster, Peter Youngs, Rachel van Aswegen, Samarth Singh, Ginger S. Watson, and Scott T. Acton. 2024c. Automated classification of elementary instructional activities: Analyzing the consistency of human annotations. *Journal of Learning Analytics*:1–18.

Megan L. Franke, Elham Kazemi, and Dan Battey. 2007. Mathematics teaching and classroom practice. In Frank K. Lester and National Council of Teachers of Mathematics, editors, *Second handbook of research on mathematics teaching and learning: a project of the National Council of Teachers of Mathematics*, pages 225–256. Information Age Publishing, Charlotte, NC.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Heather C. Hill, Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4):430–511.

Heather C. Hill, Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2):56–64.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education*, pages 60–74, Cham. Springer Nature Switzerland.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An introduction to statistical learning: With applications in R*.Springer texts in statistics. Springer, New York, NY, Second edition.

Matthew S. Johnson, Xiang Liu, and Daniel F. McCaffrey. 2022. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. *Journal of Educational Measurement*, 59(3):338–361.

Magdalene Lampert, Megan Loef Franke, Elham Kazemi, Hala Ghousseini, Angela Chan Turrou, Heather Beasley, Adrian Cunard, and Kathleen Crowe. 2013. Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, 64(3):226–243.

G. Frank Lawlis and Elba Lu. 1972. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78(1):17–20.

Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. 2020. Background Suppression Network for weakly-supervised temporal action localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11320–11327.

Shuangshuang Liu, Courtney A. Bell, Nathan D. Jones, and Daniel F. McCaffrey. 2019. Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1):61–95.

Fred M. Newmann and Associates. 1996. *Authentic achievement: Restructuring schools for intellectual quality*.The Jossey-Bass education series. Jossey-Bass, San Francisco.

Prasoon Patidar, Tricia Ngoon, Neeharika Vogety, Nikhil Behari, Chris Harrison, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. Edulyze: Learning analytics for real-world classrooms at scale. *Journal of Learning Analytics*, 11(2):297–313.

Anna-Katharina Praetorius and Charalambos Y. Charalambous. 2018. Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, 50(3):535–553.

Ourania Rotou and André A. Rupp. 2020. Evaluations of automated scoring systems in practice. *ETS Research Report Series*, 2020(1):1–18.

André A. Rupp. 2018. Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, 31(3):191–214.

Lena Schlesinger and Armin Jentsch. 2016. Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 48(1):29–40.

Vijeta Sharma, Manjari Gupta, Ajai Kumar, and Deepti Mishra. 2021. EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors (Basel, Switzerland)*, 21(17):5699.

Bo Sun, Yong Wu, Kaijie Zhao, Jun He, Lejun Yu, Huanqing Yan, and Ao Luo. 2021. Student Class Behavior Dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing and Applications*, 33(14):8335–8354.

Jessica Thompson, Mark Windschitl, and Melissa Braaten. 2013. Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal*, 50(3):574–615.

Howard E. Tinsley and David J. Weiss. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4):358–376.

Howard E. Tinsley and David J. Weiss. 2000. Interrater reliability and agreement. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, pages 95–124. Elsevier.

Temple A. Walkowiak, Robert Q. Berry, J. Patrick Meyer, Sara E. Rimm-Kaufman, and Erin R. Ottmar. 2014. Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1):109–128.

Temple A. Walkowiak, Robert Q. Berry, Holly H. Pinter, and Erik D. Jacobson. 2018. Utilizing the M-Scan to measure standards-based mathematics teaching practices: affordances and limitations. *ZDM*, 50(3):461–474.

Rose E. Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. arXiv:2306.03090 [cs].

Mark C. White. 2018. Rater performance standards for classroom observation instruments. *Educational Researcher*, 47(8):492–501.

Mark White and Matt Ronfeldt. 2024. Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality. *Educational Assessment*, 29(2):124–146.

Anne Garrison Wilhelm, Amy Gillespie Rouse, and Francesca Jones. 2018. Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research, and Evaluation*, 23(4):1–16.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Peter Youngs, Lauren Molloy Elreda, Dorothea Anagnostopoulos, Julie Cohen, Corey Drake, and Spyros Konstantopoulos. 2022. The development of ambitious instruction: How beginning elementary teachers' preparation experiences are associated with their mathematics and English language arts instructional practices. *Teaching and Teacher Education*, 110:103576.

Toward Automated Evaluation of AI-Generated Item Drafts in Clinical Assessment

Tazin Afrin[1], Le An Ha[2], Victoria Yaneva[1], Keelan Evanini[1], Steven Go[1], Michael Heilig[1], Kristine DeRuchie[1]
[1]National Board of Medical Examiners, Philadelphia, USA
{tafrin, vyaneva, kevanini, sgo, mheilig, kderuchie}@nbme.org
[2]Ho Chi Minh City University of Foreign Languages, Vietnam
anhl@huflit.edu.vn

Abstract

This study examines the classification of AI-generated clinical multiple-choice question drafts as "helpful" or "non-helpful" starting points. Expert judgments were analyzed, and multiple classifiers were evaluated—including feature-based models, fine-tuned transformers, and few-shot prompting with GPT-4. Our findings highlight the challenges and considerations for evaluation methods of AI-generated items in clinical test development.

1 Introduction

The development of high-quality standardized assessments fundamentally depends on the availability of well-crafted test items. As the demand for more efficient and scalable item development grows, many organizations are turning to large language models (LLMs) to meet this need (LaFlair et al., 2023; Song et al., 2025). LLMs offer the promise of aiding the creation of items at scale – increasing diversity and improving security by eliminating item reuse. These benefits make LLMs an attractive solution for organizations seeking to streamline the assessment development process.

However, as LLMs become more widely used for generating content across various domains, evaluating the quality of the generated output has become increasingly critical to the usability of these models. Without a reliable and scalable method for assessing quality, there is a risk of replacing one bottleneck — manual content creation—with another: sorting through a vast amount of content that varies in quality. This issue is especially challenging in fields that require specialized expertise, such as medical educational assessment, and in contexts where there is no universal agreement among experts due to the nuanced and inherently subjective nature of the criteria used to define high-quality output. To fully harness the potential of LLMs in generating exam items, it is essential to address this evaluation bottleneck.

In this study, we present one of the first explorations of automated evaluation of AI-generated items in the clinical domain, using a dataset of 512 clinical multiple-choice questions (MCQs), each rated by two experts. This work presents the following original contributions:

- We collect and analyze expert ratings of AI-generated MCQs in the context of medical education assessment.
- We evaluate a range of automated classification metrics to determine how well they predict expert judgments, identifying which metrics align most closely with human assessments; an error analysis aims to identify areas where these automated metrics fall short.
- While primarily aimed at providing practical insights in assessment development, we also discuss the implications of these findings for the broader challenge of evaluating AI-generated expert text, highlighting the need for nuanced evaluation frameworks as generative AI becomes increasingly integrated into professional workflows.

2 Related Work

Since the advent of LLMs, the medical community has had a keen interest in exploring the medical knowledge of LLMs (He et al., 2025; Singhal et al., 2023; Tang et al., 2023; Yaneva et al., 2024; Zhou et al., 2023) and generating MCQs that can be used in medical education and assessments (Artsi et al., 2024; Al Shuraiqi et al., 2024). The quality of automatically generated MCQs has been evaluated using a range of methods across multiple studies (see Table 1 for a comparison).

Cheung et al. (2023) conducted a multinational prospective study evaluating the quality of MCQs produced by ChatGPT for graduate medical examinations across Hong Kong, Singapore, Ireland, and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 172–182

Study	Evaluation Dimensions	Findings
Cheung et al. (2023)	Appropriateness, Clarity, Relevance, Discrimination, Exam Suitability	No significant difference between AI- and human-generated MCQs; AI scored slightly lower in relevance (p = 0.04)
Klang et al. (2023)	Accuracy, Terminology, Sensitivity	0.5% of questions were false; 15% required revisions due to various inaccuracies
Agarwal et al. (2023)	Validity, Difficulty, Reasoning	ChatGPT produced the least difficult questions; strong inter-rater reliability ($\kappa \geq 0.8$)
Ayub et al. (2023)	Accuracy, Complexity, Clarity	Only 40% of AI-generated questions were suitable for ABD-AE preparation
Bedi et al. (2025)	Distinguishability, Validity, Reviewer Consensus	64% of questions deemed valid; 51.8% distinguishability (random chance); reviewers took 3.2 min/question

Table 1: Summary of studies evaluating AI-generated medical MCQs

the United Kingdom. Five independent international assessors evaluated questions based on five domains: appropriateness, clarity and specificity, relevance, discriminative power of alternatives, and suitability for medical graduate examinations. The study found no significant difference in overall question quality between AI-generated and human-authored questions, except in the relevance domain, where AI-generated questions scored slightly lower (AI: 7.56 ± 0.94 vs. human: 7.88 ± 0.52; p = 0.04).

In Klang et al. (2023), GPT-4 was utilized to generate MCQs for medical examinations. Specialist physicians, blinded to the source of the questions, evaluated them for mistakes and inaccuracies. The study reported that only 0.5% of AI-generated questions required replacement, while 15% required revisions due to issues like outdated terminology and demographic sensitivities.

Agarwal et al. (2023) assessed the applicability of ChatGPT, Bard, and Bing in generating reasoning-based MCQs in medical physiology. Two physiologists rated the AI-generated questions on validity, difficulty, and reasoning ability using a 0-3 scale. ChatGPT produced the least difficult questions, and all AI models showed limitations in generating high-level reasoning questions. Inter-rater reliability was high, with Cohen's kappa (κ) values ≥ 0.8 across all parameters.

In dermatology, Ayub et al. (2023) explored ChatGPT's potential in generating board-style questions. Two board-certified dermatologists conducted a qualitative analysis of 40 AI-generated questions, assessing accuracy, complexity, and clarity. Only 40% of the questions were deemed accurate and appropriate for American Board of Dermatology Applied Exam (ABD-AE) preparation, highlighting the need for expert oversight.

QUEST-AI (Bedi et al., 2025) is an AI system for generating, verifying, and refining USMLE-style items. Three physicians and two medical students participated in a twofold assessment: distinguishing between AI- and human-generated items and evaluating the validity of AI-generated content. Participants could only distinguish between the two at a rate of 51.8%, suggesting indistinguishability of AI-generated items. Furthermore, 64% of AI-generated items were unanimously deemed correct by reviewers, while 36% were flagged for issues like multiple correct answers or incorrect AI-selected answers. The average review time per item was 3.21 minutes, indicating efficiency advantages over traditional question drafting.

Among these studies, only Bedi et al.'s (2025) included automatic evaluation, in the form of an ensemble of language models to automatically flag flawed questions. None of the studies directly evaluated the AI-generated items in the context of operational assessments.

The literature so far provides important insights across a range of use cases, highlighting both the promise and current challenges of AI-assisted item development. However, the diversity in study designs, evaluation rubrics, expert backgrounds, and question types makes direct comparison across studies difficult. Most studies rely on expert judgment, while automated evaluation remains underexplored, with only preliminary use by Bedi et al. (2025). Our study is among the first to investigate automated methods for evaluating AI-generated medical questions, aiming to complement expert review with scalable and consistent quality checks.

3 Data

The dataset used in this study comprises 512 clinical MCQs generated by GPT-4-0314, aiming to cover 26 topics across various clinical domains. These include, but are not limited to, the respiratory system, renal and urinary systems, obstetrics and gynecology, behavioral disorders, and gastroenterology. The items were evaluated by ten subject matter experts (SMEs), who were physicians with extensive experience in writing clinical MCQs for high-stakes standardized assessments. The evaluation was organized such that each item was annotated by two SMEs and each SME saw $\approx$100 items. Additionally, the annotation was organized so that 5 pairs of SMEs that each shared the same domain of expertise annotated the same set of items that were grouped by topic. This paired assignment of SMEs to topics was necessary due to the highly technical nature of the MCQ content and a need to ensure, to the extent possible, that the SMEs had the right background to evaluate the items.

The SMEs were first shown an item stem (the clinical scenario that presents the problem to be solved) along with the key (the correct answer). They were then given up to 12 distractors (incorrect answer options) and asked to select those that, collectively, could form a partial or complete option set for the item. Following this selection, the SMEs evaluated several aspects of the item drafts, including their usefulness as starting points for developing items for a high-stakes clinical assessment. Each draft was rated as either a *Helpful* starting point (requiring relatively minor changes) or a *Non-Helpful* starting point (requiring substantive revisions). Optionally, SMEs also provided rationales for their selections.

When providing their ratings, SMEs were instructed to label drafts as *Helpful* starting points if only minimal revisions were needed. This included small edits to the stem—such as adding, modifying, or removing up to three minor history or exam details for accuracy, realism, or appropriateness—or minor changes to the answer options, like adding a distractor to complete a 4–5 option set with appropriate difficulty. Drafts requiring more substantive revisions to the stem or answer options were to be labeled *Non-Helpful* starting points. These guidelines were intended as reference points, with SMEs encouraged to use their judgment in assessing the overall effort required to finalize a draft. The instructions were presented both in writing and verbally during a training session, where SMEs also rated three sample items together. The discussion with the SMEs revealed the limitations of rigid criteria based on a specific number of item edits, as SMEs noted that a single change can sometimes require significant effort, while multiple superficial edits may be quick and easy to implement.

Data	Helpful	Non-Helpful
Set 1	280	232
Set 2	280	68

Table 2: Distribution of helpful and non-helpful items with (Set 1) and without (Set 2) SME disagreements.

This study focuses on developing an automated evaluation of the quality of AI-generated drafts by using the draft item as input and predicting the labels assigned by the SMEs. These labels are defined as follows: if both SMEs agreed that a draft was a helpful starting point, it is labeled *Helpful*. If at least one of the SMEs rated the draft as not helpful, it is labeled *Non-Helpful*, because the system should preferably reject any item draft that could be labeled as Non-Helpful by human annotators in order to streamline the review process. The data distribution following this labeling method is shown in Table 2 as Set 1. In a follow-up analysis, we refine the label distribution by removing the cases where the SMEs disagreed and consider only item drafts that were labeled by both annotators as either Helpful or Non-Helpful (Set 2 in Table 2). As shown in the following sections, this reduces labeling noise caused by rater disagreement but introduces class imbalance, making the classification task more challenging.

4 Analysis of Human Annotations

Overall, 70.8% of the individual annotations provided by the SMEs characterized the item as Helpful. However, the distribution of Helpful vs. Non-Helpful annotations varied substantially across raters with 47.1% Helpful ratings for the most rigorous annotator and 88.8% Helpful ratings for the most lenient annotator. These results suggest that the SMEs had different subjective interpretations of the definitions of Helpful and Non-Helpful provided in the annotation guidelines. The inter-annotator agreement statistics provide additional evidence for the challenging nature of the annotation task. The overall inter-annotator exact agreement was 67.1% and Cohen's κ was 0.223. Across the five pairs of raters that annotated the same sets of items, exact agreement ranged from 52.0% to 73.3% and Cohen's κ ranged from 0.078 to 0.376.

5 Automated Classification of Helpfulness

We conducted three types of automated classification experiments to predict the helpfulness of generated items based on the expert judgments. These included: (1) a feature-based approach utilizing interpretable features, (2) fine-tuned transformer models, and (3) an LLM judge utilizing few-shot learning with a focus on prompt engineering to explore creative prompting strategies.

5.1 Feature-Based Classification

In our experiment with hand-crafted interpretable features, we conducted a 5-fold cross-validation experiment utilizing a Random Forest (RF) classifier implemented via the `scikit-learn` Python library (Pedregosa et al., 2011). The models were trained on two types of manually engineered features: word count-based features and readability metrics. The word count features captured surface-level textual patterns such as the total word count in the item stem, the number of words in the key, the average word count between distractors, and the maximum word count between distractors. Readability was assessed using the Flesch-Kincaid Grade Level and Flesch Reading Ease scores (Kincaid et al., 1975), which estimate the linguistic complexity of the item content. These features serve as an interpretable baseline intended to quantify the extent to which surface characteristics such as item length are predictive of the two classes.

5.2 Transformer Models

We performed a 5-fold cross-validation experiment and fine-tuned three models: BERT-base-uncased, DeBERTa-v3-base, and DeBERTa-v3-large from HuggingFace (Wolf et al., 2020). The following parameters are used to fine-tune all models: batch size of 16, learning rate of $9e^{-6}$, 50 warmup steps, and a weight decay of 0.01. The input to the models consists of the stem, answer key, and distractor list, each separated with a [SEP] token.

5.3 LLM as Judge with Few-shot Learning

We used GPT-4 (OpenAI et al., 2024) as a judge to determine the helpfulness of the generated item drafts. We employed few-shot prompting and tested the following four distinct prompt designs (refer to Appendix A for the complete prompts):

Simple Prompt: In this approach, we did not provide detailed instructions to the LLM. We instructed the LLM to take the a role of a highly knowledgeable medical educator, provided it with two labeled examples (one Helpful item requiring few edits and one Non-Helpful item requiring major edits), and then asked it to classify a third item.

Criteria-Based Prompt: In this prompting strategy, the LLM was prompted to act as a highly knowledgeable medical educator and was given a set of review criteria, including clarity, relevance, validity, formatting, cognitive level, and statistical usability. Similar to the simple prompt, two labeled examples were followed by a third item to be classified. In this case, the model was explicitly instructed not to provide an explanation.

Criteria-Based Prompt with Rationale: This prompting strategy followed the structure of the criteria-based prompt. In addition, the SME rationales for the example items in the prompt were included, and the model was instructed to provide a clear rationale for its decision.

Similarity-Based Prompt with Rationale: Building on the third prompt, this version improved the example selection process by choosing examples most similar to the target item. Similarity was computed using cosine distance between sentence-level vector embeddings of the items. The sentence vectors were extracted from the sentence transformer embedding model (Zhang et al., 2025).

6 Results

We evaluated all models using two metrics: weighted F1-score and accuracy, with results presented in Table 3. For Set 1, which includes cases that SMEs disagreed upon, the majority baseline yielded an F1-score of 0.387 and an accuracy of 0.547. For Set 2, without SME disagreements, the corresponding scores were 0.718 and 0.805.

Comparative analysis indicates that, while the feature-based Random Forest classifier outperformed the baseline in terms of F1 score, it consistently underperformed on the accuracy metric across both sets. Notably, for all feature ablation combinations, the classifier's accuracy remained below the majority baseline. Among the feature sets, word count features achieved the best performance, suggesting that item length provides a predictive signal when modeling helpfulness. To further investigate this relationship, we computed the Pearson correlation between the helpfulness label and various hand-crafted features. Interestingly, word count features did not exhibit a statistically significant correlation with helpfulness. However, the number of words in the item stem and the readability measured via the Flesch Reading Ease score showed a negative correlation (shown in Table 4).

The fine-tuned transformer models outperformed

		Set 1		Set 2	
		F1-score	**Accuracy**	**F1-score**	**Accuracy**
Baseline	Majority class	0.387	0.547	0.718	0.805
RF	Word count features	**0.505**	0.506↓	**0.741**	0.779↓
	Readability features	0.479	0.482↓	0.709↓	0.747↓
	All features	0.486	0.492↓	0.739	0.796↓
Transformers	BERT	0.558	0.561	0.769	0.802↓
	DeBERTa-base	**0.589**	**0.604**	0.718	0.805
	DeBERTa-large	0.564	0.568	**0.771**	**0.810**
GPT-4	Simple Prompt	0.465	0.575	**0.768**	**0.815**
	Criteria-Based Prompt	0.482	0.573	0.755	0.792↓
	Criteria-Based Prompt w/ Rationale	**0.537**	**0.586**	0.742	0.754↓
	Similarity-Based Prompt w/ Rationale	0.534	0.574	0.732	0.732↓

Table 3: Comparison of model performance using weighted F1-score and accuracy. Models that did not improve over the baseline are marked with the ↓ symbol. The best performing models within each type are marked in bold.

Features	Correlation	P-val Range
Word Count		
Stem	-0.062	[0.093, 0.500]
Answer	0.015	[0.523, 0.798]
Avg. Distractor	0.022	[0.401, 0.974]
Max. Distractor	0.016	[0.481, 0.944]
Readability		
Grade Level	0.053	[0.051, 0.813]
Reading Ease	-0.062	[0.019*, 0.910]

Table 4: Average magnitudes and p-value ranges for correlations between the helpfulness label and hand-crafted features over 5-fold cross validation. *$p < 0.05$

the majority baseline, with the exception of BERT-base-uncased on the accuracy metric. DeBERTa-v3-base consistently outperformed both BERT-base-uncased and DeBERTa-v3-large on Set 1, which includes item drafts with discrepant SME ratings. In contrast, DeBERTa-v3-large achieved the best performance on Set 2, where item drafts with discrepant ratings were removed.

When using GPT-4 to assess helpfulness, we evaluated its performance both with and without rationale explanations. On Set 1, which includes item drafts with discrepant SME ratings, the *Criteria-Based Prompt with Rationale* outperformed all prompting strategies. While the *Similarity-Based Prompt with Rationale* yielded competitive results, it did not surpass the performance of the *Criteria-Based Prompt with Rationale*. In contrast, for Set 2, with no discrepant SME ratings, the *Simple Prompt* achieved the highest performance. The other three prompts did not exceed the baseline accuracy on Set 2, suggesting that in the absence of discrepancy, GPT-4 performs best with simple prompts.

7 Error Analysis

The experiments presented in Section 6 show that modeling draft helpfulness is a challenging task for various classifiers. Our findings identify DeBERTa-base and the *Criteria-Based Prompt with Rationale* as the most effective approaches, which is why we focus on these two models to further understand their error patterns.

The confusion matrices in Appendix B show that both the DeBERTa-base and GPT-4 models are effective at recalling Helpful items—correctly identifying many of them as Helpful. However, both models also exhibit a tendency to incorrectly classify Non-Helpful items as Helpful. We suspect our labeling choice of marking items with annotator disagreement as Non-Helpful may have contributed to the misidentification. To better understand this effect, we analyzed how the models handled the items marked by disagreement. Of the 164 such items, DeBERTa-base predicted 103 (63%) as Helpful and 61 (37%) as Non-Helpful. The GPT-4 model using the criteria-based prompt with rationale predicted 126 (77%) as Helpful and only 38 (23%) as Non-Helpful. These findings suggest that, although items with discrepant ratings were labeled as Non-helpful, the models exhibit a consistent bias toward predicting them as Helpful.

To better understand why a Non-Helpful item might have been labeled as Helpful, we examined the comments provided by both human annotators and the rationales generated by GPT-4 as part of the *Criteria-Based Prompt with Rationale*. On average, GPT-4's rationales were 4.6 times longer than those written by humans. A qualitative analysis reveals that human annotators tend to focus more on issues related to the stem of the question, whereas GPT-4 emphasizes aspects of the distractors and the correct answer. Additionally, GPT-4 often describes items as clear, concise, and easy to understand. Human rationales also contain highly technical medical knowledge about flaws in the

Human Rationale	GPT-4 Rationale
Draft item seems to hint at diagnosis (medication known to cause acute kidney injury, AKI), and yet answer relates to hematuria and proteinuria, which may or may not be seen in AKI. Also, shortness of breath is not further elaborated on in history or physical exam. If this is fluid overload in setting of AKI, would likely have crackles or lung findings. Draft item lacks focus and is not necessarily aligned with indicated answer.	The question stem is clear, concise, and relevant to the intended learning objectives. The options are unambiguous and easy to understand. However, the correct answer should be revised to match the format of the other options. Instead of "Proteinuria and/or hematuria," the correct answer should be written as "Protein and hematuria." This change will make the correct answer consistent with the other options and maintain the clarity of the question. Additionally, the question matches the cognitive level required for the audience and is free of bias and stereotypes.

Table 5: Human annotator rationale vs. GPT-4 rationale for an example where the item draft was labeled as Non-Helpful by the SMEs but predicted as Helpful by GPT-4 using the Criteria-Based Prompt with Rationale.

item draft that the GPT-4 may not pick up.

Given that the items were generated by GPT-4 and are also being evaluated by GPT-4, this may partly explain its tendency to find the items easier to understand. Table 5 presents an example where the item was labeled as Non-Helpful by human annotators but predicted as Helpful by GPT-4, along with their respective rationales.

8 Discussion and Conclusion

This study investigated human and automated evaluations of AI-generated item drafts, intended to serve as starting points for item development. The results indicate that this is a challenging task for both experts and machines. Despite our efforts to mitigate variability—through detailed guidelines, topic-to-rater matching, and a group calibration exercise—inter-rater agreement remained modest.

A likely explanation lies in the inherently subjective nature of the task. A given draft may evoke different ideas and interpretations depending on the item writer's experience, domain-specific preferences, or approaches to the item development process. Writers may also vary in their thresholds for what they perceive as a "substantive" effort required to revise a draft. While future research should further refine rating criteria and protocols, the subjective nature of evaluating helpfulness is unlikely to be eliminated entirely.

Turning to the automated evaluation, classification models exhibited modest success. Even when analysis was limited to instances where both raters agreed—a subset of examples with arguably clearer ground truth—the performance of the classifiers remained moderate. One contributing factor was class imbalance within the dataset (while this skew affects supervised models, the GPT-4-based few-shot prompting approach used a balanced set of examples, mitigating this issue during inference). Notably, rationale-augmented prompts improved GPT-4's performance in Set 1, suggesting that structured reasoning can help guide the model's decisions in more complex cases. However, in Set 2, with no disputed labels, the simple prompt outperformed more elaborate versions—highlighting that, in low-ambiguity scenarios, additional reasoning may introduce unnecessary "cognitive" load and reduce accuracy.

Several limitations warrant consideration. First, the relatively small sample size constrains the generalizability of our findings. Model performance may differ across item formats or subject areas not represented in our dataset. There could also be potential misalignments between item content and rater expertise. For example, an item involving pediatric trauma may have been assigned to a general pediatrician, whereas the underlying clinical focus would have been more suitable for an emergency physician. In addition, the choice of particular specialists in this study were limited by the item writer's availability and addressing these limitations in future work is important for obtaining a more robust evaluation.

The practical aim of this study was to explore whether automated evaluation methods could help streamline the human review process. The extent to which this goal was achieved remains open to interpretation. On the one hand, several of the classifiers outperformed baseline models and demonstrated reasonable recall for identifying helpful items (in other words, there is lower risk of discarding helpful drafts). However, the relatively low precision in identifying unhelpful drafts—when combined with subjective preferences— limits the utility of such methods in practice. Further precision refinement is needed before automated triaging can be considered a dependable aid in clinical test development.

References

Mayank Agarwal, Priyanka Sharma, and Ayan Goswami. 2023. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. *Cureus*, 15(6):e40977.

Somaiya Al Shuraiqi, Abdulrahman Aal Abdulsalam, Ken Masters, Hamza Zidoum, and Adhari AlZaabi. 2024. Automatic generation of medical case-based multiple-choice questions (mcqs): A review of methodologies, applications, evaluation, and future directions. *Big Data and Cognitive Computing*, 8(10).

Yaara Artsi, Vera Sorin, Eli Konen, Benjamin S Glicksberg, Girish Nadkarni, and Eyal Klang. 2024. Large language models for generating medical examinations: systematic review. *BMC Medical Education*, 24(1):354.

Ibraheim Ayub, Dathan Hamann, Carsten R. Hamann, and Matthew J. Davis. 2023. Exploring the potential and limitations of chat generative pre-trained transformer (chatgpt) in generating board-style dermatology questions: A qualitative analysis. *Cureus*, 15(8):e43717.

Suhana Bedi, Scott L. Fleming, Chia-Chun Chiang, Keith Morse, Ankit Kumar, Bhavik Patel, Jindal A. Jindal, Christopher Davenport, Christina Yamaguchi, and Nigam H. Shah. 2025. QUEST-AI: A System for Question Generation, Verification, and Refinement using AI for USMLE-Style Exams. In *Proceedings of the Pacific Symposium on Biocomputing 2025*, pages 54–69.

Billy H. H. Cheung, Gary K. K. Lau, Gordon T. C. Wong, Elaine Y. P. Lee, Dhananjay Kulkarni, Choon S. Seow, Ruby Wong, and Michael Co. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE*, 18(8):e0290691.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated readability index. *Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel: Inst Sim Trng*.

E. Klang et al. 2023. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*, 23(1):1–9.

Geoff LaFlair, Kevin Yancey, Burr Settles, and Alina A von Davier. 2023. Computational psychometrics for digital-first assessments: A blend of ml and psychometrics for item generation and scoring. In *Advancing natural language processing in educational assessment*, pages 107–123. Routledge.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yishen Song, Junlei Du, and Qinhua Zheng. 2025. Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, pages 1–20.

Liyan Tang, Ziyue Sun, Betina R. Idnay, Gongbo Zhang, Yufan Zhang, Chen Wang, Yanshan Zhang, and Hong Yu. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6:158.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Victoria Yaneva, Peter Baldwin, Daniel P. Jurich, Kimberly Swygert, and Brian E. Clauser. 2024. Examining chatgpt performance on usmle sample items and implications for assessment. *Academic Medicine*, 99(2):192–197.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *Preprint*, arXiv:2412.19048.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A LLM Prompts

Simple Prompt

System:
Here are 2 examples of medical MCQ questions where the first example is *Non-helpful* and the second example is *Helpful*. Given a third example, your job is to answer if it is Helpful or Non-helpful.

User:
Example 1: [example 1 question]
Answer: [example 1 answer]
Options: [example 1 options]
Label: Non-helpful

Example 2: [example 2 question]
Answer: [example 2 answer]
Options: [example 2 options]
Label: Helpful

Example 3: [test example question]
Answer: [test example answer]
Options: [test example options]

Is the third example *Helpful* or *Non-helpful* ?"

Criteria-Based Prompt

System:
You are a highly knowledgeable medical educator and expert in medical exam question design. Your task is to review a set of Multiple Choice Questions (MCQs) intended for a medical education platform.

Criteria:

- Clarity and Conciseness: Is the question stem clear and concise, avoiding unnecessary complexity? Are the options unambiguous and easy to understand?

- Relevance and Focus: Does the question align with the intended learning objectives or topic? Is it free of irrelevant or extraneous details that might confuse the respondent?

Criteria-Based Prompt (Cont..)

- Answer Key Validity: Is the correct answer clearly supported by the question and defensible? Are distractors (incorrect options) plausible but clearly incorrect?

- Formatting and Grammar: Is the question grammatically correct, free of typos, and formatted appropriately?

- Cognitive Level: Does the question match the cognitive level (e.g., recall, application, analysis) required for the audience or context?

- Bias and Sensitivity: Is the question free of bias, stereotypes, or language that might disadvantage certain groups?

- Statistical Usability (Optional): Does the question have characteristics likely to yield good discrimination and difficulty levels if data is available?

User:
Here are two examples of well-structured MCQs where the first example is *Non-helpful* and the second example is *Helpful*:

Example 1:
- Question: [example 1 question]
- Options: [example 1 options]
- Correct Answer: [example 1 answer]
- Label: Non-helpful

Example 2:
- Question: [example 2 question]
- Options: [example 2 options]
- Correct Answer: [example 2 answer]
- Label: Helpful

Now, classify the following question:

- Question: [test example question]
- Options: [test example options]
- Correct Answer: [test example answer]

Criteria-Based Prompt (Cont..)

Instruction:
ONLY return one of the following labels:
- *Non-helpful*
- *Helpful*

Do **NOT** provide any additional explanation.

Criteria-Based Prompt with Rationale

System:
You are a highly knowledgeable medical educator and expert in medical exam question design. Your task is to review a set of Multiple Choice Questions (MCQs) intended for a medical education platform.

Criteria:

- Clarity and Conciseness: Is the question stem clear and concise, avoiding unnecessary complexity? Are the options unambiguous and easy to understand?

- Relevance and Focus: Does the question align with the intended learning objectives or topic? Is it free of irrelevant or extraneous details that might confuse the respondent?

- Answer Key Validity: Is the correct answer clearly supported by the question and defensible? Are distractors (incorrect options) plausible but clearly incorrect?

- Formatting and Grammar: Is the question grammatically correct, free of typos, and formatted appropriately?

- Cognitive Level: Does the question match the cognitive level (e.g., recall, application, analysis) required for the audience or context?

- Bias and Sensitivity: Is the question free of bias, stereotypes, or language that might disadvantage certain groups?

Criteria-Based Prompt with Rationale (Cont..)

- Statistical Usability (Optional): Does the question have characteristics likely to yield good discrimination and difficulty levels if data is available?

User:
Here are two examples of well-structured MCQs where the first example is *Non-helpful* and the second example is *Helpful*:

Example 1:
- Question: [example 1 question]
- Options: [example 1 options]
- Correct Answer: [example 1 answer]
- Label: Non-helpful
- Rationale: [example 1 rationale]

Example 2:
- Question: [example 2 question]
- Options: [example 2 options]
- Correct Answer: [example 2 answer]
- Label: Helpful
- Rationale: [example 2 rationale]

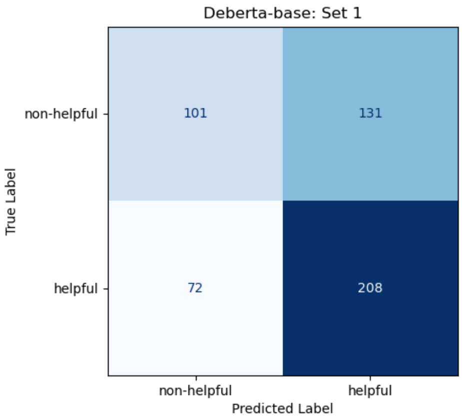
Now, classify the following question:

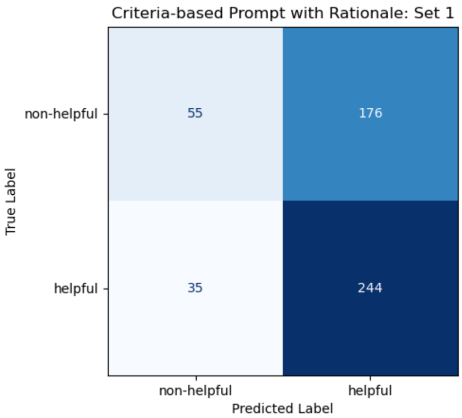
- Question: [test example question]
- Options: [test example options]
- Correct Answer: [test example answer]

Instruction:
ONLY return one of the following labels:
- *Non-helpful*
- *Helpful*

Provide a clear Rationale for your assessment, highlighting any issues related to the system criteria.

B Error Analysis Confusion Matrices





Numeric Information in Elementary School Texts Generated by LLMs vs Human Experts

Anastasia Smirnova*
San Francisco State University
smirnov@sfsu.edu

Erin S. Lee
University of California Berkeley
leoz0113@berkeley.edu

Shiying Li
San Francisco State University
sli63@sfsu.edu

Abstract

LLMs can address long-standing problems in education, such as the lack of instructional materials, by generating grade-appropriate content. We evaluate GPT-4o's ability to generate informational texts for elementary school children. We specifically focus on the model's ability to represent numeric information in text, such as fractions, ratios, and percentages, and assess it with respect to the human baseline. The analysis shows that both humans and GPT-4o reduce numeric information as texts get simplified but do so to a different degree and in a different manner: GPT-4o retains more percentages, while humans use more fractions and ratios. We suggest that these strategies provide different learning opportunities for students.

1 Introduction

Large Language Models (LLMs) have great potential to improve the quality of education (Abdelghani et al., 2024; Han et al., 2024; Yan et al., 2024). They can be used to address long-standing issues in schools, such as shortage of teachers (Edwards et al., 2024) or lack of good instructional materials (Oakes and Saunders, 2004), by generating grade and age-appropriate educational content for students (Scaria et al., 2024; Tan et al., 2025). Diliberti et al. (2024) report that among teachers who employ AI in the classroom, 48% use it to adapt content to the appropriate grade level.

In this paper, we focus on LLMs' ability to adapt informational texts for elementary school children. Informational texts contain quantitative information and expose students to mathematical concepts outside of the math curriculum. The introduction of informational texts in schools in the US was motivated by the demands of the technologically advanced society and the need to develop quantitative literacy (numeracy) in general population (Agnello and Agnello, 2019; Bookman et al., 2008; Steen, 1997, 1999).

Informational texts contain different types of numeric information, as the following passage about paleontological research demonstrates.

Reumer and two colleagues looked in the collections of the Natural History Museum and the Naturalis Biodiversity Center in Leiden, both in the Netherlands, and found 16 samples of mammoth vertebrae from the base of the neck. Seven of the samples were missing the part that would have clued the researchers in on whether a cervical rib had been attached. Of the remaining nine, six were normal and three once had a cervical rib. That worked out to an incidence of 33.3 percent.

Of particular interest here are the last two sentences that allow students to understand how proportions and percentages work.

LLMs' ability to adapt informational text for a specific grade level depends on their mathematical proficiency, their ability to understand quantitative information in text, and to represent it in the form that is appropriate for elementary school children. Mathematical proficiency can be considered an emergent ability in LLMs. McCoy et al. (2024) argue that as a consequence of their design – LLMs were trained to predict next word in text – their performance on tasks that require quantitative skills is sensitive to input probabilities. Thus, GPT-4 performs well on a standard, high-frequency task, such as Celsius-to-Fahrenheit conversion: *multiply by 9/5 and add 32*, but is likely to underperform on a task that has similar complexity but lower input probability: *multiply by 7/5 and add 31*.

Previous work on LLMs' mathematical proficiency returned mixed results. Patel et al. (2023) assessed the ability of GPT-3 model to simplify math word problems for elementary school children. They showed that GPT-3-generated texts are simpler, but noted problems with accuracy. In one instance, GPT-3 inaccurately simplified *she*

*Corresponding author

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 183–191
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

gives each student an eighth of a foot of ribbon as *she gives each student 1 inch of ribbon*. More advanced models perform better – GPT-4 shows 35% improvement in accuracy on math problems compared to GPT-3 (Mishra et al., 2024) – but not at the domain expert level. Mishra et al. (2024) demonstrated that GPT-4o tends to overrely on decimal approximation when working with fractions. Moreover, while GPT-4o showed 90% accuracy on fraction addition tasks, its performance dropped to 61% when instructed to recompute the task with one of the original fractions changed. These results suggest that LLMs' numeric competence is different from human competence (Lee et al., 2024; Lucy et al., 2024).

In what follows, we evaluate LLMs' ability to adapt numeric information in texts for elementary school children. While LLMs' numeric competence is usually assessed on benchmark math tests, we focus on LLMs' ability to convey numeric information in the context of text simplification. Text simplification involves the reduction of structural and lexical complexity of a text, while maintaining its meaning (Shardlow, 2014; Siddharthan, 2014). It is a promising technique for generating age- and grade-appropriate materials with LLMs (Patel et al., 2023).

Previous work on LLMs' ability to generate grade-appropriate content by means of text simplification mostly focuses on the overall readability metrics (Murgia et al., 2023; Patel et al., 2023) or lexical features (Valentini et al., 2023). In these studies, the complexity of a text is operationalized in terms of average word and sentence lengths (shallow textual features), as well as lexical and syntactic features. Simplified texts have shorter words and sentences, more concrete, age-appropriate vocabulary, and simple clauses. These measures, however, do not evaluate how numeric information is conveyed. Similarly to linguistic information, numeric information can be conveyed at different levels of complexity. Proportions, for example, can be represented as fractions, ratios, and percentages (Power and Williams, 2011), and the choice of representation has implications for comprehension and understanding (Bautista et al., 2011). Since numeric information is not included in the standard text complexity measures, little is known about how well LLMs can simplify numeric information in texts.

Our study addresses this gap by evaluating LLMs' ability to convey numeric information in texts and assessing their performance vis-à-vis human experts. We chose to evaluate a particular LLM, GPT-4o by OpenAI, one of the most advanced models at the time of writing. This choice is motivated by GPT-4o's superior performance on math tasks in comparison to other models (Lucy et al., 2024; Mishra et al., 2024) and its widespread use in education.[1]

Our evaluation of GPT-4o and human experts focuses on two questions:

1. How does the amount of numeric information change as texts get simplified?

2. How is difficult mathematical information (proportions) represented in simplified texts?

We report two main findings. First, as texts get simplified, the amount of numeric information is reduced in both human-simplified and GPT-4o-simplified texts. Crucially, GPT-4o reduces numeric information to a greater extent than humans do. Second, we find that humans and GPT-4o use different strategies when simplifying complex information (proportions) with GPT-4o preserving more complex numeric representation (percentages).

2 Study 1: Amount of Numeric Information

Numeric information imposes additional cognitive demands on readers, and thus increases text complexity (Agnello, 2021). We expect that the amount of numeric information will decrease as texts are adapted for lower grade levels.

2.1 Data

Our texts come from Newsela, a provider of educational materials for K-12 curriculum. The Newsela corpus (Xu et al., 2015) is a parallel corpus of informational texts, consisting of the original texts and the corresponding human-simplified texts. There are 5 levels of text complexity within the corpus, from the most complex (level 0) to the least complex texts (level 4). As Figure 1 shows, the distribution of texts by grade and complexity levels is not uniform, with the two largest subsets being texts for grade 12, level 0 (the most complex texts) and texts for grade 4, level 4 (the most simplified texts). These are the two grade levels that we choose to analyze in our study.

[1] https://openai.com/index/introducing-chatgpt-edu/

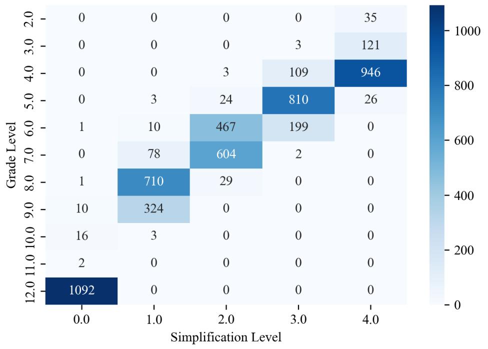


Figure 1: Distribution of texts by grade and simplification level in Newsela corpus.

2.2 Constructing Three Corpora

Our analysis is based on a subset of the Newsela corpus (original and human-simplified texts) and the corresponding GPT-4o-generated texts. For the original corpus, we randomly selected 100 Newsela texts for grade 12, level 0. To construct the human-simplified corpus, we matched these original texts with the corresponding 100 simplified texts for grade 4, level 4. We generated the corpus of GPT-4o-simplified texts by submitting 100 original texts as input to GPT-4o model with instructions to simplify.

We used OpenAI API (model = GPT-4o, temperature = 1) with zero-shot prompting strategy. The prompt was designed to match the style (informational texts), grade level (grade 4), and the average length of texts in human-simplified corpus. Thus, since the average number of words in human-simplified texts for grade 4, level 4 was 680, this length requirement was specified as part of the prompt. The prompt was formulated as follows: "In approximately 680 words, simplify the text below for a fourth-grade reading level written in the newspaper genre."

We did not specifically instruct the model to simplify numeric information. This choice is motivated by the consideration to keep instructions for LLMs and humans as similar as possible (Lampinen, 2024). Since human experts in Newsela use readability scores (Lexile) to guide their simplification process (Agnello, 2021), and since these scores do not take numeric complexity into account, numeric complexity was not referenced in the prompt to GPT-4o. Thus, neither human experts nor the model are specifically instructed to simplify numeric information.

We manually examined GPT-4o-simplified texts for hallucinations and found none. Moreover, our analyses showed that GPT-4o can adequately reduce textual complexity for a specific grade level (Smirnova et al., 2025).

2.3 Extracting Numeric Expressions

Our definition of numeric expressions is based on Agnello (2021). Numeric expressions include counts and measures, arithmetic operations, fractions, percentages, ranges, and others.[2] To extract numeric expressions from texts, we designed a regular expression-based pipeline. Texts were lightly preprocessed to normalize special characters and whitespace, while pattern matching was performed case-insensitively to maximize coverage. Regular expressions were then applied to the preprocessed texts, and sentences containing numeric matches were extracted using rule-based splitting. The results were recorded in three output files for original, human-simplified, and GPT-4o-simplified texts.

2.4 Results

We computed the average number of numeric expressions in the three corpora. The original texts (M=23.55, SD=14.42) contain more numeric expressions compared to both human-simplified (M=12.36, SD=7.15) and GPT-4o-simplified texts (M=9.45, SD=7.10) (see Figure 2). The difference in the distribution of numeric expressions in original and human-simplified texts was statistically significant on a paired t-test ($t(99)=9.042$, $p<0.00001$), and so was the difference between original and GPT-4o-simplified texts ($t(99)=11.698$, $p < 0.00001$). Importantly, the difference between GPT-4o-simplified and human-simplified texts was also statistically significant ($t(99)=4.097$, $p=0.0001$).

Fewer numeric expressions in GPT-4o-simplified texts might suggest that these texts are simpler compared to the corresponding human-simplified texts. However, the number of numeric expressions alone is not sufficient to address this question. In Study 2 we analyze how different numeric expression types are distributed in simplified texts.

[2]See https://github.com/sub-mit/numeracy for the full list of numeric expression types, the corresponding regular expressions and the examples of sentences they match.

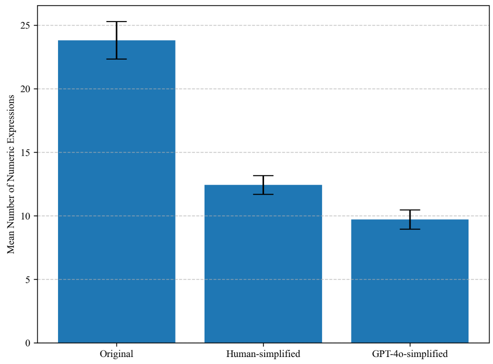


Figure 2: Numeric expression means for original, human-simplified, and GPT-4o-simplified texts. Error bars are +/-1 standard error.

3 Study 2: Complexity of Numeric Expressions

In this study we analyze how proportions are expressed in texts. Proportions can be represented as percentages (*25 percent*), fractions (*one-fourth*), and ratios (*one in four*) (Power and Williams, 2011). The analyses of educational materials and studies with human participants showed that percentages are the most complex numeric expression type (Bautista et al., 2011; Power and Williams, 2011; Wu, 2011). As texts become simplified, we expect that the percentages in the original text will be replaced with other expressions that can convey proportions.

3.1 Percentages in Original Texts

From the list of sentences with numeric expressions in original texts (Study 1), we extracted all sentences that mention "percent". There was a total of 120 unique sentences (types) from 47 texts. Several sentences contained multiple numeric expressions with "percent" in them. Each such expression was treated as an independent token. As a result, we ended up with a total of 147 token sentences referencing percentages.

3.2 Passage Alignment

In order to analyze how percentages from the original texts were represented in the corresponding human-simplified and GPT-4o-simplified texts, we implemented fine-tuned Neural Conditional Random Field (CRF) passage alignment algorithm by Jiang et al. (2020). The alignment is based on the similarity score between passages.

The Neural CRF model is specifically designed for sentence alignment tasks in text simplification (Jiang et al., 2020). It employs a linear-chain CRF integrated with neural network components to align complex and simplified text pairs. The alignment sequence is determined by combining semantic similarity scores derived from fine-tuned BERT embeddings with transition features that reflect sentence order within parallel documents. We select this model for aligning our Newsela corpora because it was trained on Newsela-style datasets, and the authors have provided a version specifically fine-tuned for Newsela sentence alignment.

Our alignment process consisted of two steps described below, data preprocessing and the identification of the most similar passage.

3.2.1 Data Preprocessing

Newsela passages are smaller than paragraphs; they contain one or more sentences. To each Newsela passage we assigned a unique passage_id, formatted as follows: `{corpus_type}__{slug}__{language}__{n_passage}`.

- corpus_type:
  - source_files_grade12: the corpus of original source texts;
  - human_simplified_grade4: human-simplified texts for grade 4;
  - llm_simplified_grade4: the corpus of GPT-4o-simplified texts;
- slug: the slug of the article from which the passage is extracted, e.g. afghan-taxidriver;
- language: the language of the passage;
- n_passage: if the passage_a is the nth passage in the article, then n_passage of passage_a is n.

3.2.2 Getting The Most Similar Passage

The program compares each passage in the original text corpus with every candidate passage in the simplified corpora using the passage-to-passage_id map which provides information about the article. We followed the notebook provided by Jiang et al. (2020) to build our own passage alignment pipeline and used their pre-trained fine-tuned Newsela sentence alignment model for our tasks. For each article within a simplified corpus, we identified and

Text type	Text passage	Numeric Type
Original text	Halloween is crucial to the company, accounting for **25 percent** of Party City's $1.6 billion in annual retail sales.	Percentages
Human-simplified	Party City is a big retailer. It has many stores. Halloween is very important to the company. **One-quarter** of its sales each year come from Halloween.	Fraction
GPT-4o-simplified	Party City is another big store. It sells Halloween items in regular stores and special Halloween City stores. Halloween makes up **25%** of Party City's sales.	Percentages

Table 1: Types of numeric information in aligned passages from original, human-simplified, and GPT-4o-simplified texts.

selected the passage that had the highest similarity score as the aligned passage. The results were recorded as an aligned triplet of original – GPT-4o-simplified – human-simplified passages with similarity scores for subsequent analysis. See Appendix A for an example of aligned triplet with similarity scores.

This implementation resulted in 147 aligned triplets across three conditions (total of 441 passages, i.e. 147 x 3).

3.3 Qualitative Analysis and Coding

We manually examined all aligned triplets. This allowed us to assess how accurately GPT-4o conveys numeric information as well as alignment accuracy. We did not find any numerical inaccuracies in GPT-4o-generated texts, but we did find mismatches in alignment. In cases of content mismatch within a triplet, we consulted full texts side-by-side and searched them for a better candidate to replace the mismatched passage in either human-simplified or GPT-4o-simplified texts. We ended up replacing 53 passages (31 passage replacements in human-simplified and 22 passage replacements in GPT-4o-simplified texts).

We manually coded how numeric expressions referencing percentages were represented in simplified texts. Based on the previous literature (Agnello, 2021; Bautista et al., 2011), we developed a coding system consisting of 5 categories: Percentages, Ratio, Fraction, Non-numeric word, and Dropped. Dropped means that the information was absent in simplified texts. Non-numeric words, such as quantifiers "few" and "many" convey information non-numerically. Of the remaining categories, fraction is the least difficult numeric expression. Ratio can be viewed as a complex fraction (Wu, 2011) but it is less complex than percentages.

Num Type	Humans	GPT-4o
Percentages	4	30
Ratio	12	9
Fraction	18	1
Non-numeric	30	32
Dropped	83	75
Total	147	147

Table 2: Representation of percentages in human-simplified and GPT-4o-simplified texts.

Percentages are the most sophisticated way to represent proportions (Bautista et al., 2011). Table 1 presents an example of aligned passages and the codes for numeric expressions.

3.4 Results: Complex Numeric Types

Chi-square test shows that there are statistically significant differences between human-simplified and GPT-4o-simplified texts in terms of the types of numeric expressions used to represent percentages ($p < 0.00001$, $\chi^2(16)=75.5$). The agreement in the choice of numeric expressions between human-simplified and GPT-4o-simplified texts is 53%. Table 2 shows the distribution by type. Both GPT-4o and humans drop a substantial number of numeric expressions with percentages. When this information is preserved, GPT-4o tends to retain the same numeric type, percentages, while humans tend to use fractions and ratios.

4 Conclusion

In this study we evaluated GPT-4o's ability to convey numeric information in texts simplified for elementary school children and compared its performance vis-à-vis human experts. Study 1 showed that both humans and GPT-4o reduce the number of numeric expressions as they simplify texts, but GPT-4o does so to a greater extent. Since numeric expressions increase text complexity, these results might suggest that GPT-4o-simplified texts are less complex. Study 2 showed that GPT-4o-simplified texts retain percentages, the most difficult numeric type, to a greater extent than human-simplified texts do.

Is GPT-4o's strategy less effective? While linguistically and numerically difficult texts can present a challenge for the reader, they can also provide a unique learning opportunity.The analysis of GPT-4o-generated passages shows that percentages are presented in a way that is easy for a child to understand. Specifically, these texts make the relationship between numbers and percentages transparent: *Some samples were missing parts, but of the nine they could study, three had a cervical rib. This means around 33% of the mammoths had these extra ribs.* From this perspective, retention of complex numeric expressions in GPT-4o-generated texts can be viewed as a learning opportunity, fostering the development of numeracy in elementary school children. At the same time, simplifications that avoid difficult content might ultimately slow down learners' progress (Crossley et al., 2014).

5 Limitations

There are several limitations that arise from the novelty and complexity of the phenomenon under consideration. First, while our choice of GPT-4o model is motivated by its capabilities and widespread application in educational context, it is not clear whether these results will generalize to other LLMs.

Second, we analyzed representation of numeric information in a context of a general text simplification task, but we did not discuss how numeric complexity is related to linguistic complexity. Just as numeric and linguistic features interact in word problem tasks (Daroczy et al., 2015, 2025), linguistic factors can affect representation of numeric information in educational texts.

Finally, we operationalized numeric complexity as (i) the amount of numeric information and (ii) the type of numeric information in texts. While these are standard measures of numeric complexity (Agnello, 2021; Bautista et al., 2011), there are limitations to the approach that is based solely on the distributional frequency of numeric expressions in text. Text comprehension by the intended end users, elementary school children, can serve as additional evaluation metrics for assessing the quality of informational texts generated by LLMs. A comprehension study can directly compare different numeric simplification strategies, contrasting texts that retain complex numeric types (percentages) with texts that represent the same information as fractions or non-numeric words.

References

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2024. Gpt-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, 34(2):483–518.

Ellen C Agnello. 2021. Simplified but not the same: Tracing numeracy events through manually simplified newsela articles. *Numeracy*, 14(2):1–20.

Ellen C Agnello and Kevin M Agnello. 2019. Crossing the final frontier: Exploring the numeracy demands of texts read in english language arts. *Numeracy*, 12(2):7.

Susana Bautista, Raquel Hervás, Pablo Gervás, Richard Power, and Sandra Williams. 2011. How to make numerical information accessible: Experimental identification of simplification strategies. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*, pages 57–64. Springer.

Jack Bookman, Susan L Ganter, and Rick Morgan. 2008. Developing assessment methodologies for quantitative literacy: A formative study. *The American Mathematical Monthly*, 115(10):911–929.

Scott A Crossley, Hae Sung Yang, and Danielle S McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

Gabriella Daroczy, Christina Artemenko, Magdalena Wolska, Detmar Meurers, and Hans-Christoph Nuerk. 2025. Are text comprehension and calculation processes in word problem solving sequential or interactive? an eye-tracking study in children. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*.

Gabriella Daroczy, Magdalena Wolska, Walt Detmar Meurers, and Hans-Christoph Nuerk. 2015. Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, 6:348.

Melissa Diliberti, Heather L Schwartz, Sy Doan, Anna K Shapiro, Lydia Rainey, and Robin J Lake.

2024. *Using Artificial Intelligence Tools in K-12 Classrooms*. RAND.

Danielle Sanderson Edwards, Matthew A Kraft, Alvin Christian, and Christopher A Candelaria. 2024. Teacher shortages: A framework for understanding and predicting vacancies. *Educational Evaluation and Policy Analysis*.

Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. 2024. Teachers, parents, and students' perspectives on integrating generative ai into elementary literacy education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476.

Unggi Lee, Youngin Kim, Sangyun Lee, Jaehyeon Park, Jin Mun, Eunseo Lee, Hyeoncheol Kim, Cheolil Lim, and Yun Joo Yoo. 2024. Can we use gpt-4 as a mathematics evaluator in education?: Exploring the efficacy and limitation of llm-based automatic assessment system for open-ended mathematics question. *International Journal of Artificial Intelligence in Education*, pages 1–37.

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. MathFish: Evaluating language model math reasoning via grounding in educational curricula. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5644–5673, Miami, Florida, USA. Association for Computational Linguistics.

Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Shubhra Mishra, Gabriel Poesia, Belinda Mo, and Noah D. Goodman. 2024. Mathcamps: Fine-grained synthesis of mathematical problems from human curricula. *Preprint*, arXiv:2407.00900.

Emiliana Murgia, Zahra Abbasiantaeb, Mohammad Aliannejadi, Theo Huibers, Monica Landoni, and Maria Soledad Pera. 2023. Chatgpt in the classroom: A preliminary exploration on the feasibility of adapting chatgpt to support children's information discovery. In *UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 22–27, New York, NY, USA. Association for Computing Machinery.

Jeannie Oakes and Marisa Saunders. 2004. Education's most basic tools: Access to textbooks and instructional materials in california's public schools. *Teachers College Record*, 106(10):1967–1988.

Nirmal Patel, Pooja Nagpal, Tirth Shah, Aditya Sharma, Shrey Malvi, and Derek Lomas. 2023. Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning*, 39(3):804–822.

Richard Power and Sandra Williams. 2011. Generating numerical approximations. *Computational Linguistics*, 38(1):113–134.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Anastasia Smirnova, Kyu beom Chun, Wil Louis Rothman, and Siyona Sarma. 2025. Text simplification for children: Evaluating llms vis-à-vis human experts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Lynn Arthur Steen. 1997. *Why numbers count: Quantitative literacy for tomorrow's America*. New York: College Entrance Examination Board.

Lynn Arthur Steen. 1999. Numeracy: The new literacy for a data-drenched society. *Educational Leadership*, 57:8–13.

Kehui Tan, Jiayang Yao, tianqi pang, Chenyou Fan, and Yu Song. 2025. Elf: Educational llm framework of improving and evaluating ai generated content for classroom teaching. *J. Data and Information Quality*.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore. Association for Computational Linguistics.

Hongxi Wu. 2011. *Understanding Numbers in Elementary School Mathematics*. American Mathematical Society.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

A Appendix: Passage Alignment

The program compares each passage in the original text corpus with every candidate passage in the simplified corpora based on the similarity scores. Table 3 presents an example of an aligned triplet with similarity scores.

Slug	**Original passage**	**Human-simplified passage**	**Human-simplified: Similarity Score**	**GPT-4o-simplified passage**	**GPT-4o-simplified: Similarity Score**
predatoryfish-decline	The removal of top predators has been called "humankind's most pervasive influence on nature," and it is as detrimental in the sea as it is on land. Consumers prefer predatory fish like grouper, tuna, swordfish and sharks to species lower on the food chain such as anchovies and sardines, providing strong incentives for fishermen to catch the bigger fish. Going after the more valuable predators first, fishing them until there aren't enough left to support a fishery and then moving on to species lower in the food chain, a pattern sometimes observed in global fisheries, has been called "fishing down the food web."	The result is something called "fishing down the food web." Fishermen go after the more valuable predators first. They fish them until there aren't enough left. Then they move on to smaller fish that are lower on the food chain. The bigger fish start to disappear	0.9997756	Overfishing big, important fish in the sea is causing trouble. People like to eat big fish like tuna, swordfish, and sharks. These are called predatory fish because they eat smaller fish. Because people want to eat these fish, fishermen catch a lot of them. Once there aren't many big fish left, they move on to catching smaller fish like anchovies and sardines. This is known as "fishing down the food web."	0.9971335

Table 3: Example of passage alignment and similarity scores.

Towards evaluating teacher discourse without task-specific fine-tuning data

Beata Beigman Klebanov, Michael Suhan, Jamie N. Mikeska
ETS Research Institute
bbeigmanklebanov,msuhan,jmikeska@ets.org

Abstract

Teaching quality is one of the determinants of student outcomes. Teaching simulations with feedback are one way to provide teachers with practice opportunities to help improve their skill. We investigated methods to build evaluation models of teacher performance in leading a discussion in a simulated classroom with the goal of providing feedback, particularly for tasks with little performance data.

1 Introduction

Teaching quality is one of the determinants of student outcomes (Blömeke et al., 2016; Fauth et al., 2019). The theory of practice-based teacher education (Ball and Cohen, 1999) argues that teachers need opportunities to practice core teaching skills, such as engaging students in the disciplinary discourse practices and leading classroom discussions, in situations of reduced complexity (Grossman, 2021; Forzani, 2014). For such practice opportunities to be successful and impactful, they need to be flexible, target specific difficulties, and provide learning support, in the form of timely feedback (McDonald et al., 2013; Mikeska et al., 2024).

Simulated classrooms are one environment providing such opportunities. They allow for strategic reduction in task complexity so that aspects of teaching can be isolated and practiced separately. Simulations are used in teacher education in various forms, including peers role-playing students (Davis et al., 2017; Masters, 2020), mixed-reality simulations where trained actors play the students (Bondie et al., 2021; Dieker et al., 2019), as well as emerging work where AI agents role-play students to help train teachers or tutors (Lim et al., 2025; Pan et al., 2025; Markel et al., 2023). Across all these forms, feedback to the teacher on their performance that would point out strengths and areas for growth in a constructive and actionable manner is critical (Cohen et al., 2020; Mikeska et al., 2023a).

Until recently, a bottleneck for creating automated feedback was acquiring a substantial amount of data of learners performing the simulation. Such data, with human scores, enabled the creation of machine learning based evaluation models to power automated feedback. With the advent of few-shot learning with large language models, there is an opportunity to mitigate the bottleneck, since only a handful of examples might suffice for these models to be able to evaluate a new learner's performance.

The goal of this paper is to start exploring this opportunity through two research questions: (**RQ1**) How accurately can a few-shot LLM evaluate a teacher's performance in a simulation? (**RQ2**) How do these results compare to an alternative method – a *generic* model fine-tuned on data from other tasks and used to evaluate performance in a new task? The latter approach has been successful in large-scale essay scoring, where a model trained on essays responding to a variety of essay prompts is used to evaluate essays from new, unseen prompts (Ramineni and Williamson, 2013). However, results might depend on data representation. The relatively content-agnostic essay scoring features may have been responsible for the success; recent reports suggest that transformer-based fine-tuned models do not generalize well across prompts (Shermis, 2024). To our knowledge, this is the first exploration of few-shot vs. generic fine-tuned models for evaluating teacher discourse in the absence of fine-tuning data for a task-specific model.

2 Related work

2.1 Digital teaching simulations with feedback as a learning opportunity for teachers

Using digital teaching simulations within teacher education and professional development programs can improve teachers' instructional skills, beliefs, and knowledge (Francis et al., 2018; Lee et al., 2024). Simulations are typically integrated into

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 192–200
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

these learning environments via cycles where teachers prepare for, engage in, and reflect on their performance, as well as receive formative feedback on how well they have enacted specific aspects of teaching (Mikeska et al., 2023b; Pecore et al., 2023). Recent research has shown that timely, personalized feedback is important to propel teachers' learning from digital teaching simulations (Cohen et al., 2020; Garrett et al., 2020; Mikeska et al., 2023a). Yet, such feedback is hard to scale, as generating it relies on extensive human resources.

2.2 Automated evaluation of teacher discourse

Recent work on automated evaluation of classroom discourse using pre-trained LLMs has explored fine-tuning (Kupor et al., 2023; Nazaretsky et al., 2023; Xu et al., 2024; Ilagan et al., 2024), zero/few-shot learning (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024; Hou et al., 2024; Asano et al., 2025), or both (Chen, 2023; Tran et al., 2024). Xu et al. (2024) noted that results are better for aspects of teaching that require less pedagogical expertise. None of these studies systematically investigated generalization across content domains, topics of discussion, or other aspects of classroom discourse.

3 Feedback in a teaching simulation

The context of this paper is ongoing work on developing new tasks for digital teaching simulations focused on the core teaching competency of leading a math or science argumentation-based discussion in an elementary classroom. After engaging in a simulation, teachers receive an automated feedback report. The report was designed by teacher education researchers and professionals to cover indicators of teaching quality (Mikeska et al., 2024). For each indicator, the report provides a comparison to a typical high-quality discussion and shows utterances where the target behavior did and did not occur; see Figure 1. The high-quality discussions are those that received a high score on a holistic rubric such as shown in Table 1. The comparison to high-quality discussions shows whether the teacher has engaged in the target behavior often enough.

4 Method

4.1 Data

We use data collected in multiple studies where a simulation was used as part of pre-service teacher coursework (Mikeska et al., 2023b, 2022). Before the simulation, the teacher is shown the prior work of two or three groups of simulated students; each group is designed to have a specific knowledge profile. For example, in task S1 students need to identify the mystery powder – one of six known powders – and the properties needed for the identification. The groups differ in what they think the powder is and what properties are needed for identification, as they explain in their prior work. Teachers are given up to a week to prepare to lead a 20-min discussion with a specific learning goal. In S1, the goal is to have the class reach consensus on the mystery powder and the necessary properties to identify it. In another task (M1), the goal is reaching consensus on methods for ordering fractions with different numerators and denominators.

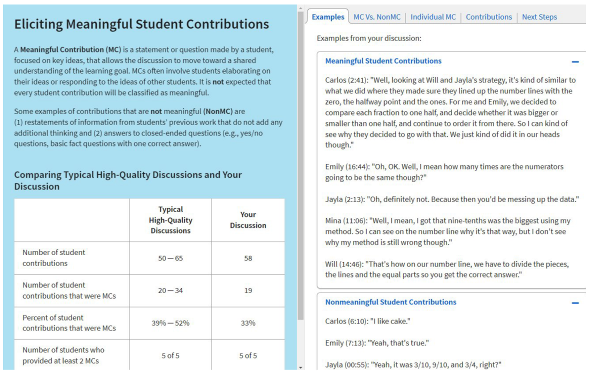


Figure 1: A screenshot of a part of a feedback report for indicator 1b – elicitng meaningful student contributions.

1	[Beginning] The teacher does not make any use of student ideas during the lesson.
2	[Developing] The teacher makes use of some student ideas in a limited way. This can mean: A missed opportunity to move the lesson forward; Only occasiona use of students' ideas when there were multiple opportunities throughout the lesson; An attempt was made to use the students'ideas, but the teacher did not do so in a way that moved the lesson forward.
3	[Well-prepared] The teacher makes use of student ideas to move the lesson forward.

Table 1: Holistic scoring rubric, indicator 1c.

Each discussion transcript was scored using a multi-dimensional rubric (Mikeska et al., 2021). Dimension 1 focused on the teacher's skill in attending to students' ideas equitably. Dimension 1 covered three indicators: how well the teacher (1a) incorporated all the key ideas that appear in the students' prior work into the discussion; (1b) elicited meaningful contributions from all students; and (1c) attended to and made use of each of the relevant student ideas shared during the discussion. We focus on 1b and 1c, for which raters scored teacher performance on the scale of 1–3 or 1–4 (allowing scores like 1.3 or 2.7) and provided justifications by selecting one or more specific teacher (1c) or student (1b) utterances in the transcript where the target behavior clearly did (label 1) or did not (label 0) occur. These justifications form the bulk of the utterance-level annotations used in this study; some additional selections were made by research staff. Table 6 in the Appendix shows a snippet of a discussion, with justifications. We collapsed the top two levels on the 1-4 scale into score 3, as the fourth level was originally added to separate out the strongest performances. Table 1 shows the rubric for indicator 1c. The inter-rater reliability was estimated on double-scored data from task S1: r = 0.52 for 1b and r = 0.53 for 1c, indicating moderate agreement (Dancey and Reidy, 2007).

We use data from six discussion tasks, two in science (S1, S2) and four in math (M1 through M4). Table 2 shows short descriptions. For three tasks – S1, M1, M2 – we have fine-tuning data. For S2, M3, and M4 we have only data to evaluate models; these three tasks will serve as the new tasks to answer RQ2. Table 3 describes the test sets. Data used to develop few-shot models and to fine-tune the BERT model will be described with the models.

4.2 Models

4.2.1 Few-shot models

We implemented the model setup found to be the best for assessing various aspects of classroom discourse in the literature: Tran et al. (2024) investigated task formulations, zero vs few-shot, random or selected examples, and length of context for scoring models implemented using Mistral and Lacuna LLMs. Across multiple constructs and both LLMs, the authors found that an utterance-level classifier with ten few-shot examples (4 positive, 3 negative, and 3 hard negative) and with four prior utterances as context resulted in the best prediction of human holistic scores when aggregated into transcript-level scores. After sampling the test data described in Table 3, we sampled 3 transcripts per task to serve as sources of the ten examples for the tasks S2, M3, and M4 for which we had little data available. For tasks S1, M1, and M2, we sampled one transcript at a time from a larger set until the target 10 examples were identified. It took 8, 6, and 5 transcripts for the three tasks, respectively. Teacher education researchers and assessment developers selected the examples.

We use a state-of-art LLM, **GPT-4o**, and an open source smaller model, DeepSeek-R1:7b (**DS-R1**), distributed through Ollama 0.5.1,[1] both with temperature of 1.0 and default settings for all hyperparameters. Prediction is done on each utterance three times; majority vote decides the final label.

The prompt is a template into which task-specific information is put when the model is used to evaluate data from that task. The template elements are the domain (math or science), task information (e.g., identifying the mystery powder), learning goal (see Section 4.1), and few-shot examples. Below is the template of the GPT-4o prompt for 1c, with few-shot examples appended as user and assistant turns in the messages array sent to the model:

Instructions
Answer yes or no to the following question:
Given the dialogue between a teacher and students in a *{domain}* classroom about *{task_info}*, in the last turn, did the teacher attempt to make use of students' ideas to move the discussion towards the learning goal?
Learning Goal
{learning_goal}
Student names
Jayla, Will, Emily, Mina, Carlos
Output structure
Output must be one of the following words:
yes
no

To take advantage of DeepSeek-R1's "thinking", the examples are included in the system prompt, and the instructions for output structure state that the answer should be on the last line of the output.

4.2.2 Fine-tuned models

We use the utterance-level binary classifiers for indicators 1b and 1c originally developed for the S1 task by Nazaretsky et al. (2023). For indicator 1c, the teacher's utterance to be classified is represented as an embeddings vector and enriched by the embeddings vector of the preceding student utterance as context. For indicator 1b, we are classifying the students' utterances as providing or not providing a meaningful contribution, as evidence for the teacher's success in eliciting such contributions. Therefore, for indicator 1b, the target utterances are students' and the context is the preceding teacher or another student's utterance. The models use Hugging Face DistilBERT[2] base model (un-

[1] https://ollama.com
[2] https://huggingface.co/docs/transformers/model_doc/distilbert

M1	The teacher leads a discussion of three student-generated strategies for ordering the given fractions from least to greatest.
M2	The teacher leads a discussion with the students about an unconventional student-generated method for generating fractions between two given fractions. The discussion is focused on the strengths and weaknesses of the strategy, and its applicability to other situations.
M3	This discussion is grounded in students' work on a story problem in which they have used fraction multiplication. Prior to the discussion, the students individually critiqued one another's work, making the critique aspect of argumentation more clearly available to the teacher.
M4	This discussion focuses on students' work to generate meaningful understandings and representations of division by a fraction.
S1	The discussion focuses on reaching group consensus on the identity of an unknown powder based on its properties and what is known about a set of common powders. In addition to identifying the mystery powder, students discuss which properties are most useful and why.
S2	In this task, the teacher supports the students in discussion whether the amount of matter is conserved during a physical change, in this case, the mixing of ingredients to produce lemonade.

Table 2: Task descriptions.

Task	M1	S1	M2	S2	M3	M4
#transcripts	40	34	40	40	37	35
			Indicator 1c:			
#utts. (K)	1.8	1.5	1.5	1.8	1.8	1.5
#labeled utts. (K)	.39	.29	.46	.63	.49	.89
Average score	2.3	2.4	2.4	2.5	2.4	2.4
Std of scores	.60	.58	.60	.51	.50	.58
%1 in labeled utts.	52	67	71	27	60	69
			Indicator 1b:			
#utts. (K)	2.1	1.9	1.7	2.1	2.1	1.9
#labeled utts. (K)	.60	.43	.62	.77	.74	.65
Average score	2.3	2.4	2.4	2.5	2.4	2.4
Std of scores	.65	.57	.65	.45	.57	.68
%1 in labeled utts.	36	59	72	53	60	45

Table 3: Description of the test data. For each indicator, we show the number of teacher (1c) or student (1b) utterances, overall and labeled. Next are average and std of the holistic transcript scores. The last rows show the percentage of labeled utterances that have the label 1 (where the target behavior occurs).

cased, 66M parameters) (Sanh et al., 2019) with PyTorch 2.2.2 (Paszke et al., 2019). Details of the fine-tuning process and licensing information can be found in the Appendix (see Technical Details). We fine-tuned the classifiers on data from 120 transcripts – 40 from each of M1, S1, and M2 – that were sampled after the test data was partitioned out.

4.3 Model Evaluation

We evaluate the models in two ways. First, we use the utterance level data classified as 1 or 0 using rater justifications to evaluate the accuracy of utterance-level predictions. We think about these as "easy" data, in the sense used in the NLP and machine learning literature to refer to clear-cut, uncontroversial cases – as opposed to "hard" cases where human annotators disagree or where the correct label might be genuinely unclear (Leonardelli et al., 2021; Loukina et al., 2018; Jamison and Gurevych, 2015; Beigman Klebanov and Beigman, 2014). Having just "easy" evaluation data does not allow for a comprehensive evaluation of utterance-level predictions, but being able to classify the easy cases correctly can serve as a first-cut test, as a model that can't handle the easy cases would have low face validity. This evaluation most directly supports the feedback component where example class 1 and class 0 utterances are shown (see Figure 1).

The second evaluation is at the level of the transcript, where we derive a holistic score from utterance-level predictions (number of utterances classified as 1) and compare it to the holistic score provided by human raters. In the easy-vs-hard framework, this evaluation includes both easy and hard instances, since predictions are made on all utterances, some of which are expected to be harder than others. In the NLP literature, there is evidence that for some tasks, training a model on only the easy data results in better performance on not just the easy test cases but on the hard ones, too (Jamison and Gurevych, 2015), presumably because the system isn't getting confused by training examples that could be ambiguous or controversial. The transcript-level evaluation supports the feedback where the overall statistics of the target behavior in the current teacher's discussion are compared to those in high-quality discussions (see Figure 1).

5 Results

5.1 Utterance-level

Table 4 shows the results for the utterance level classification. To answer RQ1: GPT-4o has average accuracy of 0.73 across two indicators $\times$ six tasks, range = 0.62–0.81, std = 0.064. DeepSeek-R1 is much less successful, with average accuracy of only 0.56, range = 0.38–0.69, std = 0.10.

To answer RQ2, we compare the performance of GPT-4o on tasks S2, M3, and M4 to that of the BERT model fine-tuned on data from tasks M1, S1, and M2. The average accuracy of the BERT model on two indicators $\times$ three new tasks is 0.76, range = 0.54–0.91, std = 0.13. The average accuracy of the GPT-4o model on the same data is 0.72, range = 0.63–0.81, std = 0.07. While BERT has a higher average, it is more volatile, with a poor performance of 0.54 on indicator 1c for task S2.

Comparing GPT-4o and BERT on the three tasks on which BERT was fine-tuned – M1, S1, and M2 – we observe that BERT shows stronger performance, as expected. BERT's average accuracy on two indicators $\times$ three tasks is 0.81, range = 0.70–0.89, std = 0.08. GPT-4o's average accuracy on the same data is 0.74, range = 0.62–0.79, std = 0.06. These results indicate that it is worthwhile to fine-tune a model on available data for scoring new performances belonging to the tasks on which the model was fine-tuned. For utterance-level scoring of data from new tasks, one might want to go with GPT-4o, as its performance is comparable to BERT's on average but more stable across tasks.

	M1	S1	M2	S2	M3	M4
			Indicator 1c:			
BERT	**.73**	.70	**.87**	.54	**.91**	**.81**
DS-R1	.53	.52	.69	.38	.47	.46
GPT-4o	.62	**.75**	.79	**.63**	.73	.68
			Indicator 1b:			
BERT	**.84**	**.81**	**.89**	.67	**.81**	**.84**
DS-R1	.50	.65	.69	.57	.69	.58
GPT-4o	.77	.75	.78	**.69**	**.81**	.79

Table 4: Accuracy of classifying teacher utterances as making use of student ideas or not (Indicator c) and student utterances as providing a meaningful contribution or not (Indicator b), on labeled test data. Best performance per indicator per task is boldfaced. Gray background marks BERT performance on tasks on which the BERT model was fine-tuned.

5.2 Transcript-level

Table 5 shows the correlations between the human holistic indicator score and the number of teacher (1c) or student (1b) utterances that were classified as 1 (exhibiting the target behavior). To answer RQ1: GPT-4o averages r = 0.46 across the two indicators $\times$ the six tasks, range = 0.14–0.73, std = 0.18. DeepSeek-R1 averages r = 0.44 for the same data, range = .22–.64, std = 0.15. Thus, the two few-shot models show comparable moderate performance and substantial volatility across tasks.

To answer RQ2, we compare the few-shot models to BERT on the three new tasks – S2, M3, and M4. BERT performs at r = 0.39 on average across two indicators $\times$ three tasks, range = 0.19–0.57, std = 0.14. GPT-4o average performance on the same data is r = 0.32, range = 0.14-0.52, std = 0.13. DeepSeek-R1 averages r = 0.34, range = 0.22-0.51, std = 0.10. At the aggregate level of the full transcripts, the generic fine tuned model tends to show stronger performance than few-shot models, although all the models achieve only low-medium correlations with the human holistic scores and are quite volatile.

Across the three tasks on which the BERT model was fine-tuned (tasks M1, S2, M2), it outperforms the few-shot models: BERT averages r = 0.67, range = 0.55–0.79, std = 0.09. GPT-4o averages r = 0.59, range = 0.45–0.73, std = 0.09. DeepSeek-R1 averages r = 0.54, range = 0.33–0.64, std = 0.11. For the transcript level, the results suggest that the fine-tuned generic model is the model of choice.

	M1	S1	M2	S2	M3	M4
			Indicator 1c:			
BERT	**.74**	.59	**.55**	.32	.19	**.41**
DS-R1	.33	**.63**	.53	.22	**.27**	.32
GPT-4o	.45	.62	**.55**	**.35**	.14	.22
			Indicator 1b:			
BERT	**.79**	**.70**	**.64**	**.50**	.34	**.57**
DS-R1	.64	.59	.54	.32	**.37**	.51
GPT-4o	.73	.62	.59	.36	.31	.52

Table 5: Pearson's correlation between the human holistic indicator score and the number of teacher utterances automatically labeled exhibiting the target behavior. Best performance per indicator per task is boldfaced. Gray background marks BERT performance on tasks on which the BERT model was fine-tuned.

6 Discussion

Our results suggest that a fine-tuned generic model is worth creating if only to score the data from the tasks it was fine-tuned on, as it shows stronger performance than few-shot models in these cases, both at the uttereance and at the transcript level. However, for evaluating data from new tasks for which it is not feasible to fine-tune a model due to lack of data, the situation is less clear-cut. In particular, at the utterance level, the generic fine-tuned model shows more volatile performance across tasks than the few-shot one, failing to capture the "easy" distinctions between utterances in one of the six evaluated cases (accuracy = 0.54 on task S2).

It could be that the S2 data is particularly difficult to classify; however, since GPT-4o shows much stronger performance, it is likelier that the class distribution difference between the fine-tuning data and the S2 data is to blame for BERT's failure of generalization for S2. Indeed, Table 3 shows that S2 data had an exceptionally low proportion of 1s – only 27%. This occasional generalization failure of a generic fine-tuned model illustrates its weakness compared to a few-shot model, namely, its dependence on class distribution in the fine-tuning data, to which the few-shot models are more robust.

Another interesting finding is that the wide gap between GPT-4o and DeepSeek-R1 at the utterance level (average accuracy of 0.73 vs 0.56, respectively) is closed almost entirely at the transcript level (average correlations of 0.46 and 0.44, respectively). Thus, while DeepSeek-R1 has worse face validity, as it isn't able to consistently tell apart clear-cut cases of 0s and 1s, its aggregate performance is similar to GPT-4o's. To gain further insight into this result, we checked the proportion of utterances classified as 1. GPT-4o classified 69% of all teacher utterances as 1 (indicator 1c). The number is 57% for DeepSeek-R1. DeepSeek-R1 predicts many fewer 1s than GPT-4o, including mis-classifying more "easy" 1s as 0s: The ratio of the number of false negative predictions for the labeled utterance-level data for DeepSeek-4o to that of GPT-4o is 3.5:1, while the ratio of false positives is 1:1. Still, the two systems make similar relative judgments of which transcripts have more utterances with the target behavior. Indeed, GPT-4o's and DeepSeek-R1's transcript-level predictions correlate at r = 0.72 on average across tasks for indicator 1c – a much stronger correlation than either of them has with the human holistic scores.

Finally, we observe that transcript-level performance on M1, S1, and M2 is stronger than on S2, M3, and M4, for both GPT-4o and BERT. It is not the case at the utterance level, apart from S2. The models were able to classify the "easy" utterance-level labeled data, but that was not always sufficient to be able to classify all cases – easy and hard – in a reasonable way, that is, in alignment with the tendency expected based on the holistic score. For GPT-4o, limiting the number of transcripts to draw the ten few-shot examples from to 3 may have resulted in examples of lower quality – when we were not limited by dataset size, we went through 5-8 transcripts to find good examples for S1, M1, and M2. Going with fewer than 8-10 transcripts per new task may not be advisable. For BERT, it is apparently not enough to fine-tune on "easy" cases to handle not only the "easy" cases for the new tasks but the hard ones, too. In future work, we will explore automated detection of harder examples. This should help focus the utterance-level models on the easy ones (where the accuracy is high) for picking examples for feedback. Identifying harder cases in the unlabeled utterances from the 120 fine-tuning transcripts, labeling them, and adding to the fine-tuning data might help improve BERT's transcript-level performance on new tasks.

The current study has a number of limitations. First, we experimented with a limited range of models; it is possible that results would change with more effective prompts or different LLMs or more sophisticated data representation for fine-tuning. Second, we considered only two teaching quality indicators; while it is encouraging that the results are aligned between these two, further work is necessary to evaluate robustness of the findings.

7 Conclusion

We investigated two approaches for evaluating teacher performance in leading a discussion in a simulated classroom in the context where no data for fine-tuning on the specific discussion task is available. One approach uses a few-shot LLM. The other approach is a "generic" model fine-tuned on data from other discussion tasks. We found that the few shot model (GPT-4o) may be preferable for analyzing utterance-level data, due to its more stable performance across tasks, while the fine-tuned BERT model performed better in the aggregate, transcript-level evaluation. Our results thus point towards a way to capitalize both on few-shot learning and on previously collected data in order to supply the most effective learning opportunity – the one with timely automated feedback – even when little prior data is available for the current performance task.

References

Yuya Asano, Beata Beigman Klebanov, and Jamie Mikeska. 2025. Exploring task formulation strategies to evaluate the coherence of classroom discussions with GPT-4o. In *Proceedings of the 20th Workshop* *on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 716–736, Vienna, Austria. Association for Computational Linguistics.

Deborah Ball and David Cohen. 1999. Developing practice, developing practitioners: Toward a practice-based theory of professional education. In *Teaching as the learning profession: Handbook of policy and practice*, pages 3–32. San Francisco: Jossey Bass.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396.

Sigrid Blömeke, Rolf Vegar Olsen, and Ute Suhl. 2016. Relation of student achievement to the quality of their teachers and instructional quality. *Teacher quality, instructional quality and student outcomes*, 2:21–50.

Rhonda Bondie, Zid Mancenido, and Chris Dede. 2021. Interaction principles for digital puppeteering to promote teacher learning. *Journal of research on technology in education*, 53(1):107–123.

Gaowei Chen. 2023. Can ChatGPT detect student talk moves in classroom discourse? A preliminary comparison with BERT. In *Proceedings of the 16th International Conference on Educational Data Mining*.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational evaluation and policy analysis*, 42(2):208–231.

Christine Dancey and John Reidy. 2007. *Statistics without maths for psychology*. Pearson education.

Elizabeth A Davis, Matthew Kloser, Andrea Wells, Mark Windschitl, Janet Carlson, and John-Carlos Marino. 2017. Teaching the practice of leading sense-making discussions in science: Science teacher educators using rehearsals. *Journal of Science Teacher Education*, 28(3):275–293.

Lisa A Dieker, Carrie Straub, Michael Hynes, Charles E Hughes, Caitlyn Bukathy, Taylor Bousfield, and Samantha Mrstik. 2019. Using virtual rehearsal in a simulator to impact the performance of science teachers. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 11(4):1–20.

Benjamin Fauth, Jasmin Decristan, Anna-Theresia Decker, Gerhard Büttner, Ilonca Hardy, Eckhard Klieme, and Mareike Kunter. 2019. The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and teacher education*, 86:102882.

Francesca M Forzani. 2014. Understanding "core practices" and "practice-based" teacher education: Learning from the past. *Journal of teacher education*, 65(4):357–368.

Anthony Tuf Francis, Mark Olson, Paul J Weinberg, and Amanda Stearns-Pfeiffer. 2018. Not just for novices: The programmatic impact of practice-based teacher education. *Action in Teacher Education*, 40(2):119–132.

Rachel Garrett, Toni Smith, Melinda Griffin, and Melissa Yisak. 2020. A randomized field study of a teacher professional development program using mixed-reality simulation to develop instructional practice. *Society for Research on Educational Effectiveness*.

Pam Grossman. 2021. *Teaching core practices in teacher education*. Harvard Education Press.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and ChatGPT. In *International Conference on Artificial Intelligence in Education*, pages 60–74. Springer.

Michael Ilagan, Beata Beigman Klebanov, and Jamie Mikeska. 2024. Automated evaluation of teacher encouragement of student-to-student interactions in a simulated classroom discussion. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 182–198, Mexico City, Mexico. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 291–297.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.

Tammy D Lee, Carrie Lee, Mark Newton, Paul Vos, Jennifer Gallagher, Daniel Dickerson, and Camryn Regenthal. 2024. Peer to peer vs. virtual rehearsal simulation rehearsal contexts: Elementary teacher candidates' scientific discourse skills explored. *Journal of Science Teacher Education*, 35(1):63–84.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieun Lim, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, Haewon Jung, Yoonsun Jang, Sanghyeok Lee, and Jewoong Moon. 2025. Development and implementation of a generative artificial intelligence-enhanced simulation to enhance problem-solving skills for pre-service teachers. *Computers & Education*, 232:105306.

Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 1–12.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Heidi Masters. 2020. Using teaching rehearsals to prepare preservice teachers for explanation-driven science instruction. *Journal of Science Teacher Education*, 31(4):414–434.

Morva McDonald, Elham Kazemi, and Sarah Schneider Kavanagh. 2013. Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of teacher education*, 64(5):378–386.

Jamie Mikeska, Heather Howell, Joseph Ciofalo, Adam Devitt, Elizabeth Orlandi, Kenneth King, and G Simonelli. 2021. Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics: The mystery powder task. *ETS Research Memorandum no. RM-21-06)*.

Jamie Mikeska, Jonathan Steinberg, Pamela Lottero-Perdue, Dante Cisterna, Devon Kinsey, and Heather Howell. 2023a. Using simulated classrooms to examine elementary teachers' perceptions about, attention to, and use of formative feedback to improve their ability to facilitate science discussions. *Contemporary Issues in Technology and Teacher Education*, 23(1):48–83.

Jamie N Mikeska, Heather Howell, and Devon Kinsey. 2023b. Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education*, 74(5):422–436.

Jamie N. Mikeska, Beata Beigman Klebanov, Alessia Marigo, Jessica Tierney, Tricia Maxwell, and Tanya Nazaretsky. 2024. Exploring the potential of automated and personalized feedback to support science teacher learning. In *Artificial Intelligence in Education*, pages 251–258, Cham. Springer Nature Switzerland.

Jamie N. Mikeska, Calli Shekell, Adam V. Maltese, Justin Reich, Meredith Thompson, Heather Howell, Pamela S. Lottero-Perdue, and Meredith Park Rogers. 2022. Exploring the potential of an online suite of practice-based activities for supporting preservice elementary teachers in learning how to facilitate argumentation-focused discussions in mathematics and science. In *Proceedings of Society for Information Technology Teacher Education International Conference 2022*.

Tanya Nazaretsky, Jamie N Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with AI: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 122–132.

Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

John L Pecore, Corey Nagle, Tadlee Welty, Minkyoung Kim, and Melissa Demetrikopoulos. 2023. Science teacher candidates' questioning and discussion skill performance in a virtual simulation using experiential deliberate practice. *Journal of Science Teacher Education*, 34(4):415–435.

Chaitanya Ramineni and David M Williamson. 2013. Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1):25–39.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mark D Shermis. 2024. Automated scoring for NAEP short-form constructed responses in reading. In *The Routledge International Handbook of Automated Essay Evaluation*, pages 117–140. Routledge.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2):304–335.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4375–4389.

A Appendix

A.1 Example of the data

T:	Mina and Will, why did you choose weight as an important property?
M:	Because it falls under some of the things that we can see and measure.
T:	Carlos, do you want to explain to them about why you thought that weight wasn't important?
C:	**Sure. Well, actually I don't think weight is really that important, because the weight of the object doesn't really change what the object is. If you were to add more powder, it would change the weight, but that doesn't change what the powder is.**
M:	I guess I see what you mean by that, but I still think that we found the correct thing.
T:	Jayla and Emily, do have any other points to make on the conversation of whether or not weight was important?
E:	Well, we did test the weight in ours because we thought that testing all the properties would be important, but now I'm starting to think about it. I guess weight doesn't really matter, since if we were to add more or take away some of the powder the weight would change, but it wouldn't change what the powder was, like Carlos was saying. So I get that now.
T:	Right. Will, Jayla, do you have any other points that you want to make?
W:	I guess I'm starting to see what Carlos means by that.
T:	Jayla?
J:	Oh yeah, I can see where he was coming from.

Table 6: A snippet of a Mystery Powder (S1) discussion. The blue-boldfaced teacher's utterance was marked by a human rater as an example of the teacher using a student's idea to move the discussion forward (indicator 1c), whereas the black-boldfaced utterance from Carlos was marked as an example of a meaningful student contribution (indicator 1b). T: Teacher. M: Mina. C: Carlos. J: Jayla. W: Will. E: Emily.

A.2 Technical Details of the Models

Fine-tuning. The BERT models were fine-tuned (including all transformer layers, the pooling layer, and the final dense output layer) with Adam optimizer (learning rate = 1e-5, learning warmup = 600) to minimize the binary cross-entropy loss. We used a grid search across 15 epochs with batch sizes 1, 4, and 8 for indicator 1c and across 10 epochs with batch sizes 1, 4 and 8 for indicator 1b. The parameters were tuned using 7-fold cross-validation on the fine-tuning data. For indicator 1c, we used 7 epochs with batch size 4. For indicator 1b, we used 3 epochs with batch size 4.

GPU hours. For indicator 1c, DistilBERT fine-tuning was run locally on a desktop PC with an NVIDIA GeForce RTX 3050 GPU and 16gb physical memory. All fine-tuning, including grid search for all models, took 5 hours and 14 minutes. For indicator 1b, DistilBERT finetuning was run locally on a MacBook Pro with an Apple M2 Pro chip (integrated GPU) and 16gb physical memory, and took 14 hours and 57 minutes.

For both indicators, DeepSeek-R1 test set predictions were run on the same PC, taking on average 97 minutes per transcript for indicator 1c and 66 minutes for indicator 1b. Predictions for fine-tuned models were run on the same MacBook Pro, taking on average 20 seconds per transcript for indicator 1c and 66 seconds per transcript for indicator 1b. GPT-4o predictions were generated using the Batch API via the Microsoft Azure OpenAI Service under our organization's subscription, which provides a 24-hour target turnaround for batch jobs.

Licensing of artifacts. The instance of GPT-4o used is a proprietary AI model accessible via Microsoft's Azure OpenAI Service, subject to Microsoft's licensing terms. Ollama and DeepSeek-R1 are licensed under the MIT License. DistilBERT is licensed under the Apache License, Version 2.0. PyTorch is licensed under the BSD-3-Clause.

Linguistic Proficiency of Humans and LLMs in Japanese: Effects of Task Demands and Content

May Reese and **Anastasia Smirnova**
San Francisco State University
1600 Holloway Avenue, San Francisco, California, USA
mreese@mail.sfsu.edu
smirnov@sfsu.edu

Abstract

We evaluate linguistic proficiency of humans and LLMs on pronoun resolution in Japanese, using the Winograd Schema Challenge dataset. Our main research question is whether task demands and content effects affect performance in these two target groups. First, we found that in the baseline condition, humans outperform LLMs. This finding is consistent with the observation that the language of evaluation is important and that humans perform better than LLMs in lower-resourced languages. Second, we find strong evidence for the effect of task demands in both humans and LLMs. As task demands increase due to syntactic incongruencies in the input, accuracy rates fall for both groups. Third, we found evidence for content effects. In the relevant condition, the content of the scenarios referenced US culture, a favorable condition for LLMs and an adversarial condition for Japanese speakers. We found that LLMs outperformed humans, providing strong evidence for content effects.

1 Introduction

Large Language Models (LLMs) display an impressive set of abilities that require proficiency in human language. They perform well on text summarization (van Schaik and Pugh, 2024), translation (Wang et al., 2023), and writing (Herbold et al., 2023). On many of these tasks, LLMs perform better than humans. For example, Herbold et al. (2023) asked professional evaluators to assess argumentative essays generated by ChatGPT and by humans. The results suggested that the GPT-generated essays consistently achieved higher rankings and were deemed by experts to be of higher quality.

These results are encouraging. However, when LLMs are evaluated on seemingly simpler tasks targeting basic linguistic proficiency, such as the ability to distinguish grammatical sentences from ungrammatical, or meaningful expressions from nonsensical, the results are mixed. Dentella et al. (2024) found that the ability of LLMs to decide whether a sentence is grammatical is much worse than that of humans. While GPT-4 achieved significantly higher accuracy, LLMs performed at the chance level when results were averaged across all tested models. Moreover, LLM responses displayed errors that humans would never make. The authors concluded that LLMs' understanding and performance on tasks involving grammar is not human-like (cf. also Katzir, 2023). In another study, Riccardi et al. (2024) evaluated the ability of LLMs to detect whether a two-word combination is meaningful (*baby clothes*) or nonsensical (*clothes baby*). In humans, this judgement requires knowledge of syntax and semantics. The rightmost word is the syntactic head, and it determines the meaning of the construction: *baby clothes* are a type of clothes. The same rule would make *clothes baby* nonsensical. Riccardi et al. (2024) found that even the most advanced models, such as GPT-4, performed poorly compared to humans. One interesting tendency was for LLMs to err on the side of interpreting nonsensical phrases as meaningful.

The discrepancies between LLMs and humans on basic linguistic tasks have implications for LLM integration in everyday life. There are many applied contexts in which it is highly desirable for LLMs to behave similarly to humans with respect to language understanding. For example, if LLMs' abilities are leveraged in educational contexts to provide feedback on children's writing or on L2 learners' essays, LLMs' assessment of what is grammatical and what is not should parallel the assessment of human experts. Riccardi et al. (2024) identified similar challenges for a workplace context. If the task description or a request does not make sense, be it due to human error or malicious intent, LLMs should behave like a professional human expert would – by asking for clarification or by denying the request, not by interpreting it as

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 201–211

sensible across the board, a tendency that LLMs in their study displayed.

Studies that found performance differences between LLMs and humans on basic linguistic tasks were criticized for using evaluation methods that disadvantage LLMs (Lampinen, 2024; Hu et al., 2024). For example, Lampinen (2024) found that when LLMs are provided with a sufficient number of examples as part of the prompt, they achieve human-like performance when distinguishing grammatical sentences from ungrammatical sentences. Another criticism pertained to the use of metalinguistic prompts, which disadvantage LLMs (Hu et al., 2024). These authors argue that subpar performance should not be interpreted as lack of competence. In fact, studies suggest that LLMs and humans perform similarly on tasks that target basic linguistic proficiency. Hu and Frank (2024) argue that increasing the task demand can lead to lower accuracy in LLMs, just like an increased cognitive load leads to worse performance in humans. Lampinen (2024), focusing on reasoning tasks, also found that the content of the task can either facilitate or hinder performance, and that humans and LLMs show similar content effects.

Our work continues the line of research evaluating LLMs performance vis-à-vis humans on tasks requiring linguistic proficiency. To address the most recent debate about the effect of task demand and content on LLMs and humans, we evaluate their performance as we manipulate these conditions. Unlike previous studies, we focus on Japanese.

2 Evaluating Linguistic Proficiency With the Winograd Schema Challenge

2.1 The Winograd Schema Challenge as a Test of Linguistic Proficiency

In this study we use the Winograd Schema Challenge (WSC). The WSC was originally designed to evaluate machine intelligence as an alternative to the Turing test (Levesque et al., 2012). However, despite its promise and widespread application as a benchmark for commonsense reasoning, it is now generally acknowledged in the literature that the test falls short of assessing machine intelligence (Kocijan et al., 2023). At best, it is a test of linguistic proficiency (Browning and LeCun, 2023), and we use it as such.

The test consists of different scenarios, each of which has a pair of sentences. The classic example in (1) shows that each sentence introduces two entities, the city councilmen (A) and the demonstrators (B), and includes an ambiguous pronoun *they* that refers to one of the entities. The task is to establish the correct referent for the pronoun. We refer to this task as pronoun resolution. The interpretation of the pronoun arises from the meaning of the words *fear/advocate*. In the first sentence, the state of fear is attributed to the city councilmen (they = city councilmen), and in the second example, the action of advocating violence is attributed to the demonstrators (they = demonstrators).

(1a) The city councilmen (A) refused the demonstrators (B) a permit because they **feared** violence.

(1b) The city councilmen (A) refused the demonstrators (B) a permit because they **advocated** violence.

The authors of the WSC assumed that humans would perform at an accuracy level close to 100% (Levesque et al., 2012). Empirical studies revealed a different picture. Bender (2015) showed that human participants achieve 92% accuracy on well-crafted WSC sentences in English. Participants reported several difficulties, including unfamiliarity with certain concepts, such as *crop duster* or *bassinet*. Unfamiliar words and concepts can lead to an increase in task demand and possibly lower accuracy rates. Moreover, the content of the question and whether it aligns with or contradicts participants' expectations and personal experience can also have an effect on accuracy. In one of the scenarios, oatmeal cookies were preferred to chocolate cookies. Some participants found this unnatural and chose chocolate cookies as the answer to the pronoun resolution task, even though this incorrect answer contradicted information in the scenario (see Bender (2015) for discussion).

When LLMs were evaluated on the original WSC datasets, they performed worse than humans. However, training on larger datasets and fine-tuning helped. Language models gradually reached an accuracy of 90% (Sakaguchi et al., 2021). The most recent LLMs perform at 94% accuracy levels when evaluated in English (Artkaew, 2025).

2.2 The WSC in Other Languages: Human and LLM Performance

As the use of the WSC for evaluation benchmarks grew in popularity, the original WSC datasets developed for English were translated into other languages. However, translations to typologically different languages proved to be challenging. One set of difficulties pertained to typological and grammatical differences between the source language (English) and other target languages. English does not encode grammatical gender, animacy, or formality levels, and this presents a translation challenge. Research teams approached these challenges in varying ways. For example, when translating the WSC to French, Amsili and Seminck (2017) made changes to the original examples to achieve naturalness. The same strategy is reported by Artkaew (2025) for Thai. On the other hand, the authors of the Japanese Winograd Schema Challenge, WSCR-ja, noted that some translations resulted in ungrammatical examples due to structural differences between English and Japanese, but they decided to keep the examples in the dataset (Shibata et al., 2015).

Another translation difficulty pertains to cultural knowledge. Artkaew (2025) observed that an English scenario about playing cards uses the expression 'run of good luck', which is not natural in Thai. Another example pertained to a game of tag and how the chaser can be identified. In both cases, Artkaew (2025) chose to modify the original scenarios to make them more culturally appropriate. In their discussion of the Japanese WSC, Shibata et al. (2015) also acknowledged culturally inappropriate examples.

Comparison between human performance on the translated sets and human performance on the comparable dataset in English reveals differences in accuracy rates. Artkaew (2025) found that humans achieve 88% accuracy on the Thai WSC, which is lower than the 92% accuracy level reported for the English WSC (cf. Bender, 2015). Artkaew (2025) suggests that these differences should be attributed to translation effects and the difficulty of adapting scenarios from English to other languages.

There are also interesting differences in how language models perform on translated datasets compared to models evaluated on the original English datasets. Hashimoto et al. (2023) use the WSCR-ja by Shibata et al. (2015) to fine-tune BERT, a language model. Model fine-tuning helps increase accuracy on certain tasks, such as pronoun resolution. They found that the accuracy level increased from 57% to 58%, a modest gain. In comparison, fine-tuning the English model on the corresponding dataset in English leads to more significant gains. In the case of model evaluation, different factors might affect performance, including model size and architecture. Hashimoto et al. (2023) explicitly discuss the quality of the translated examples in the WSCR-ja, including cases of mistranslation, unfamiliar words, and cultural concepts, as a possible reason for smaller accuracy gains of their model after fine-tuning on the WSCR-ja. Results of evaluating more recent models on translated WSC datasets also show that they underperform compared to the base rate for English models. Artkaew (2025) reports that the accuracy of the best performing LLM on the Thai WSC is only 79.65% (Claude-3-Opus), compared to 94% on the English WSC.

3 Study

In this study, we use the WSC in Japanese to evaluate LLM and human performance on pronoun resolution. We focus on three conditions: (i) the baseline condition, (ii) a condition that manipulates task demands and (iii) a condition that manipulates content effects. The results on the baseline condition allow us to establish how LLMs perform vis-à-vis humans in the default setting. The null hypothesis is that LLM performance will parallel human performance. In the condition that manipulates task demands, we create adversarial conditions for both humans and LLMs and predict that this will negatively affect their performance. In the condition that manipulates content effects, we create favorable conditions for LLMs but adversarial conditions for humans, and we expect that it will increase LLM accuracy rates.

3.1 Materials

Our stimuli are derived from the WSCR-ja set (Shibata et al., 2015). This dataset is a translation of the Definite Pronoun Resolution (DPR) set (Rahman and Ng, 2012). Unlike other WSC datasets, the DPR set scenarios were crowdsourced from undergraduate students in the US and many of the original criteria of the test were relaxed. WSCR-ja consists of 941 question pairs which are split into a train set (659 pairs) and a test set (282 pairs). We performed a qualitative analysis of the entire WSCR-ja test set. Consistent with the observations in the previous literature, we found that the data was not homogeneous. In addition to grammatical and well-formed sentences, there were ungrammatical sentences and scenarios that reference cultural concepts that might not be familiar to Japanese speakers with limited experience of US culture. The US cultural references are a carry-over from the original dataset and the translated dataset reflects the cultural background of the content creators. These different categories of data correspond to the three conditions we define below. In addition, we noticed substantial variation between stimuli. Some stimuli were syntactically complex, or had multiple pronouns and/or referents with non-core grammatical roles (indirect object or object of preposition). Since all these factors can affect performance on pronoun resolution (Kehler et al., 2008), we excluded such items from consideration. (See Appendix A.1 for examples of rejected stimuli and explanations for rejection). Based on our qualitative analysis, we formulated specific criteria for selecting stimuli for the three conditions and validated selected examples and their appropriateness for each condition with a linguistic consultant, a native speaker of Japanese. For each condition, we selected five scenarios for a total of ten sentences per condition. See Appendix A.2 for the full list of stimuli.

3.1.1 The 'Good' Condition

The baseline condition measures the performance on the pronoun resolution task in the absence of any other interfering factors. The stimuli in this condition adhered to the original considerations for the WSC dataset (Levesque et al., 2012). Specifically, subjects should not be able to induce correct answers based on selectional restrictions or word co-occurrence statistics. The scenario in (2) (from Kocijan et al., 2023) violates the selectional criteria rule since only women can be pregnant and only pills can be carcinogenic. The choice of the correct referent becomes trivial in this context. The scenario in (3) is ruled out based on statistical co-occurrence considerations (*racecar* and *fast* frequently co-occur)(from Kocijan et al., 2023).

> (2) The women stopped taking the pills because they were [pregnant/carcinogenic].
>
> (3) The racecar zoomed by the school bus because it was going so [fast/slow].

(4) is an example scenario selected for this category.

> (4a) ジョーはアダムより良い香りがする。彼は日頃からシャワーを浴びるからだ。 'Joe smells better than Adam since he showers regularly.'
>
> (4b) ジョーはアダムより良い香りがする。彼はめったにシャワーを浴びないからだ。 'Joe smells better than Adam since he hardly ever showers.'

3.1.2 The 'Grammar' Condition

The stimuli in this condition are designed to measure the effect of task demand on performance. There are different ways to manipulate task demand, but here we focus on the effect of grammar. Specifically, we hypothesized that syntactically incongruent stimuli will increase task demand and reduce accuracy rates. Scenarios were selected for the 'grammar' condition if at least one sentence in the pair is grammatically unacceptable or has been mistranslated so that the meaning is significantly different. Sentences may also not adhere to the original WSC constraints. (5) is an example set from this scenario. While the pronoun *she* might be an acceptable pronoun for a car in English, this is not the case for Japanese, resulting in (5a) being ungrammatical.

> (5a) シーラは古いポンコツ車を修理しようとした。彼女は30年も車に取り組んでいなかったにも拘らずだ。 'Sheila tried to repair the old jalopy, even though she had not worked on cars in three decades.'
>
> (5b) シーラは古いポンコツ車を修理しようとした。彼女は30年も走っていなかったにも拘らずだ。 'Sheila tried to repair the old jalopy, even though she had not run in three decades.'

3.1.3 The 'Culture' Condition

This condition is designed to test content effects on performance. Familiarity with specific cultural concepts as well as the lack thereof can affect accuracy ratings. For this condition, we selected scenarios that referenced US cultural concepts. We hypothesized that such scenarios will align with LLMs' competence, thus boosting their performance, but would disadvantage Japanese speakers. (6) is an example scenario selected for this category. In this scenario, 'Autobot', 'Decepticon' and the world of the Transformers movies are references from US pop culture, which might not be familiar to speakers of Japanese. While an English speaker unfamiliar with Transformers may be able to associate 'Decepticon' with evil motives because of the similarity to 'deceive', Japanese speakers may not benefit from this clue.

(6a) オートボットはデセプティコンを食い止めようとする。彼らは世界の人々が平和に暮らすことを望んでいるのだ。 'The Autobots try to stop the Decepticons since they want the world to live in peace.'

(6b) オートボットはデセプティコンを食い止めようとする。彼らは世界を破壊したがっているからだ。 'The Autobots try to stop the Decepticons since they want to destroy the world.'

3.2 Participants

23 native Japanese speakers participated in the study. Participants were recruited via academic snowballing in Japan with two starting nodes. The average age was 29. Nine participants were male, eight were female and six did not state their sex.

3.3 Design and Procedure

Human participants accessed the survey hosted on Qualtrics via an anonymized link. They provided consent to participate in research and confirmed that they were of age and native speakers of Japanese. The participants saw 30 questions that tested their performance on the pronoun resolution task.[1] Participants saw the stimuli presented in random order and had to pick one of two answer options. The answer options were also presented in random order. There were no filler items, and participants were not given any training examples to maintain consistency with the LLMs' evaluation format. Task instructions and an example question can be found in Appendix A.3.

3.3.1 Collecting Data From LLMs

Our LLM data came from the responses of the GPT-4o model, the most advanced LLM at the time of research, collected from the OpenAI API. We chose the API rather than the chat interface because it allows us to control the model parameters (GPT-4o, temperature=1). We used the same design as in the study with human participants. The same stimuli were submitted to the OpenAI API. The order of questions and order of answers were randomized. We ran the code 30 times. Recent studies emphasize the need for a 'fair' evaluation of humans and machines with the emphasis on the same training and conditions for both groups (Lampinen, 2024). We follow this recommendation here. LLMs were evaluated zero-shot, and humans did not receive any prior training.

[1] We also collected naturalness judgements and recorded reaction time, but these data are not the main focus of the paper.

3.4 Results

We coded all correct answers as 1 and all incorrect answers as 0 for both humans and LLMs. Comparison of the means showed that in the 'good' condition, humans outperformed GPT-4o on the pronoun resolution task ($M_{good\_human} = 0.92$; $M_{good\_GPT} = 0.79$). In the 'grammar' condition, humans and LLMs performed similarly ($M_{grammar\_human} = 0.63$; $M_{grammar\_GPT} = 0.61$), and in the 'culture' condition, GPT-4o outperformed humans ($M_{culture\_human} = 0.92$; $M_{culture\_GPT} = 0.97$). The means and standard deviations are shown in Table 1.

To analyze the data, we applied a mixed-effects model, using the "lmerTest" package in R (Kuznetsova et al., 2017). Subject id and question were entered as random intercepts, while condition (good, grammar, culture) and source (human, GPT) were entered as fixed factors. The statistical analysis revealed a significant interaction between the two fixed factors (z(1523) = 2.68, p<.01). Follow-up tests showed that the statistical interaction was coming from the better performance of humans in the good condition (z(506) = 4.95, p<.001) and the better performance of GPT-4o in the culture condition (z(511) = -2-53, p<.05). The results from the study are presented in Figure 1.

3.5 Discussion

We observe that the overall accuracy (83%) displayed by human subjects in Japanese is lower than that reported for humans in English (92%). While this aligns with the lower accuracy levels reported for Thai (88%), it is important to point out that human performance in our study varies significantly depending on the condition. On well-formed grammatical examples in the baseline condition, the accuracy rates are 92%, similar to what is reported for English. Our study reveals that the translated dataset is not homogeneous and that examples with syntactic incongruencies can dramatically affect

	Humans	GPT-4o
Good	M=0.92 (SD=0.13)	M=0.78 (SD=0.35)
Grammar	M=0.63 (SD=0.44)	M=0.61 (SD=0.51)
Culture	M=0.92 (SD=0.08)	M=0.97 (SD=0.08)
Overall	M=0.83 (SD=0.30)	M=0.79 (SD=0.38)

Table 1: Means and standard deviations for humans and GPT-4o across the three conditions

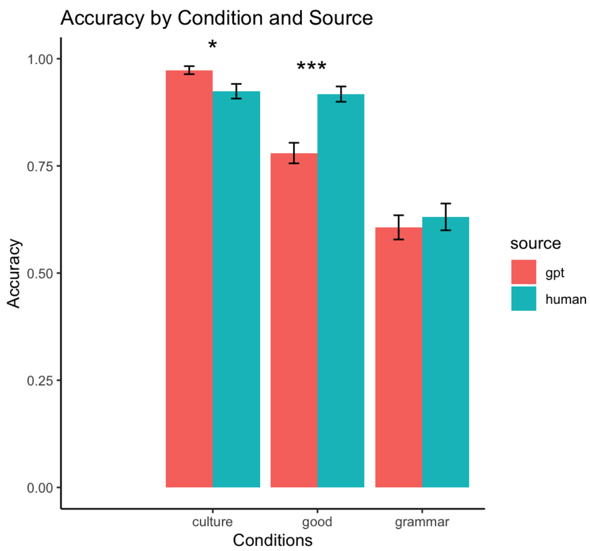


Figure 1: Accuracy of human and GPT judgements as a function of condition. Humans outperformed GPT in the 'good' condition, while the pattern was reversed for the 'culture' condition. No statistical difference was observed in the 'grammar' condition, where both sources performed poorly. The error bars represent +/-1 standard error. The significance tests are based on a mixed-effect model: * p<0.05, *** p<.001.

accuracy rates. These factors should be taken into account when evaluating either humans or LLMs on translated datasets.

Another implication from our study pertains to the potential applications of LLMs in contexts that require proficiency in Japanese. Previous studies have discussed leveraging LLMs' knowledge of Japanese in educational contexts for student writing assessment (Li and Liu, 2024, Takeuchi and Okgetheng, 2024) or example sentence generation (Benedetti et al., 2024). Our study demonstrates that the most advanced language models, such as GPT-4o, perform similarly to humans on tasks that require linguistic proficiency, which opens the opportunity for their integration in everyday life. However, the findings by Riccardi et al. (2024) that LLMs tend to interpret nonsensical input as meaningful, suggest that we should be cautious in applying them not only in language education contexts, but also in other linguistic tasks, such as text summarization (Gu et al., 2024) and annotation (Nishikawa and Koshiba, 2024).

Finally, we note that more insights could be gained from a systematic analysis of LLM mistakes. While this is outside of the scope of this paper, future work should look at these trends in more detail and compare the capabilities of different models, particularly those fine-tuned for Japanese.

4 Conclusions

In this study we compared the performance of LLMs and humans on a pronoun resolution task. We manipulated task demands and content effects and compared how they affect LLMs and humans. We found that in the baseline condition, humans outperform GPT-4o. These findings align with the results in Reese and Smirnova (2024) for Japanese, and with the results for Thai reported in Artkaew (2025). They suggest that in lower-resourced languages, humans still perform better than LLMs, even when competing with the most advanced models, such as GPT-4o.

Our results also provide evidence for task demands and content effects. In the relevant condition, task demands increased because of incongruent syntax/bad grammar. This manipulation negatively affected both human and LLM performance. Our results align with the observation in Hu and Frank (2024), who demonstrated that as task demands increase, LLM performance suffers, by analogy to how increased cognitive load in humans leads to reduced accuracy.

We manipulated content effects through cultural references. We selected scenarios with US cultural references, thus creating favorable conditions for LLMs, which were likely exposed to this information during training, and adversarial conditions for humans, as Japanese speakers might not be familiar with these references. We found that the changes in performance followed our predictions. In this condition, LLMs outperformed humans, providing evidence for content effects.

References

Pascal Amsili and Olga Seminck. 2017. A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Phakphum Artkaew. 2025. Thai Winograd schemas: A benchmark for Thai commonsense reasoning. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 42–51, Online. Association for Computational Linguistics.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference 2015, Greensboro, NC, USA, April 25-26, 2015*, volume 1353 of *CEUR Workshop Proceedings*, pages 39–45. CEUR-WS.org.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024. Automatically suggesting diverse example sentences for l2 japanese learners using pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131. Association for Computational Linguistics.

Jacob Browning and Yann LeCun. 2023. Language, common sense, and the winograd schema challenge. *Artificial Intelligence*, 325:104031.

Vittoria Dentella, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2024. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14(1):28083.

Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. An ai-resilient text rendering technique for reading and skimming documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ryo Hashimoto, Masashi Takeshita, Rafal Rzepka, and Kenji Araki. 2023. Development of japanese wsc273 winograd schema challenge dataset and comparison between japanese and english bert baselines. In *In the proceedings of the Language Technology Conference (LTC'23)*, pages 91–95.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University. url: https://lingbuzz. net/lingbuzz/007190*.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the winograd schema challenge. *Artificial Intelligence*, 325:103971.

Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561, Rome, Italy. AAAI Press.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Kai Nishikawa and Hitoshi Koshiba. 2024. Exploring the applicability of large language models to citation context analysis. *Scientometrics*, pages 1–27.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

May Lynn Reese and Anastasia Smirnova. 2024. Comparing chatgpt and humans on world knowledge and common-sense reasoning tasks: A case study of the japanese winograd schema challenge. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Nicholas Riccardi, Xuan Yang, and Rutvik H Desai. 2024. The two word test as a semantic benchmark for large language models. *Scientific Reports*, 14(1):21593.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2015. Nihon go winograd schema challenge no kochiku to bunseki. In *Proceedings of the 21th Annual Meeting of the Association for Natural Language Processing*, pages 493–496, Kyoto. The Association for Natural Language Processing.

K Takeuchi and B Okgetheng. 2024. Estimating japanese essay grading scores with large language models. *Lan-guage Resources and Evaluation*, 58(2):345–367.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

A Appendices

A.1 Examples of Rejected Items

Rejected Item: リチャードはカーソン上院議員を脅した。彼の沈黙が守られるように。 Richard blackmailed Senator Carson so that his silence would be secured.
Reason for Rejection: This item was considered for the good category, but rejected based on the intuition of our Japanese native speaking consultant that 沈黙, silence, is used in an unnatural way.

Rejected Item: 火事についての記事によれば、それらによってシアトルの大部分に火がついた。 The article about the fires said that they torched most of Seattle.
Reason for Rejection: This item was considered for the syntax category but was rejected because while ‘the article’ has the syntactic function of subject, ‘the fires’ is neither the subject nor the object of the first clause.

Rejected Item: ハーヴィー・デントは恋人を失ったことをバットマンのせいにする。彼が彼女を救出しなかったからだ。 Harvey Dent blames the Batman for losing his love because he did not save her.
Reason for Rejection: This item was considered for the culture category, but was rejected because there are multiple pronouns in the second clause.

Rejected Item: 私たちは人間に果物を与えた。それらが熟していたからだ。 We gave the fruit to the humans because they were ripe.
Reason for Rejection: This item was rejected because in Japanese, it has a selectional clue to the answer. それ, it, can only be used to refer to inanimate objects.

Rejected Item: 猫が人間を襲った。彼らは野生化していたのだ。 The cats attacked the humans because they were feral.
Reason for Rejection: This item was rejected because it can be solved by co-occurrence statistics. The adjective feral is more often associated with cats than humans.

A.2 List of Stimuli

Good Category Stimuli

学外のアパートは学内のアパートより好まれた。それらの方が安かったからだ。安かった方はどちらですか？ The off-campus apartments were preferred to the on-campus apartments because they were cheaper. Which were cheaper?
Answer: 学外のアパート the off-campus apartments

学外のアパートは学内のアパートより好まれた。それらの方が高かったからだ。高かった方はどちらですか？ The off-campus apartments were preferred to the on-campus apartments because they were more expensive. Which were more expensive?
Answer: 学内のアパート the on-campus apartments

ジョーはアダムより良い香りがする。彼は日頃からシャワーを浴びるからだ。日頃からシャワーを浴びるのは誰ですか？ Joe smells better than Adam since he showers regularly. Who showers regularly?
Answer: ジョー Joe

ジョーはアダムより良い香りがする。彼はめったにシャワーを浴びないからだ。めったにシャワーを浴びないのは誰ですか？ Joe smells better than Adam since he hardly ever

showers. Who hardly ever showers?
Answer: アダム Adam

ジャックはジョンより多く得票した。人々は彼を好んだ。人々が好んだのは誰ですか？ Jack got more votes than John because the people liked him. Who did the people like?
Answer: ジャック Jack

ジャックはジョンより多く得票した。人々は彼を好まなかった。人々が好まなかったのは誰ですか？ Jack got more votes than John because the people did not like him. Who did the people not like?
Answer: ジョン John

アダムはアレクサンダーの殺害に失敗した。そこで彼は再度のために暗殺者を雇った。再度のために暗殺者を雇ったのは誰ですか？ Adam failed to kill Alexander, so he hired an assassin for the second attempt. Who hired an assassin for the second attempt?
Answer: アダム Adam

アダムはアレクサンダーの殺害に失敗した。そこで彼は再度を恐れてボディーガードを雇った。再度を恐れてボディーガードを雇ったのは誰ですか？ Adam failed to kill Alexander, so he hired a bodygaurd in case of a second attempt. Who hired a bodyguard in case of a second attempt?
Answer: アレクサンダー Alexander

トニーはジェフを手伝った。彼は手伝いたかったのだ。手伝いたかったのは誰ですか？ Tony helped Jeff because he wanted to help. Who wanted to help?
Answer: トニー Tony

トニーはジェフを手伝った。彼は手助けが必要だったからだ。手助けが必要だったのは誰ですか？ Tony helped Jeff because he needed help. Who needed help?
Answer: ジェフ Jeff

Syntax Category Stimuli

バットはボールに当たった。それが軌道を描くように飛んだからだ。軌道を描くように飛んだのは何ですか？ The bat hit the ball because it flew in the way of the trajectory. What flew in the way of the trajectory?
Answer: バット the bat
Note from the translators: ボールでも？ The ball too?

バットはボールを打った。それは可哀想な動物に向かってまっしぐらにとんだからだ。可哀想な動物に向かってまっしぐらにとんだのは何ですか？ The bat hit the ball because it flew straight at the poor animal. What flew straight at the poor animal?
Answer: ボール the ball

シーラは古いポンコツ車を修理しようとした。彼女は30年も車に取り組んでいなかったにも拘らずだ。30年も車に取り組んでいなかったのはどちらですか？ Sheila tried to repair the old jalopy, even though she had not worked on cars in three decades. Who had not worked on cars in three years?
Answer: シーラ Sheila

シーラは古いポンコツ車を修理しようとした。彼女は30年も走っていなかったにも拘らずだ。30年も走っていなかったのはどちらですか？ Sheila tried to repair the old jalopy, even though she had not run in three decades. Who had not run in three decades?
Answer: 古いポンコツ車 the old jalopy

りんご酒がわたしの口に入った。それは美味しかったから。美味しかったのは何ですか？ The apple wine entered my mouth because it tastes good. What tastes good?
Answer: りんご酒 the apple wine

りんご酒がわたしの口に入った。それは一杯ではなかったから。一杯ではなかったのは何ですか？ The apple wine entered my mouth because it was not full. What was not full?
Answer: わたしの口 my mouth

雇用主はケイティに仕事を提供した。彼女はインタビューが好きだったからだ。インタビューが好きだったのは誰ですか？ The employer offered Katie a job, because she liked the interview. Who liked the interview?
Answer: 雇用主 the employer

雇用主はケイティに仕事を提供した。彼女が会社にぴったりだったからだ。会社にぴったりだったのは誰ですか？ The employer offered Katie a job, because she was a fit for the

company. Who was a fit for the company?
Answer: ケイティ Katie

ジョーはマイクに倒れ掛かった。彼は眠る場所が必要だった。眠る場所が必要だったのは誰ですか？ Joe crashed into Mike because he needed a place to sleep. Who needed a place to sleep?
Answer: ジョー Joe

ジョーはマイクに衝突した。彼は損害分を支払わなくてはならなかった。損害分を支払わなくてはならなかったのは誰ですか？ Joe crashed into Mike and he had to pay for the damage. Who had to pay for the damage?
Answer: マイク Mike
Note from translators: ジョーでも？ Could be Joe too?

Culture Category Stimuli

ワトソンはジオパディでケンを負かした。彼は優れた機械だ。優れた機械は誰ですか？ Watson beat Ken at Jeopardy because he is a superior machine. Who is a superior machine?
Answer: ワトソン Watson

ワトソンはジオパディでケンを負かした。彼は劣った人間だからだ。劣った人間は誰ですか？ Watson beat Ken at Jeopardy because he is an inferior human. Who is an inferior human?
Answer: ケン Ken

ビリーはスクラブルでトミーを負かした。あの新入りには運がついていた。運がついていたのは誰ですか？ Billy beat Tommy at Scrabble because that newbie had all the luck. Who had all the luck?
Answer: ビリー Billy

ビリーはスクラブルでトミーを負かした。あの新入りには能力がなかったから。能力がなかったのは誰ですか？ Billy beat Tommy at Scrabble because that newbie had no skill. Who had no skill?
Answer: トミー Tommy

オートボットはデセプティコンを食い止めようとする。彼らは世界の人々が平和に暮らすことを望んでいるのだ。世界の人々が平和に暮らすことを望んでいるのは誰ですか？ The Autobots try to stop the Decepticons since they want the world to live in peace. Who wants the world to live in peace?
Answer: オートボット the Autobots

オートボットはデセプティコンを食い止めようとする。彼らは世界を破壊したがっているからだ。世界を破壊したがっているのは誰ですか？ The Autobots try to stop the Decepticons since they want to destroy the world. Who wants to destroy the world?
Answer: デセプティコン the Decepticons

メアリはジョーが好きだ。彼女は女性が好きだからだ。女性が好きなのは誰ですか？ Mary likes Joe because she likes females. Who likes females?
Answer: メアリ Mary

メアリはジョーが好きだ。彼女は名前が素敵だからだ。名前が素敵なのは誰ですか？ Mary likes Joe because she has a cool name. Who has a cool name?
Answer: ジョー Joe

カリフォルニアの人の方がニューヨークの人より良い。彼らにはハリウッドがあるから。ハリウッドがあるのは誰ですか？ Californians are better than New Yorkers because they have Hollywood. Who has Hollywood?
Answer: カリフォルニアの人 Californians

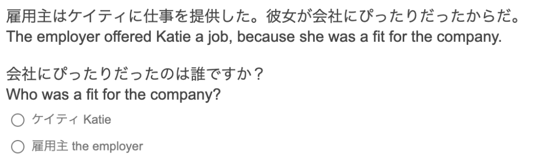
カリフォルニアの人の方がニューヨークの人より良い。彼らには映画を作ってくれるハリウッドの連中がいないからだ。映画を作ってくれるハリウッドの連中がいないのは誰ですか？ Californians are better than New Yorkers because they do not have Hollywood lots to produce movies. Who does not have Hollywood to produce movies?
Answer: ニューヨークの人 New Yorkers

A.3 Task Instructions and Prompt Examples

Task Instructions for Human Participants:

> 2つの日本語の文章と、その文章の内容に関する質問と２つの答えが表示されています。2つの答えの内正しいと思う方を選んでください。どちらの答えも妥当と思われる場合

は、最も適切と思われる方を選んでください。
You will be shown two Japanese sentences and a question with two answers about the content of the sentences. Please choose the answer you think is correct. If both options seem right, please pick the one you think is the most fitting.


雇用主はケイティに仕事を提供した。彼女が会社にぴったりだったからだ。
The employer offered Katie a job, because she was a fit for the company.

会社にぴったりだったのは誰ですか？
Who was a fit for the company?

◯ ケイティ Katie

◯ 雇用主 the employer

Figure 2: Example survey question shown to human participants. (The English translation was not shown to participants)

Example Prompt for GPT-4o:

Japanese Prompt:
学外のアパートは学内のアパートより好まれた。それらの方が安かったからだ。 次の問題をAかBで答えてください。安かった方はどちらですか？A.学外のアパート B.学内のアパート
English Translation:
The off-campus apartments were preferred to the on-campus apartments because they were cheaper. Answer the following question with A or B. Which were cheaper? A. The off campus apartments B. The on campus apartments

(The English translation was not given in the prompt.)

Generative AI Teaching Simulations as Formative Assessment Tools within Preservice Teacher Preparation

Jamie N. Mikeska[1], Aakanksha Bhatia[2], Shreyashi Halder[1], Tricia Maxwell[1], Beata Beigman Klebanov[1], Benny Longwill[1], Kashish Behl[1], Calli Shekell[3]
[1]ETS Research Institute, [2]ExcelOne, [3]Pennsylvania Western University, Clarion campus
jmikeska,shalder001,tmaxwell,bbeigmanklebanov,blongwill,kbehl@ets.org;
aakankshabhatia01@gmail.com;shekell_c@pennwest.edu

Abstract

This paper examines how a generative AI (GenAI) teaching simulation can be used as a formative assessment tool to gain insight into preservice teachers' (PSTs') instructional abilities. Our team investigated the teaching moves PSTs used to elicit student thinking in a GenAI simulation and their perceptions of the simulation's usefulness.

1 Introduction and Study Aims

Most applications of GenAI in educational contexts during the last year have occurred within K-12 settings, where the primary focus has been on applications that directly support student learning (Chiu, 2025; Mintz et al., 2023). Yet, GenAI also has potential to provide meaningful learning opportunities to teachers to support them in improving their instructional skills, knowledge, and abilities (Lee & Yeo, 2022; Lim et al., 2025; Mikeska & Bhatia, 2025). In this study, our cross-disciplinary team of researchers in teacher learning and educational technology, assessment developers, AI engineers, subject matter experts, and teacher educators collaborated on developing and deploying a GenAI teaching simulation where PSTs could prepare for, engage in, and reflect on their ability to engage in one core teaching practice: elicit and attend to student thinking.

Our team examined how this GenAI teaching simulation could be used as a formative assessment tool to identify the nature of the teaching moves that the PSTs used to elicit and attend to student thinking and the PSTs' perceptions of the simulation's usefulness to support PST teacher learning when integrated within an educator preparation program. By formative assessment, we focus on how the GenAI teaching simulation can be used to gather evidence that can help PSTs understand their instructional strengths and areas for growth and to determine how they could adjust their teaching moves in future instruction (Irons & Elkington, 2021). The main research questions addressed in this study are: (1) What are the teaching moves that elementary PSTs use to elicit and attend to student thinking in a GenAI teaching simulation? and (2) What are PSTs' perceptions of the simulation's usefulness?

2 Background

2.1 Using Digital Teaching Simulations to Support Teacher Learning

While digital teaching simulations can vary in format and structure, most provide PSTs and in-service teachers with opportunities to try out aspects of the teaching within settings of reduced complexity (Dieker et al., 2014; Ersozlu et al., 2021). Digital teaching simulations have been used to support PSTs and in-service teachers in learning how to elicit student thinking, facilitate productive discussions, manage the classroom, and engage with students who are multilingual learners or have special needs (Bondie et al., 2021; Lee et al., 2024; Mikeska et al., 2021). For example, TeachLivE and Mursion use an online simulated classroom that is comprised of up to five student avatars who can interact in real time with the teacher and each other verbally; currently there are multiple simulated classrooms available including an early childhood classroom, upper elementary classroom, middle school classroom, and high school classroom. Other teaching simulations, such as SchoolSims, use an online environment where teachers read through specific scenarios and then are provided

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 212–220
October 27-29, 2025

opportunities to make a series of instructional decisions via text-based choices and observe the impact of those decisions.

During the last couple decades, a growing number of research studies have provided empirical evidence illustrating how digital teaching simulations can be integrated productively within educator preparation programs and professional development contexts. Studies have shown that these simulations can be used to improve several outcomes including PSTs' and in-service teachers' ability to engage in core teaching practices, their instructional beliefs, and their content knowledge for teaching (Mikeska et al., 2023; Pecore et al., 2023; Straub et al., 2015). Other studies have suggested that it is important to embed the use of such simulations within learning cycles where teachers have opportunities to prepare for, engage in, and reflect on their simulated teaching experiences, as well as to provide formative feedback to teachers so they can understand and reflect on their instructional strengths and areas for growth (Cohen et al., 2020; McDonald et al., 2013; Mikeska et al., 2021). However, one challenge across this line of research has been the fact that the current simulations require significant human resources to develop and deploy, especially since many of them require a human-in-the-loop to power the student avatars. The recent advances in GenAI offer a potential solution to this challenge – one which we explore in this study by examining the potential of a GenAI teaching simulation as a formative assessment tool with an elementary mathematics methods course.

2.2 Evaluating Teacher Performance

Skilled teaching is critical for positive student outcomes (Blömeke et al., 2016; Fauth et al., 2019). The need for reliable instruments for measuring teacher performance to help them improve has been recognized as a major issue in teacher education research (Correnti et al., 2015). One of the more influential frameworks in this area is the Accountable Talk Theory (Michaels et al., 2008) that provides a protocol for classifying teacher and student contributions to classroom discourse into categories defined by the purpose of each 'move'. For example, teacher talk moves include repeating what the student said and pressing the student for reasoning, while student talk moves include asking for information and relating to what another student said. Talk moves can be reliably identified (Suresh et al., 2022a). More recently, there is work on automating the coding of talk moves and similar constructs to support feedback to teachers (Demszky 2023; Nazaretsky et al., 2023; Suresh et al., 2022b; Tran et al., 2024). Since the protocols are designed to apply across a variety of classroom discussions, eliciting student thinking is only a part of what the teacher does in the bigger picture of facilitating classroom discussions. In this work, we "zoomed in" on the elicitation activity in more detail, since this specific practice is the focus of the simulation. Furthermore, differently from a real classroom, we control the "students" in the simulation by giving them task-specific knowledge profiles that include specific understandings and misunderstandings. As such, we are in a position to evaluate which of the specific points the teacher actually elicited. We therefore used a protocol that combined general categories similar to those in the talk moves literature that pertain to elicitation (e.g., ask questions tied to student actions) and highly content specific categories that focus on unlocking points of understanding or misunderstanding in the simulation (e.g., the student does not understand the commutative property in addition); we call this protocol an "evidence inventory." This two-pronged approach is designed to support feedback both about general tendencies (how often the teacher attends closely to the students' ideas) and about the effectiveness of the elicitation – whether the teacher actually identified the specific pre-designed aspects of the GenAI student's thinking.

3 Study Methodology

3.1 Study Sample

Ten elementary PSTs who were enrolled in an elementary mathematics methods course as part of their educator preparation program at a U.S. university located in the Northeast participated in this study. All PSTs were between 18 to 24 years old and spoke English as their first language. Half of the PSTs had some previous teaching experience via substitute teaching (2 PSTs), as an after school coordinator (1 PST), or as a mentor to elementary students (2 PSTs). None of the PSTs had any previous experience participating in professional learning focused on AI, educational technology, or digital teaching simulations.

3.2 Data Collection

In this study, the elementary teacher educator integrated the GenAI teaching simulation into their elementary mathematics methods course in Spring 2025 at two different timepoints within a two-week window. At each timepoint, the PSTs had a chance to prepare for, engage in, and then reflect on their GenAI simulated teaching session. Details about the preparation and reflection activities are reported in Mikeska, Beigman Klebanov et al. (2025). Each session used the same GenAI teaching simulation, which we call the Strategies for Adding task.

In this task, PSTs learn that a class of first grade students have been working on learning about strategies for adding numbers within 20 and one student named Cecilia recently solved the following problem: *Mike has 6 crayons. Ann has 8 crayons. How many crayons do they have in all?* The PST's goal in the GenAI simulation is to: (1) ask questions to elicit what the student (Cecilia) did to produce the answer given and (2) probe to understand why the student (Cecilia) performed the particular steps and what conceptual understanding the student has and does not have regarding addition and regarding adding numbers within 20. As part of their preparation, each PST is instructed to review Cecilia's written work (see Figure 1) and prepare by considering ways they could elicit the following: what Cecilia did to produce the answer given, why Cecilia performed the particular steps, and what conceptual understanding Cecilia does and does not have regarding adding numbers within 20, including posing other problems to elicit or confirm Cecilia's understanding.

When ready the PST enters the online environment and begins having a verbal conversation with Cecilia to practice eliciting her thinking about the problem she solved and her understanding in this topic area. Figure 1 shows an image of the Strategies for Adding GenAI simulation interface, as well as shows Cecilia's written work where she drew 8 circles, put dots in three of the circles, and wrote a number sentence underneath the picture (6+2=8).

During the GenAI simulation, all of Cecilia's responses are powered by GenAI. Our team used prompt engineering via GPT-4o on Microsoft's Azure OpenAI service to develop and deploy this GenAI simulation. One of the key resources we leveraged was an already developed human-led simulation task and training protocols, from a previous project, which we then used to develop the initial generation prompt. The initial generation prompt included two parts – instructions and few-shot examples -- to create the response that Cecilia, the GenAI student, would provide during the simulation. Details about the specific prompt used

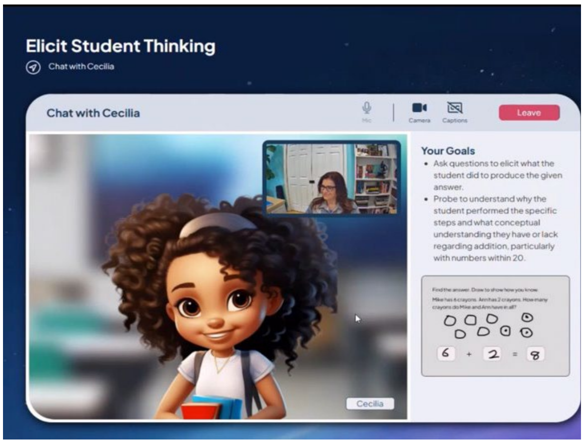


Figure 1. Strategies for Adding Online Interface

and user testing that our team engaged in to refine the prompt for use within teacher learning contexts can be found in Mikeska, Beigman Klebanov et al. (2025). Previous research indicated that the GenAI student's (Cecilia's) responses in the simulation were: consistently aligned with Cecilia's conceptual understanding and addition problem solving process; age and grade level appropriate; responsive to the teachers' questions and prompts; and coherent across the conversation (Mikeska, Beigman Klebanov et al., 2025).

Chatbot response generation using GPT-4o (v2024-08-06) followed a structured pipeline designed to ensure safety, contextual relevance, and alignment with pedagogical constraints. It began with a request to Microsoft Azure's Chat Completions API, using a system prompt tailored to the GenAI student's profile, few-shot dialogue examples to model interaction style, and the full chat history for context. The API would return a response that has already passed a built-in moderation filter for harmful content. This output then underwent additional validation and transformation steps to reinforce behavioral consistency, ensure educational appropriateness, and reduce the unpredictability of large language model outputs before being presented to the PST.

Primary data sources for this study included written transcripts from each PST's GenAI simulation conversation and survey responses after each session. Each transcript included the utterances from the PST and Cecilia during the conversation (see Appendix A for one example conversation). After each of the two reflection activities, our research team administered an online survey to the PSTs that used both Likert and open-ended questions to gather data about the PSTs' understanding of the GenAI student's thinking and their perceptions of the simulation's authenticity, usability, and usefulness. This study reports on findings from survey questions that asked about the PSTs' perceptions on the usefulness of GenAI teaching simulations within PST learning contexts. Most questions used a Likert scale with Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree as choices for PSTs to select in response to specific statements (e.g., GenAI teaching simulations are a useful tool to support elementary PSTs' learning) while one was an open-ended question about what improvements were needed to the GenAI simulation to best support PST learning.

3.3 Data Analysis

Since each PST engaged in the GenAI simulation at two different timepoints, there were a total of 20 transcripts and survey responses used in the data analysis. To address the first research question, our team used a previously developed evidence inventory rubric to code for the presence or absence of key teaching moves that the PSTs could use in this GenAI simulation to engage in productive aspects of eliciting student thinking. For example, PSTs could use questions or prompts to elicit information about several aspects of Cecilia's problem solving process, including eliciting that Cecilia drew 6 circles and then 2 circles or that Cecilia solved the problem by counting on from 6 (e.g., How did you count to figure out how many crayons they had in all?), and her understanding within this topic area, including that she cannot fluently add the numbers, does not understand the commutative property, and understands what the six, two, and eight represent. These very content-specific categories were coded for Cecilia's turns, namely, where Cecilia's utterance provides evidence that the teacher has successfully elicited this particular element of Cecilia's mathematical thinking. In parallel, PSTs could also use various teaching moves to attend to Cecilia's responses and use them as a basis for further questions, such as asking questions tied to specific things that Cecilia did (e.g., Why did you count on from 6?), and to use follow-up questions or prompts to provide opportunities for Cecilia to explain her reasoning or understanding, such as having Cecilia describe her work and explain aloud (e.g., Why did you only draw dots in three of the circles?). These categories were annotated for the PST's turns and were not tied to the specifics of the mathematical knowledge involved (e.g., "ask Cecilia to explain her reasoning" would be marked the same whether it is about the order of the addends or the use of dots in the circles).

Two raters used the evidence inventory rubric to code for the absence or presence of 18 different teaching moves within the 20 transcripts. If raters noted that specific teaching moves were present in a particular transcript, then they also identified the specific utterances in the transcript that served as evidence of each teaching move. The coding process involved the two raters initially meeting to collectively score one transcript to develop a shared understanding of the 18 teaching moves and the coding process. Then, each rater individually coded the remaining 19 transcripts and then met to reconcile and reach consensus on any individual code applications where they initially disagreed. Overall, the two raters achieved 96.5% exact agreement on the code applications for the presence or absence of these teaching moves across the 19 transcripts and 84.8% agreement for identifying the specific utterances for each teaching move that was identified as present. Finally, we calculated the number and percentage of transcripts that had these teaching moves represented at each timepoint and used the descriptive frequencies to identify the PSTs' strengths and areas for growth within and across timepoints.

To address the second research question, we calculated descriptive frequencies of PSTs' responses to the Likert scale questions about the GenAI simulation's usefulness. Then, we conducted qualitative content analysis (Schreier, 2012) of the PSTs' responses to the open-ended question and calculated descriptive frequencies by codes applied to identify patterns in their responses.

4 Results

4.1 Teaching Moves Used in a GenAI Simulation

Tables 1 and 2 provide the results for the extent to which these PSTs engaged in specific teaching moves, as evidenced by the GenAI student utterances or PST utterances, respectively. These teaching moves were used by the PSTs to elicit the GenAI student's thinking about the process she used and her conceptual understanding about strategies for adding within 20 (Table 1) and to attend to and follow-up on the student's reasoning (Table 2). The results indicate the number and percentage of PSTs (out of 10 PSTs) at each timepoint who exhibited the specific teaching moves in their conversation with the GenAI student. These results indicate several strengths and areas of growth across this group of PSTs.

Table 1. *Teaching Moves to Elicit the GenAI Student's Thinking*

	Teaching Moves (evidenced by the GenAI student utterances)	Timepoint 1 (n=10 PSTs) n (%)	Timepoint 2 (n=10 PSTs) n (%)
Focused on the Student's Process	Elicits that the student draws 6 circles and then 2 circles	5 (50%)	6 (60%)
	Elicits that the student draws Mike's crayons first because that is the first number in the problem	0 (0%)	0 (0%)
	Elicits that the student draws Ann's crayons second because that is the second number in the problem	0 (0%)	0 (0%)
	Elicits that the student solves the problem by counting on from 6	5 (50%)	10 (100%)
	Elicits that the student always counts on from one of the numbers in the problem	6 (60%)	10 (100%)
Focused on the Student's Understanding	Elicits that the student cannot fluently add the numbers	3 (30%)	5 (50%)
	Elicits the student's understanding of the commutative property	1 (10%)	0 (0%)
	Elicits the student's understanding of what the 6 represents	8 (80%)	10 (100%)
	Elicits the student's understanding of what the 2 represents	6 (60%)	9 (90%)
	Elicits the student's understanding of the 8	2 (20%)	1 (10%)
	Elicits the student's understanding that the first addend name summarizes the procedure of counting all of the circles representing that addend	5 (50%)	10 (100%)
	Elicits the student's understanding of the plus symbol	1 (10%)	0 (0%)

First, results suggest that by the second timepoint, all PSTs were able to engage in one or more productive teaching moves to elicit information about the GenAI student's process and conceptual understanding. In particular, the PSTs were most likely to be able to elicit: (a) how Cecilia always counted on from the first addend to solve the addition problem, (b) Cecilia's understanding of what the two addends (six and two) represent, and (c) Cecilia's understanding that the first addend name (six) summarizes the procedure of counting all the circles representing that addend.

For example, one PST asked Cecilia about how she solved the problem and counted; Cecilia replied, "I drew 6 circles for Mike's crayons. Then I drew 2 circles for Ann's crayons. Then I counted 6, 7, 8." Similarly, another PST prompted Cecilia to talk about what the 6 and 2 represented in the number sentence to which Cecilia explained that "Mike's crayons were the first six circles and Ann's were the next 2 circles."

Table 2. *Teaching Moves Used to Attend to and Follow-up on Student's Reasoning*

	Teaching Moves (evidenced by the PST's utterances)	Timepoint 1 (n=10 PSTs) n (%)	Timepoint 2 (n=10 PSTs) n (%)
Focused on the Student's Process	Asks questions tied to specific things that the student did	9 (90%)	10 (100%)
	Attends to and makes use of specific ideas from what the student says	9 (90%)	9 (90%)
Focused on the Student's Understanding	Has the student show work and describe/explain aloud	9 (90%)	10 (100%)
	Poses one or more additional tasks that are clearly useful for the student to solve	1 (10%)	4 (40%)
	Asks questions that lead the student to a particular answer *	3 (30%)	5 (50%)
	Fills in answers for the student (e.g., a contribution that provides information that should have been elicited or probed for) *	1 (10%)	0 (0%)

*These teaching moves do not support the practice of eliciting student thinking.

Second, the results also highlight how these PSTs were quite adept – both at the first and second timepoints – at attending to the GenAI student's idea by asking questions about what Cecilia did and making use of specific ideas that Cecilia shared, as well as using questions to prompt Cecilia to describe and explain aspects of her work. For example, PSTs used various prompts to learn about the steps Cecilia took to solve this addition word problem by asking questions like: "I'd really like to learn too. Can you show me how you're working this problem? What's the first step?"; "Why did you choose that strategy?"; "Can you explain to me why you did the steps you did?"; and "So tell me how did you count on from six?"

Third, the results indicate that there are several areas of growth evident in these PSTs' ability to elicit and attend to student thinking. One of the most striking patterns is that the PSTs were less likely to elicit ideas related to gaps in the GenAI student's conceptual understanding. For example, only one PST (at timepoint 1) was able to successfully elicit that Cecilia did not understand the commutative property (e.g., that 6 + 2 is the same as 2 + 6). Similarly, only 3 PSTs and 5 PSTs at timepoints 1 and 2, respectively, were able to elicit that Cecilia could not do mental math and add numbers fluently in her head; instead, Cecilia always had to draw a picture to represent the addition word problem and then count on from the first addend to solve it.

4.2 Perceptions of the Usefulness of GenAI Simulations

Across both timepoints, most of the PST survey responses to the Likert scale questions indicated that they agreed that GenAI teaching simulations, like the one used in this study, are a useful tool to support elementary PSTs' learning (70% or 14 of 20 PST survey responses across simulation rounds) and can be used to help elementary PSTs better understand student thinking and students' learning needs (85% or 17 of 20 PST survey responses across simulation rounds). There was also strong support that the experience of eliciting student thinking in the simulation closely resembled the work that elementary teachers do to support teaching in real classrooms (75% or 15 of 20 PST survey responses across simulation rounds) and the content addressed in the GenAI simulation was appropriate for elementary PSTs (85% or 17 of 20 PST survey responses across simulation rounds).

In terms of improvements needed to make the GenAI teaching simulation a more effective tool to support PST learning, the qualitative content analysis identified three main ideas. First, in 7 of the 20 survey responses across simulation rounds, PSTs indicated that decreasing the GenAI simulation's latency so that the GenAI student responded more rapidly to the PSTs' questions and prompts would make this tool a more effective one. As one PST noted, "…the only improvement would be the time it took her to respond. The first time, I thought she might not have heard me." Second, in 10 of the 20 survey responses across simulation rounds, PSTs noted that it would be important in GenAI teaching simulations to increase variation in the GenAI student's profile and include additional simulations where students have different conceptual understanding. Finally, in 3 of the 20 survey responses across simulation rounds, PSTs mentioned that the GenAI student's actual responses could be improved to make the simulation more effective, such as by not having Cecilia "repeat herself as much."

5 Conclusion

This study serves as part of broader efforts in the field to determine how GenAI can be used in responsible ways for formative use. Our research context --the use of GenAI to power interactive, online simulations where PSTs can practice and receive formative feedback about their instructional strengths and areas for growth – is one that is currently underexplored, as most research in educational contexts focuses on developing and deploying GenAI tools to support K-12 student learning and outcomes. To ensure that such tools can be used responsibly for formative assessment within teacher learning contexts, a critical first step is ensuring that PSTs' interactions within the GenAI teaching simulations can provide information about PSTs' instructional strengths and areas for growth. It is also important to examine PSTs' perceptions of such tools, as they are more likely to engage with innovative tools if they view them as supportive of their learning.

Findings from this study suggest that GenAI teaching simulations have the potential to be used as formative assessment tools that can be integrated into PST learning contexts. In particular, in this study we developed and deployed a GenAI simulation that provided learning opportunities for PSTs to practice eliciting and attending to student thinking. The study's findings provided empirical evidence of the varied teaching moves these PSTs were able to use successfully to elicit information about the process the GenAI student used to solve the addition word problem and key aspects of her conceptual understanding in this topic area. In addition, the GenAI simulation helped to highlight areas of growth for these PSTs – namely in being able to better elicit gaps in a student's understanding. These findings align with previous research that has indicated teachers struggle to be able to pinpoint challenges that students have and sometimes fail to elicit nuanced information about students' conceptual understanding (Shaughnessy & Boerst, 2018; Sleep & Boerst, 2012).

Results were also promising in terms of the PSTs' mostly positive perceptions about the GenAI simulation's usefulness. Similar results have been reported regarding the use of human-in-the-loop teaching simulations, with results indicating that PSTs and in-service teachers value these simulations to provide content-focused practice spaces where they can improve their instructional capabilities without harming any real students. Ensuring that GenAI simulations provide authentic learning spaces for PSTs that mimic aspects of real classroom interactions is an important step to being able to integrate such tools into PST learning contexts.

Collectively, outcomes from this study suggest that GenAI can be used responsibly to provide a practice-based setting where PSTs can practice eliciting and attending to student thinking, and the outputs of the simulation interaction can be assessed formatively to identify the nature of the teaching moves that the PSTs use – or fail to use – to engage in this instructional practice. This formative information could be used in varied ways to support PST learning, such as incorporating the information into personalized feedback reports for PSTs or having PSTs reflect on the teaching moves they did and did not use to elicit and attend to student thinking after each simulation session. Future research can explore how PSTs make sense of and use this kind of formative information from GenAI teaching simulations to impact their instructional decision-making, can investigate the use of various large language models to power the GenAI student responses, and can examine the use of similar approaches in other content disciplines and topics.

References

Sigrid Blömeke, Rolf Vegar Olsen, and Ute Suhl. 2016. Relation of student achievement to the quality of their teachers and instructional quality. *Teacher Quality Instructional Quality and Student Outcomes*, 2: 21–50.

Rhonda Bondie, Zid Mancenido, and Chris Dede. 2021. Interaction principles for digital puppeteering to promote teacher learning. *Journal of Research on Technology in Education*, *53*(1), 107-123.

Thomas KF Chiu. 2025. Reform, challenges, and future research on AI for K-12 education. *Empowering K-12 Education with AI*. Taylor & Francis.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*: 208-231.

Richard Correnti, Mary Kay Stein, Margaret S. Smith, James Scherrer, Margaret McKeown, James Greeno, and Kevin Ashley. 2015. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In *Socializing Intelligence through Academic Talk and Dialogue*: 315-334.

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.

Lisa Dieker, Jacqueline A. Rodriguez, Benjamin Lignugaris/Kraft, Michael C. Hynes, and Charles E. Hughes. 2014. The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, 37: 21-33.

Zara Ersozlu, Susan Ledger, and Linda Hobbs. (2021). Virtual simulation in ITE: Technology driven authentic assessment and moderation of practice. In *Authentic Assessment and Evaluation Approaches and Practices in a Digital Era*: 53-68.

Benjamin Fauth, Jasmin Decristan, Anna-Theresia Decker, Gerhard Büttner, Ilonca Hardy, Eckhard Klieme, and Mareike Kunter. 2019. The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education*, 86: 102882.

Alastair Irons and Sam Elkington. 2021. Enhancing earning through formative assessment and feedback (2nd ed.). Routledge.

Dabae Lee and Sheunghyun Yeo. 2022. Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education*, 191: 104646.

Tammy Lee, Carrie Lee, Mark Newton, Paul

Vos, Jennifer Gallagher, Daniel Dickerson, and Camryn Regenthal. 2024. Peer to peer vs. virtual rehearsal simulation rehearsal contexts: Elementary teacher candidates' scientific discourse skills explored. *Journal of Science Teacher Education*, 35: 63-84.

Jieun Lim, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, Haewon Jung, Yoonsun Jang, Sanghyeok Lee, and Jewoong Moon. 2025. Development and implementation of a generative artificial intelligence-enhanced simulation to enhance problem-solving skills for pre-service teachers. *Computers & Education,* 232: 105306

Morva McDonald, M., Elham Kazemi, and Sarah Schneider Kavanagh. 2013. Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*: 378-386.

Sarah Michaels, Catherine O'Connor, and Lauren B. Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education,* 27: 283-297.

Jamie N. Mikeska, Beata Beigman Klebanov, Aakanksha Bhatia, Shreyashi Halder, Calli Shekell, Heather Jorgenson, Tricia Maxwell, and Benny Longwill. 2025. Using generative AI digital teaching simulations as practice spaces to support personalized and adaptive learning for preservice teachers in an elementary math methods course. [Manuscript submitted for publication.] Research Institute, ETS.

Jamie N. Mikeska and Aakanksha Bhatia. 2025. Using digital teaching simulations powered by generative artificial intelligence to propel teacher learning. *Journal of the Chartered College of Teaching*. Online.

Jamie N. Mikeska, Dionne Cross Francis, Pamela Lottero-Perdue, Meredith Park Rogers, Calli Shekell, Pavneet Kaur Bharaj, Heather Howell, Adam Maltese, Meredith Thompson, and Justin Reich. 2025. Promoting preservice teachers' facilitation of argumentation in mathematics and science through digital simulations. *Teaching and Teacher Education*, 15: 104858.

Jamie N. Mikeska, Heather Howell, Lisa Dieker, and Mike Hynes. 2021. Understanding the role of simulations in K-12 mathematics and science teacher education: Outcomes from a teacher education simulation conference. Contemporary *Issues in Technology and Teacher Education,* 21: 781-812.

Jamie N Mikeska, Heather Howell, and Devon Kinsey. 2023. Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education*, 74: 422–436.

Joseph Mintz, Wayne Holmes, Leping Liu, and Maria Perez-Ortiz. 2023. Artificial intelligence and K-12 education: Possibilities, pedagogies and risks. *Computers in the Schools,* 40: 325-333.

Tanya Nazaretsky, Jamie N. Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pp. 122-132. 2023.

John Pecore, Corey Nagle, Tadlee Welty, Minkyoung Kim, and Melissa Demetrikopoulos. 2023. Science teacher candidates' questioning and discussion skill performance in a virtual simulation using experiential deliberate practice. *Journal of Science Teacher Education,* 34: 415-435.

Margrit Schreier. 2012. Qualitative content analysis in practice.

Meghan Shaughnessy and Timothy A. Boerst. 2018. Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student's thinking. *Journal of Teacher Education,* 69: 40-55.

Laurie Sleep and Timothy A. Boerst. 2012. Preparing beginning teachers to elicit and interpret students' mathematical thinking. *Teaching and Teacher Education,* 28: 1038-1048.

Carrie Straub, Lisa Dieker, Michael Hynes, and Charlie Hughes. 2015. Using virtual rehearsal in TLE TeachLivE™ mixed reality classroom simulator to determine the effects on the performance of science teachers: A follow-up study. 2015 TeachLivE™ National Research Project: Year 2 Findings.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022a. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022b. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsey Claire Matsumura. 2024. Multi-dimensional performance analysis of Large Language Models for Classroom Discussion Assessment. *Journal of Educational Data Mining 16*(2), 304-335.

A Appendix

A.1 Example of one PST's transcript

Speaker	Utterance
Teacher	*Hi, Cecilia.*
Cecilia	*Hello!*
Teacher	*Can you walk me through the steps you did to solve the problem?*
Cecilia	*I drew circles for Mike's 6 crayons, then 2 circles for Ann's crayons. Then, I counted them: 6, 7, 8.*
Teacher	*Can you tell me why you solved the problem this way?*
Cecilia	*I used circles to show the crayons and counted to find out how many there are altogether.*
Teacher	*How did you know that that would work?*
Cecilia	*I just know it works to count them.*
Teacher	*Why did you decide to use that strategy?*
Cecilia	*Because it helps me see the crayons and count them.*
Teacher	*Can you explain why addition was used here, like needed here?*
Cecilia	*Adding helps me put together Mike's and Ann's crayons.*
Teacher	*So what do you think the problem is really asking?*
Cecilia	*The problem is asking how many crayons Mike and Ann have all together.*
Teacher	*What do you think would happen if you use a different strategy?*
Cecilia	*I don't know how to do that.*
Teacher	*Why did you use counting on here?*
Cecilia	*It makes counting faster. I start at the first number.*
Teacher	*Can you explain a little bit more?*
Cecilia	*I start with Mike's crayons, which are 6, and then count Ann's from there by saying 6, 7, 8.*
Teacher	*So why did you draw five circles with no dots and three with dots?*
Cecilia	*I didn't draw like that. I drew 6 circles and then 2 more. I put dots when I counted Ann's crayons after Mike's.*
Teacher	*It looks like in the picture that you drew, it looked like you did five plus three equals eight instead of drawing six plus two.*
Cecilia	*I drew 6 circles for Mike, then 2 for Ann. I just counted on from 6.*
Teacher	*Okay. Good job.*
Cecilia	*Thanks!*

Note. Minor edits have been made to address transcription errors.

Acknowledgement

This work is part of broader efforts in the Praxis program within ETS's K-16 Strategic Business Unit to ensure that GenAI teaching simulations provide productive learning opportunities for educators, including elementary paraeducators, preservice teachers, and in-service teachers. In future work, the Praxis program plans to incorporate a version of the GenAI teaching simulation used in this study into an online learning module designed to support elementary paraeducators in learning how to engage students with subject matter. We appreciate the PSTs who participated in the study and provided us with feedback on the simulation's usefulness.

Using LLMs to identify features of personal and professional skills in an open-response situational judgment test

Cole Walsh, Rodica Ivan, Muhammad Zafar Iqbal, and Colleen Robb
Acuity Insights Inc., Toronto, ON, Canada

Abstract

Academic programs are increasingly recognizing the importance of personal and professional skills and their critical role alongside technical expertise in preparing students for future success in diverse career paths. With this growing demand comes the need for scalable systems to measure, evaluate, and develop these skills. Situational Judgment Tests (SJTs) offer one potential avenue for measuring these skills in a standardized and reliable way, but open-response SJTs have traditionally relied on trained human raters for evaluation, presenting operational challenges to delivering SJTs at scale. Past attempts at developing NLP-based scoring systems for SJTs have fallen short due to issues with construct validity of these systems. In this article, we explore a novel approach to extracting construct-relevant features from SJT responses using large language models (LLMs). We use the Casper SJT to demonstrate the efficacy of this approach. This study sets the foundation for future developments in automated scoring for personal and professional skills.

1 Background

A longstanding challenge in academia is selecting qualified and professional candidates from a larger applicant pool for professional training programs. Decision makers have traditionally relied on measures of hard skills and cognitive ability to make these decisions (Eva et al., 2009), often relying on grade point average (GPA) and standardized tests such as the Scholastic Aptitude Test (SAT), Graduate Record Exam (GRE), Medical College Admission Test (MCAT), and Graduate Management Admission Test (GMAT). Personal and professional skills such as communication, teamwork, problem-solving, and critical thinking, although recognized as predictive of future success in education and industry (Heckman and Kautz, 2012), have been more difficult to measure for a number of reasons including lack of standardization and scalability (Patterson et al., 2016). Admissions committees have commonly used reference letters, personal essays, and interviews as a proxy for applicants' personal and professional skills, but these processes do not meet the psychometric standards that we would expect from tools used in high-stakes decision-making (Kuncel et al., 2014; Patterson et al., 2016). Additionally, as the adoption of generative AI spreads, there are increased concerns about the authenticity of reference letters and personal essays (Chen et al., 2024), further exacerbating the need for valid and reliable tools to measure personal and professional skills.

Recognizing the limitations of other admissions tools (i.e., reference letters, personal essays) (Patterson et al., 2016), higher education programs have been increasingly turning to a more reliable and standardized tool, Situational Judgment Tests (SJTs), to assess applicants' personal and professional skills as part of their admissions process (Webster et al., 2020; Nadmilail et al., 2023). Though they may be delivered in different formats, including fixed-response and open-response, SJTs generally involve simulated situations and questions designed to elicit how a respondent would likely react in the situation (Lievens, 2013). Fixed-response SJTs typically require respondents to select or rank possible actions based on their effectiveness in a given situation and show stronger relationships with measures of cognitive ability, rather than personal or professional skills (McDaniel et al., 2007). Open-response SJTs, on the other hand, are more conducive to measuring behavioral tendencies (i.e., how the respondent would likely react in the given situation) and tend to show stronger relationships with personal and professional skills relative to fixed-response SJTs (McDaniel et al., 2007).

Although open-response SJTs have proven effective in evaluating personal and professional skills in a standardized and reliable manner, there are

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 221–230

challenges in executing these types of assessments at scale. Open-response SJTs are primarily scored by human raters who require extensive training to become proficient at evaluating responses (Shipper et al., 2017). Additionally, performing this level of scoring at scale requires many trained human raters operating in parallel, which presents further operational barriers. These challenges are not unique to SJTs; developers of other open-response assessments have faced similar obstacles and overcome them with automated scoring systems such as Natural Language Processing (NLP) algorithms (Valenti et al., 2003). NLP-based scoring systems of this kind have been shown to achieve strong psychometric results in writing and language proficiency tests (Chodorow and Burstein, 2004; Ramineni et al., 2012; Cardwell et al., 2022), as well as short-answer tasks (Leacock and Chodorow, 2003).

While there is a growing literature on NLP-based scoring systems for open-response assessments, few studies have investigated their efficacy specifically for SJTs (Bulut et al., 2022; Walsh et al., 2022). One issue is that insights from other automated scoring systems may not be immediately transferable to SJTs given the difference in the measured construct: while other open-response assessments may focus on language proficiency or content-mastery, SJTs measure personal and professional skills (e.g., teamwork, problem solving, critical thinking) (Lievens and Motowidlo, 2016). These differences in the measured construct influence the kinds of features used as inputs to the scoring system. In particular, NLP-based scoring systems for writing and language proficiency typically use features related to coherence, grammar, and organization (Chodorow and Burstein, 2004; Ramineni et al., 2012; Cardwell et al., 2022), features which have no direct link with most constructs assessed by SJTs. Any valid NLP-based scoring system should exhibit construct relevance through the features used as inputs to said system (McCaffrey et al., 2022), making existing approaches to NLP-based scoring largely inapplicable to SJTs. Additionally, because open-response SJTs allow respondents to describe actions that they would take or have taken in the past, these assessments are designed to allow for complexity and response diversity, and thus there is generally no single correct answer (Dore et al., 2017). This characteristic of SJTs makes scoring responses based on "correctness" or similarity with other responses impractical as well.

2 Aims

In this study, we investigate the feasibility of identifying and extracting construct-relevant features from SJT responses. We build on the work of Iqbal *et al.*(Iqbal et al., 2025) who used a mixed-methods approach to identify nine construct-relevant features that influenced raters' evaluations of an open-response SJT. We probe whether and to what extent we can identify these features in SJT responses automatically using NLP-based approaches. Recognizing the complex and nuanced nature of these features, we decided to use Large Language Models (LLMs) for this task. Recent studies have demonstrated strong performance of LLMs for essay scoring (Lee et al., 2024) even in domains like divergent thinking (Organisciak et al., 2023), which, similar to SJTs, have been notoriously difficult to automatically score because of the complex nature of the construct. This work sets the stage for future endeavors to build an automated scoring system for open-response SJTs and similar assessment types.

3 Sample

We used data from the Casper SJT in this study. Casper is an open-response SJT that purports to measure respondents' personal and professional skills along the following competencies: collaboration, communication, empathy, ethics, fairness, motivation, problem solving, resilience, and self-awareness (Dore et al., 2017; Saxena et al., 2024). Casper presents respondents with a series of hypothetical scenarios that include either a text-based or video-based prompt. Text-based prompts include a short description of a situation while video-based prompts include trained actors enacting a scripted scenario. The respondents are then asked questions related to the prompt and given a fixed amount of time to respond. The data we used in this study included responses to both types of scenarios: video-based and word-based. An example of a situation depicted in a video-prompt scenario is given below:

> Chris and Jane are sitting together in a small meeting room. Their manager, Gary, enters to deliver a few brief comments before retreating to an adjoining work space. Chris gets up to approach Gary when he notices Gary focused on his phone instead of work. Jane tells Chris that she sees Gary on his phone very often and that overall he does not

do a lot of work. Chris says it does not seem fair for someone like Gary, who is senior to them in the company, to do less work and be paid a lot more. In addition, Gary takes their hard work to present as his own, taking credit for their efforts.

Respondents were instructed that they were a coworker of Chris and Jane in this scenario and asked the following three questions:

1. How would you handle this situation with Gary, your manager? Explain your response.
2. Imagine that Gary was completing his work in a timely manner outside of normal hours, but still behaving inappropriately while in the office. Would this change your opinion? Why or why not?
3. Describe some serious issues that can occur when supervisors are not present for their team.

In addition to the different types of scenario prompts, Casper also includes two distinct response formats: respondents are either required to type their responses to the associated questions within the allocated time or record a video of themselves responding to each question. To simplify this study, we only examined scenarios with the typed-response format as analyzing video responses would have required either transcribing the responses or passing the video media itself to a multimodal AI model, fundamentally altering the procedure employed here. We leave the analysis of video responses to a future study.

Casper is a completely human-rated assessment where a respondent's responses to a scenario are rated together holistically on a 1-9 Likert scale. Trained human raters are provided with scoring guidelines which help them contextualize Casper competencies for each scenario and determine how effectively the responses addressed the questions asked. Additionally, Casper is norm-referenced, which means that raters are also instructed to score each response relative to the other responses they are reviewing for the same scenario. Raters do not, however, use an analytical rubric when rating in order to encourage diverse perspectives and interpretation during the rating process. This rating approach allows for responses exhibiting different characteristics to still receive high scores as the context and reasoning provided by the respondent are also taken into account.

The diversity in scenarios and responses makes Casper ideally suited to study underlying features of personal and professional skills in SJT responses. Previously, Iqbal *et al.* identified nine construct-relevant features that influenced Casper ratings. For the purposes of the current study, we selected seven of these features to investigate the applicability of LLMs for feature extraction, omitting two features related to competencies associated with specific scenarios. Given that different Casper competencies are probed in each scenario, we omitted these two features from our investigation as their analysis would have required prompting with more scenario and competency-specific information, which would have extended the complexity of this pilot study.

Table 1 shows the seven features we selected for this study. Iqbal *et al.* previously analyzed 27 responses from each of three Casper typed-response scenarios, ensuring a uniform distribution of responses at each scoring level (i.e., three responses for each 1-9 score assigned by human raters). Two researchers independently classified the construct-relevant features present in all responses using the levels noted in Table 1.

For the present study, we doubled the size of the dataset used, re-using the dataset collected by Iqbal *et al.* while adding 27 responses from each of three additional scenarios, which were again classified by the same two human raters as in the original study. Thus, the complete dataset in this study comprised 162 responses across six distinct Casper scenarios. We report agreement for the researchers' classifications of the features across all 162 responses in the last column of Table 1 for each of the seven features. We used Cohen's κ with quadratic weighting (after mapping features to a numeric scale) to measure agreement. In the case of binary features, the quadratic-weighted κ is identical to an unweighted κ.

4 Methods

We used LLMs as classifiers to replicate the work of the human raters in identifying construct-relevant features in Casper responses. Previous studies of LLMs for essay scoring have identified performance gains when LLMs are allowed to specifically evaluate one aspect of writing at a time (Lee et al., 2024). We applied a similar principle here by prompting LLMs to evaluate only one feature at a time for a given response.

We conducted two separate analyses. In our first study we compared several LLMs based on how well their classifications aligned with those of human raters for each feature. We used five state-of-the-art LLMs, listed in Table 2, including a mix of reasoning and non-reasoning, open and closed-source models.

For each LLM and feature, we generated classifications for all 162 Casper responses, then computed the κ between model classifications and each human rater's classifications. We then averaged the results to obtain one average κ for each LLM and feature. We used the same zero-shot prompt for all LLMs; we did not provide the LLMs with any examples within the prompts. Further, we wanted to compare how well each LLM performed using the same prompt without tailoring the prompt to work better with one LLM or another, so we provided only the necessary information to carry out the task within the prompt. Below is a minimal reproducible example of the system prompt we used:

> *You are a helpful assistant that analyzes users' responses to an ethical dilemma.*
>
> *The user was given the following prompt: "{context}".*
> *They were asked to respond to the following questions related to this prompt: "{questions}".*
>
> *Your task is to analyze {feature_description}.*
> *Return your response as a JSON object with the following keys:*
> *"decision": <{feature_levels}>*
> *"reasoning": <Reasoning for decision>*

where curly braces denote variable inputs in the prompt. {context} is the text prompt (for text-based scenarios) or the summary of the video (for video-based scenarios), {questions} is an enumerated list of the questions asked, {feature_description} is a description of the feature being evaluated (see Table 1), and {feature_levels} are the possible values that the LLM may assign.

We required the LLMs to return a "reasoning" for their decisions. For example, using the scenario and questions described above with the PERSP feature, o4-mini as the LLM, and the following response:

1. I'd invite Gary to a friendly chat, share what Jane and I feel, and listen to his pressures. Together we'd create small habits—visible credit tags and quick check-ins—so everyone's effort shines and Gary can still balance his senior duties.

2. If Gary finishes work later, I'd appreciate his flexibility, yet gently note that daytime presence and fair credit nurture team spirit. Leadership isn't only outcomes; it's modeling engagement and celebrating others while they're present.

3. When supervisors vanish, teammates can feel adrift: goals drift, growth pauses, mistakes linger, and trust erodes. Consistent guidance, recognition, and availability are anchors that keep performance, learning, and well-being afloat.

we get the following output:

> *"decision":"Thoughtfully considers and empathizes with multiple perspectives", "reasoning":"The user's responses address the concerns of the co-workers (feeling undervalued and needing recognition), invite Gary to share his pressures and perspective, and propose collaborative solutions that respect his senior role. They also weigh the team's needs for presence, credit, and guidance, demonstrating balanced empathy for both the manager and the team."*

We do not inspect "reasoning" outputs in this study, though such components can provide important validity evidence for LLM-based scoring systems (Casabianca et al., 2025), hence why we include this element in our prompting strategy.

In our second study, we investigated whether and to what extent we could improve LLM-human agreement by providing further details, including inclusion and exclusion criteria, for each level of a feature. Previous studies have indicated that this prompting strategy can effectively improve LLM performance in essay evaluation tasks (Lee et al., 2024). We worked exclusively with o4-mini for this analysis because it offered the best combination of throughput, cost, and performance that was ideally suited for this iterative pilot study.

5 Results

5.1 Comparison of LLMs with zero-shot prompt

Results are shown in Table 3, while Figure 1 shows the average κ agreement between each LLM and the two human raters on each feature using the zero-shot prompt. We find that Claude Sonnet 4 generally outperforms the other models, achieving the highest agreement with human raters on four out of seven features, while achieving the second highest agreement on two other features (JUST and CREAT). LACKINF was the lone feature where Claude Sonnet 4 did not rank among the top two LLMs, but even in this case the model achieved near human-level agreement ($\kappa_{\text{Claude Sonnet 4}} = 0.566$ compared to $\kappa_{\text{Humans}} = 0.640$).

o4-mini provides similar results to Claude Sonnet 4 in most cases, but notably struggles to identify responses that "provide insightful, novel, or creative arguments to address the questions" (CREAT). DeepSeek-R1, on the other hand, excels at identifying responses that fit this definition. GPT-4o mini is generally outclassed by the other models, but does reach super-human agreement in identifying responses that "state that they do not have enough information to make a decision" (LACKINF). Among all features explored in this study, LACKINF is the most likely to be identifiable through the use of particular words or phrases. For example, the string "gather more information" appears in seven responses. One human rater marked four of these responses as exhibiting the LACKINF feature while the other human rater marked all seven responses as exhibiting this feature. GPT-4o mini, meanwhile, classified six out of the seven responses as exhibiting LACKINF. Features such as this one that may be identifiable through keyword or semantic relationships likely see smaller benefits from using LLMs and, especially, more advanced reasoning models.

For four out of seven features, the top performing LLM achieved $\kappa > 0.4$. However, none of the LLMs approached human-level agreement on any feature outside of LACKINF. The disparities between LLM-human agreement and human-level agreement ranged from 0.209 to 0.352 ($\kappa_{\text{Humans}} - \kappa_{\text{LLM}}$). This result is not surprising given the sparse instructions provided to the LLMs in the zero-shot prompt. In the second part of this study, we investigate whether and to what extent we can close this gap via prompt engineering.

5.2 Improving LLM-Human Agreement

Disagreement between LLMs and human raters generally stems from lack of alignment on thresholds separating the levels of a feature. Table 4 shows the proportion of classifications made by the two human raters and o4-mini for each level of each feature. We can see that o4-mini is typically misaligned with the human raters in terms of how to separate the levels of a feature. For example, while human raters label a response as "fail[ing] to acknowledge or show sensitivity towards the legitimate concerns or feelings of one of the parties involved" (DISRES) only $5 - 6\%$ of the time, o4-mini classified 22.2% of responses as such. Similarly, o4-mini classified 59.9% of responses as having "Reasonable Justification", while classifying 0% and 0.6% of responses as displaying "No Justification" and "Clear and Compelling Justification", respectively. Human raters, meanwhile, provided more classifications at these extreme ends of the ordinal scale at the expense of labels in the middle of the scale. This result reflects an overall pattern we see across all non-binary features: o4-mini tended to provide more classifications in the middle of an ordinal scale than we observed with human raters. We used these results to motivate our prompt engineering strategy and further delineate feature levels.

We focused on six features for prompt engineering, omitting LACKINF where o4-mini was already achieving close to human-level performance. Results are displayed in the last column of Table 3 as well as Figure 2. We find that including additional details about the levels for a feature in the prompt effectively improves the LLM's agreement with humans. For most features we saw improvements of $0.08 < \Delta\kappa < 0.1$, but for DISRES we saw gains of $\Delta\kappa = 0.206$. With these prompts, o4-mini would've performed higher than all LLMs tested in the first experiment on all features except LACKINF (which we did not investigate improving) and CREAT, where o4-mini performs better, but is still outclassed by most other models.

6 Conclusions and Future Work

This study evaluated the feasibility of using LLMs to extract construct-relevant features from the Casper SJT. We found that reasoning models like OpenAI's o4-mini and Anthropic's Claude Sonnet 4 generally performed best at identifying these complex and nuanced constructs in responses, even

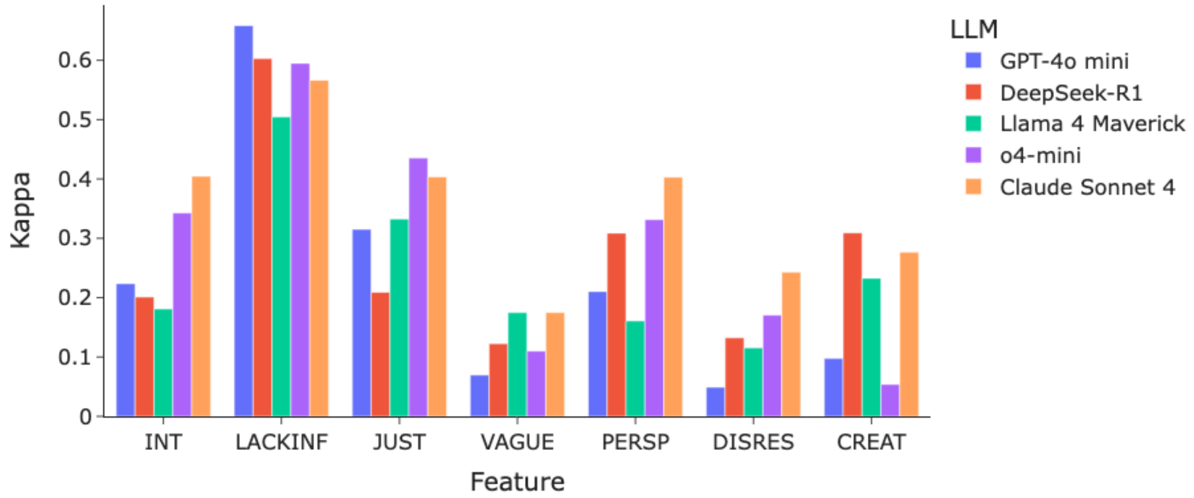


Figure 1: Average κ with human raters using zero-shot prompt.

with limited instructions. Additionally, we found that each LLM that we tested achieved the strongest performance on at least one feature. This result indicates that a future automated scoring solution using the same feature extraction method may be best served by using different LLMs for different features rather than forcing a single universal LLM. We could also consider using multiple LLMs for the same criteria and instituting a voting system resembling traditional machine learning ensemble methods to produce more accurate and reliable results (Dietterich, 2000). Overall, our results suggest a promising avenue for extracting construct-relevant features from SJTs and similar open-response assessments.

Prior to engaging in prompt engineering to improve performance, we had already reached close to human-level agreement in extracting one feature, whether a response "state[d] that they do not have enough information to make a decision." In this particular case, we hypothesize that features such as this one may be extractable by simpler models and methods such as keyword and semantic search.

For other features we fell well short of human-level agreement using zero-shot prompt classification, but we demonstrated that LLMs can be instructed to behave closer to expectations by giving further details about the levels for a feature. We found that providing these details had varying effects on our classification performance for different features, indicating that different features may be more susceptible to influence from this type of prompt engineering. Future work could explore other approaches to prompt engineering including few-shot prompting as well as fine-tuning to further improve performance.

We were also limited by small sample sizes in this study, owing to the effort and expertise required to annotate datasets such as these based on the features we explored. Future work will extend this study to explore a larger dataset from the Casper SJT as well as additional features. We plan to investigate the use of these features in an eventual automated scoring system for the Casper SJT. Such work would have important consequences, potentially extending the scalability and standardization of open-response assessments of personal and professional skills.

An automated scoring system based on the approach demonstrated here would also open avenues for future work in formative assessments by providing real time evaluation and feedback to respondents. We used a system prompt in this study that returned both a "decision" and "reasoning". We did not inspect the "reasoning"s in this study, but future work could use these "reasoning" fields to generate personalized and direct feedback for respondents. This method of extracting features from text could also be extended beyond assessments to other pieces of written text such as personal essays and reference letters.

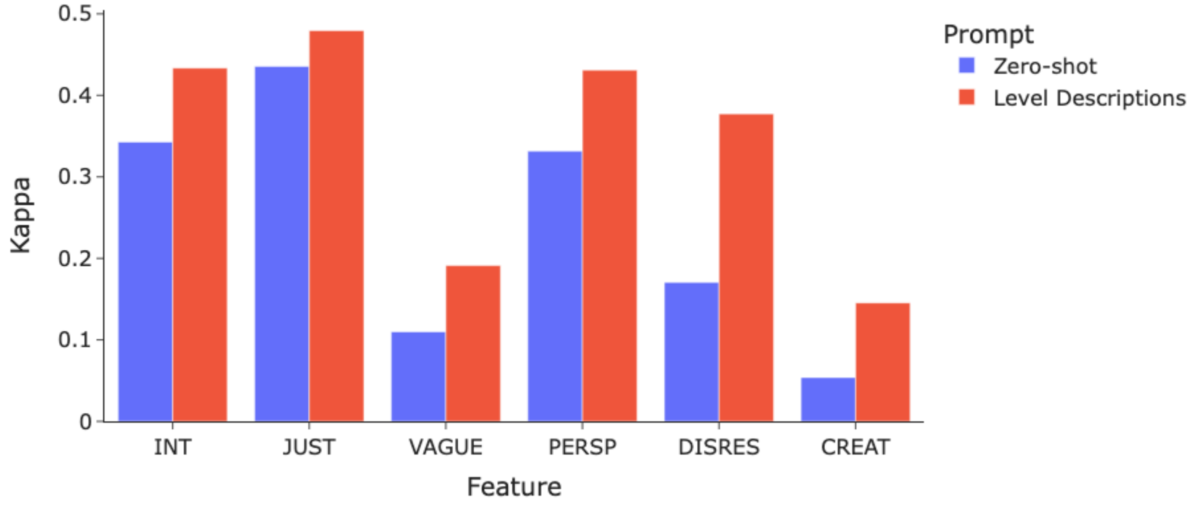


Figure 2: Average κ with human raters using o4-mini with zero-shot prompting and prompting with additional details for each level of a feature. Human-LLM agreement improves when providing additional level details in the prompt.

Acknowledgments

We would like to acknowledge Jillian Derby, Alexander MacIntosh, Josh Moskowitz, Gill Sitarenios, Susha Suresh, and Yiyu Xie for thoughtful discussions on the ideas presented in this manuscript.

References

Okan Bulut, Alexander MacIntosh, and Cole Walsh. 2022. Leveraging natural language processing for quality assurance of a situational judgement test. In *International Conference on Artificial Intelligence in Education*, pages 84–88. Springer.

Ramsey Cardwell, Geoffrey T LaFlair, and Burr Settles. 2022. Duolingo english test: technical manual. *Duolingo Research Report*.

Jodi M Casabianca, Daniel F McCaffrey, Matthew S Johnson, Naim Alper, and Vladimir Zubenko. 2025. Validity arguments for constructed response scoring using generative artificial intelligence applications. *arXiv preprint arXiv:2501.02334*.

Jeffrey Chen, Brendan K Tao, Shihyun Park, and Esta Bovill. 2024. Can chatgpt fool the match? artificial intelligence personal statements for plastic surgery residency applications: a comparative study. *Plastic Surgery*, page 22925503241264832.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Kelly L Dore, Harold I Reiter, Sharyn Kreuger, and Geoffrey R Norman. 2017. Casper, an online preinterview screen for personal/professional characteristics: prediction of national licensure scores. *Advances in Health Sciences Education*, 22:327–336.

Kevin W Eva, Harold I Reiter, Kien Trinh, Parveen Wasi, Jack Rosenfeld, and Geoffrey R Norman. 2009. Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical education*, 43(8):767–775.

James J Heckman and Tim Kautz. 2012. Hard evidence on soft skills. *Labour economics*, 19(4):451–464.

Muhammad Zafar Iqbal, Rodica Ivan, Colleen Robb, and Jillian Derby. 2025. Evaluating factors that impact scoring an open response situational judgment test: a mixed methods approach. *Frontiers in Medicine*, 11:1525156.

Nathan R Kuncel, Rachael J Kochevar, and Deniz S Ones. 2014. A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment*, 22(1):101–107.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Filip Lievens. 2013. Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. *Medical education*, 47(2):182–189.

Filip Lievens and Stephan J Motowidlo. 2016. Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9(1):3–22.

Daniel F McCaffrey, Jodi M Casabianca, Kathryn L Ricker-Pedley, René R Lawless, and Cathy Wendler. 2022. Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1):1–58.

Michael A McDaniel, Nathan S Hartman, Deborah L Whetzel, and W LEE GRUBB III. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60(1):63–91.

Azad Iqram Nadmilail, Mohd Effendi Ewan Mohd Matore, Siti Mistima Maat, and Lynn Sheridan. 2023. Broad vs. narrow traits: a scoping review of measuring personality traits in teacher selection using the situational judgment test. *Frontiers in Psychology*, 14:1217321.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Fiona Patterson, Alec Knight, Jon Dowell, Sandra Nicholson, Fran Cousans, and Jennifer Cleland. 2016. How effective are selection methods in medical education? a systematic review. *Medical education*, 50(1):36–60.

Chaitanya Ramineni, Catherine S Trapani, David M Williamson, Tim Davey, and Brent Bridgeman. 2012. Evaluation of the e-rater® scoring engine for the gre® issue and argument prompts. *ETS Research Report Series*, 2012(1):i–106.

Anurag Saxena, Loni Desanghere, Kelly Dore, and Harold Reiter. 2024. Incorporating a situational judgement test in residency selections: clinical, educational and organizational outcomes. *BMC Medical Education*, 24(1):339.

Edward S Shipper, Laura M Mazer, Sylvia Bereknyei Merrell, Dana T Lin, James N Lau, and Marc L Melcher. 2017. Pilot evaluation of the computer-based assessment for sampling personal characteristics test. *Journal of Surgical Research*, 215:211–218.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Cole Walsh, Alexander MacIntosh, Okan Bulut, and Jinnie Shin. 2022. What are we measuring?: A topic modeling framework to map professionalism aspects to responses in a situational judgment test. In *Companion Proceedings of the 12th International Conference on Learning Analytics Knowledge LAK22*, pages 39–41.

Elin S Webster, Lewis W Paton, Paul ES Crampton, and Paul A Tiffin. 2020. Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10):888–902.

Key	Description	Levels	Cohen's κ (Humans)
INT	Grasps and addresses the complex social and emotional dynamics present in the ethical dilemma.	• Limited Interpretation • Adequate Interpretation • Excellent Interpretation	0.700
LACKINF	States that they do not have enough information to make a decision.	True/False	0.640
JUST	Justifies the course of action suggested.	• No Justification • Superficial Justification • Reasonable Justification • Clear and Compelling Justification	0.788
VAGUE	Vague or unclear.	True/False	0.356
PERSP	Considers the perspectives of the different parties involved in the dilemma.	• Considers one perspective • Briefly considers multiple perspectives • Thoughtfully considers multiple perspectives	0.722
DISRES	Fails to acknowledge or show sensitivity towards the legitimate concerns or feelings of one of the parties involved.	True/False	0.647
CREAT	Provides insightful, novel, or creative arguments to address the question.	True/False	0.510

Table 1: Features identified by Iqbal *et al.* as influencing Casper scores and used in the present study. Cohen's κ is reported between two independent human raters' classifications of these features across 162 Casper responses.

Table 2: LLMs explored in this study including their providers, whether they were reasoning models, and whether model weights were open or closed-source.

Name	Provider	Reasoning Model (Y/N)	Open/closed-source
GPT-4o-mini	OpenAI	N	Closed
DeepSeek-R1	DeepSeek	Y	Open
Lllama 4 Maverick	Meta	N	Open
o4-mini	OpenAI	Y	Closed
Claude Sonnet 4	Anthropic	Y	Closed

Feature	Zero-shot GPT-4o mini	DeepSeek-R1	Llama 4 Mav.	o4-mini	Sonnet 4	Level Desc. o4-mini
INT	0.224	0.201	0.181	0.343	**0.404**	0.434
LACKINF	**0.658**	0.603	0.505	0.595	0.566	-
JUST	0.315	0.209	0.333	**0.436**	0.404	0.479
VAGUE	0.070	0.123	**0.175**	0.110	**0.175**	0.191
PERSP	0.210	0.309	0.161	0.332	**0.403**	0.431
DISRES	0.049	0.132	0.116	0.171	**0.243**	0.377
CREAT	0.098	**0.309**	0.233	0.054	0.277	0.145

Table 3: Average Cohen's κ agreement with human raters for each LLM on each feature using the zero-shot prompt. The last column shows the average κ for o4-mini after modifying the prompts to include level descriptions for each feature. We did not explore the LACKINF feature in this second experiment because we achieved close to human-level agreement with the zero-shot prompt.

Key	Level	Proportion Selected Human 1	Human 2	o4-mini
INT	Limited Interpretation	0.435	0.377	0.327
	Adequate Interpretation	0.447	0.475	0.642
	Excellent Interpretation	0.118	0.148	0.031
LACKINF	False	0.944	0.889	0.864
	True	0.056	0.111	0.136
JUST	No Justification	0.062	0.056	0
	Superficial Justification	0.358	0.333	0.395
	Reasonable Justification	0.358	0.469	0.599
	Clear and Compelling Justification	0.222	0.142	0.006
VAGUE	False	0.790	0.883	0.568
	True	0.210	0.117	0.432
PERSP	Considers one perspective	0.302	0.407	0.149
	Briefly considers multiple perspectives	0.407	0.549	0.758
	Thoughtfully considers multiple perspectives	0.290	0.272	0.093
DISRES	False	0.938	0.951	0.778
	True	0.062	0.049	0.222
CREAT	False	0.833	0.796	0.994
	True	0.167	0.204	0.006

Table 4: Proportion of responses where each feature level was selected by human raters and o4-mini (with zero-shot prompt).

Automated Evaluation of Standardized Patients with LLMs

Andrew Emerson[1], Le An Ha[2], Keelan Evanini[1], Su Somay[1], Kevin Frome[1], Polina Harik[1], Victoria Yaneva[1]
[1]National Board of Medical Examiners, Philadelphia, USA
{aemerson, kevanini, ssomay, kfrome, pharik, vyaneva}@nbme.org
[2]Ho Chi Minh City University of Foreign Languages, Vietnam
anhl@huflit.edu.vn

Abstract

Standardized patients (SPs) are essential for clinical reasoning assessments in medical education. This paper introduces evaluation metrics that apply to both human and simulated SP systems. The metrics are computed using two LLM-as-a-judge approaches that align with human evaluators on SP performance, enabling scalable formative clinical reasoning assessments.

1 Introduction

Clinical reasoning (CR) skills are fundamental to accurate diagnosis and effective patient care; accordingly, their systematic instruction and assessment constitute a critical component of undergraduate medical education (Harden, 1988). One of the most widely adopted formats for evaluating clinical reasoning competencies is the Objective Structured Clinical Examination (OSCE). Repeated, structured interactions within OSCEs have been shown to effectively promote the development of clinical reasoning skills (Laschinger et al., 2008). A central feature of OSCEs is the simulated clinical encounter, in which learners engage in a clinical case by interviewing a patient to elicit diagnostically relevant information, including symptoms, medical history, and context. Traditionally, OSCEs employ laypersons trained to portray clinical scenarios—referred to as standardized patients (SPs)—who are tasked with consistently enacting specific patient personas to support teaching, learning, and assessment. SPs adhere to detailed case scripts and standardized response protocols to ensure realism, reliability, and reproducibility across encounters.

While human SPs provide a realistic and safe environment for learners to practice clinical skills, their use is resource-intensive, requiring substantial investments in training, coordination, and examination delivery (Rau et al., 2011). In recent years, large language models (LLMs) have been explored as a way to design simulated standardized patients (SSPs), offering a promising alternative to traditional human SPs. SSPs offer several advantages, including scalability to larger cohorts of learners, increased availability for on-demand practice, and enhanced flexibility in portraying a wide range of patient personas. However, maintaining fidelity to the prescribed patient script and ensuring consistent persona representation remain significant challenges in the deployment of SSPs.

Large language models (LLMs) can be guided to portray specific patient characteristics (e.g., via prompting); however, the reliability and precision of such portrayals remain active areas of research (Cook, 2024; Schmidgall et al., 2024; Shindo and Uto, 2024). Fortunately, existing frameworks for evaluating the performance of SPs can be adapted and automated for use with SSPs (Geathers et al., 2025). The evaluation of SP or SSP performance generally falls into three categories: (1) human evaluation of the responses to physician questions; (2) traditional machine learning methods that are trained on labeled datasets of patient responses; and (3) LLMs that judge the quality of the patient responses with little or no prior training. To ensure the accuracy of methods (2) and (3), human evaluations (Category 1) are typically employed as a reference standard to produce ground-truth labels.

In this paper, we introduce a novel LLM-based evaluation framework to automatically evaluate SPs, applicable to both humans and SSPs. The framework classifies SP responses into one of five performance categories, developed as part of this work: *Correct, Too Much Information, Too Little Information, Incorrect,* or *Not Applicable*. These categories can be used as metrics to monitor the SP or SSP performance over time, across different SPs or SSP systems, and ensure that students are able to engage appropriately with the SP or SSP.

The contributions of this work are as follows:

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 231–238
October 27-29, 2025

1. We perform extensive human annotation of a set of 41 transcripts of student-patient interactions from four clinical cases to serve as the ground truth to validate our proposed automated evaluation approaches.
2. We introduce two methods of classifying SP responses:
   In *Method 1*, the LLM uses the case guidelines, the conversation up to this point, and the current physician question to classify the current response into the appropriate category. In *Method 2*, the LLM first uses the case guidelines, the conversation up to this point, and the current physician question to generate the prescribed SP response. The LLM then uses the prescribed SP response, current physician question, and conversation up to this point to compare with the actual SP response and classify it into the appropriate category.
3. We validate these proposed methods of SP evaluation by comparing the classification results of each method with the human evaluation, assessing alignment with human expert judgment. We discuss the implications of these results for evaluating both human and simulated SPs.

2 Related Work

2.1 Evaluation of Human SPs

SPs are typically evaluated on dimensions such as realism, accuracy, consistency, and communication. Realism and communication are commonly assessed through structured observations by faculty and peers, and student feedback immediately after encounters (Gonullu et al., 2023; Erby et al., 2011). Accuracy of performance—the physical, emotional, and cognitive portrayal of the clinical case—is often evaluated using third-party observations and SP self-assessment checklists. Post-encounter self-checklists help SPs reflect on the fidelity of their performance, improving reliability over time (Erby et al., 2011). Consistency of performance, a defining characteristic of SP programs, refers to the uniform delivery of case prompts and behaviors across all student interactions for a given case, and is evaluated through a mix of live observation, video review, and checklists (Lewis et al., 2017; Erby et al., 2011). Overall, the evaluation of human SPs' performance remains mainly manual, involving faculty and peer observation, student feedback, and SP self-assessment, offering a multi-angled view essential for maintaining high standards in simulated clinical environments.

2.2 LLMs as SSPs and their Evaluation

Recent research has explored the potential of LLMs in simulating patient interactions (LLMs as SSPs) to support both clinical skill development and performance evaluation. For instance, Li et al. (2024) examined how SSPs can support clinical inquiry skills, while Holderried et al. (2024) focused on improving medical history-taking skills. Yamamoto et al. (2024) expanded this to encompass general medical interview skills, and Sardesai et al. (2024) applied LLM-based simulations to anesthesia training. Gray et al. (2024) investigated the use of LLMs in guiding prenatal counseling, whereas Tu et al. (2024) worked on advancing AI diagnostic agents to improve their clinical utility.

Human judgment remains the most widely used reference for assessing chatbot-generated interactions. Specialists (Chen et al., 2023; Gray et al., 2024) and students (Fan et al., 2024) have been engaged to evaluate the realism, appropriateness, usability, relevance, rationality, and honesty of chatbot outputs. User surveys—such as Likert-scale questionnaires (Sardesai et al., 2024), the Chatbot Usability Questionnaire (Holderried et al., 2024), and the Simulation-Based Training Quality Assurance Tool (Yamamoto et al., 2024)—have been leveraged to evaluate perceived usability, intuitiveness, accuracy, comfort, and overall user experience. Automated metrics, like algorithmically derived conversational dimensions (Liao et al., 2024), a GPT-4–based chatbot arena framework (Li et al., 2024), and quantitative scoring of chatbot responses (Chen et al., 2023), have been used to enable scalable and objective evaluation of chatbot performance. These metrics assess factors such as accuracy, honesty, focus, passivity, cautiousness, and guidance. Finally, outcome-based evaluations have also been conducted; for example, Yamamoto et al. (2024) compared formal exam performance between students who used SSPs during their preparation and those who did not.

The above studies have identified several limitations of LLMs, including their tendency to produce hallucinated content, overly formal or repetitive responses, and unnaturally polite dialogue (Sardesai et al., 2024). Gray et al. (2024) highlight the importance of expert oversight when using AI-generated content in educational settings. Additionally, current LLM-based systems provide limited support for nonverbal communication skills, which are essential for effective medical interviewing (Yamamoto et al., 2024).

3 Data

3.1 Clinical Interviews with Standardized Patients

The data were drawn from a prior study in which participating students interacted with four human SPs, each portraying a distinct case. Each scenario was developed along with case-specific guidelines and training protocols designed to elicit observable clinical reasoning behaviors from the students. Students were randomly assigned to begin with one of the four cases and subsequently completed the remaining three cases in order. All encounters were recorded using Recollective[1], a qualitative research platform that supports live and asynchronous (i.e., pre-recorded) video interactions. The software first recorded the conversations and then produced transcriptions in separate files, differentiating the student and SP speech.

Participating Students. A total of 76 post-clerkship medical students were recruited from four U.S. medical schools. 32 were in their third year of medical school, 2 were transitional students between their third and fourth years, and 42 were in their fourth year.

Standardized Patients. Standardized patients were recruited from local training programs and partner institutions affiliated with the study sites. Each SP received standard compensation for participation in both training and assessment activities. The medical assessment organization personnel conducted the SP training following established industry protocols, general guidelines, and case-specific requirements. SPs underwent both individual and group training sessions to ensure consistency and reliability across performances. For each clinical case, a minimum of three SPs were trained to serve not only as actors in student encounters but also as peer evaluators, providing feedback and quality assurance for fellow SPs.

Clinical Cases. Four clinical cases were developed as part of a prior study to support the standardized evaluation of medical students' clinical reasoning skills. For each case, both general training protocols and case-specific instructions were designed to guide SP behavior and ensure that student–SP interactions elicited diagnostically productive lines of questioning. SPs were explicitly instructed to refrain from offering suggestions or guidance on how students should conduct the encounter. In instances where students inquired about symptoms not included in the case script, SPs were instructed to deny the presence of such symptoms to preserve case fidelity. Each encounter began with a standardized opening statement delivered by the SP, introducing the primary reason for visiting the clinic. To maintain consistent interactions, boilerplate responses were developed for addressing routine or general questions. For open-ended inquiries, SPs were provided with a sequenced set of acceptable responses, structured to disclose relevant clinical information with the goal of not revealing too much information at once. This approach was intended to support the development of student inquiry skills while preserving the realism and educational value of the simulation. Case contents were designed to be both realistic and engaging for the student. Case 1 consists of a 33-year-old woman who has been experiencing shortness of breath; Case 2 consists of a 40-year-old man who has been continuously vomiting; Case 3 consists of a 46-year-old woman who has been experiencing weakness; and Case 4 consists of a 65-year-old man who has had trouble sleeping.

3.2 Patient Response Annotation

Two human experts who were familiar with the case contents and evaluation criteria annotated transcripts of the conversations in order to evaluate the quality of the SP responses. Each response from the human SPs was labeled with one of the five discrete labels in Table 1. The label categories were derived based on a combination of insights from literature that evaluates human SPs and practical guidance by members of the team who have trained human SPs. The annotation process consisted of several steps to increase agreement between annotators and to ensure high-quality annotations. First, each annotator independently annotated the SP responses in an adjudication set of four transcripts sampled from the same case. Subsequently, the annotators reviewed any discrepant annotations together with other team members and agreed upon adjudicated annotations. Revisions to the annotation guidelines were made accordingly based on these conversations. The final calibration set included 162 *Correct*, 44 *Not Applicable*, 13 *Too Much Information*,

[1]https://www.recollective.com/qualitative-research-recollective

Label	Description
Correct	The response is accurate and appropriate given the instructions contained in the case training guidelines. The response is relevant to the physician's question and contains the appropriate amount of content based on the specific question the physician asked.
Too Much Information	The response is relevant to the physician's question, but it contains more information than is justified based on the specific question that the physician asked. This can occur when the patient provides additional information from the case materials that wasn't prompted by a question from the physician.
Too Little Information	The response is relevant to the physician's question, but it contains less information than is expected based on the specific question that the physician asked. This can occur when the patient omits relevant content from the case materials and provides a generic answer.
Incorrect	The response is not accurate or is inappropriate given the instructions contained in the case training guidelines. This can occur when the patient provides a response that is irrelevant or off-topic, when the patient volunteers made-up information about topics that are not covered in the training guidelines, when the patient provides specific details about their condition that are not specified in the case materials, etc.
Not Applicable	The question is not applicable to the case document and results in a non-clinical or irrelevant response.

Table 1: Annotation guidelines given to annotators for evaluating each SP response.

7 *Too Little Information*, and 7 *Incorrect* responses. After the adjudication round, the two annotators independently annotated the same 20 transcripts (five randomly sampled from each case). Finally, 17 additional transcripts were single-annotated by the annotators. Transcripts of entire conversations were annotated to allow for contextual information to be available to annotators. Table 2 shows the annotation distribution for each annotator on transcripts that were not in the calibration set. The calibration, double-annotation, and single-annotation sets yielded 41 transcripts with an average of 54.8 annotated question-response pairs (*SD*=16.4). This produced 2248 question-response pairs in total. See Appendix A for exemplar labeled patient responses.

Label	Annotator 1 Count	Annotator 2 Count
Correct	1407 (72%)	748 (65%)
TMI	59 (3%)	14 (1%)
TLI	40 (2%)	2 (0%)
Incorrect	38 (2%)	29 (3%)
NA	414 (21%)	365 (32%)

Table 2: Distribution of annotations by annotator.

The annotation guidelines were developed based on industry practice, SP training guidelines, and conversational agent literature. The Cohen's Kappa value denoting the inter-annotator agreement on the double-annotated set of 20 transcripts was 0.501, and the agreement percentage was 76.5% (*n*=1104). For case 1, the agreement percentage was 70.8% ($\kappa = 0.436$, *n*=281). For case 2, the agreement percentage was 80.2% ($\kappa = 0.481$, *n*=217). For case 3, the agreement percentage was 82.8% ($\kappa = 0.604$, *n*=332). For case 4, the agreement percentage was 71.9% ($\kappa = 0.457$, *n*=274).

4 LLM-as-a-Judge Evaluation

To automatically evaluate SP responses to student questions, we employed a technique that leverages LLMs called LLM-as-a-judge (Gu et al., 2025). For all evaluations, we used OpenAI's GPT-4o (version: *2025-01-01-preview*) as the judge. In this paper, we introduce two methods of using LLMs to judge the SP responses. *Method 1* uses a single request to the LLM to categorize the SP response using the case-specific guidelines, the conversation up to this point, and the current physician question. *Method 2* uses two requests to the LLM, in which the first request generates a prescribed patient response based on the case-specific guidelines, the conversation up to this point, the current physician question, and the second request compares the prescribed and actual SP response to categorize the SP response. *Method 2* was chosen over alternative methods that leverage prior data (e.g., few-shot learning or fine-tuning) to first attempt to solve this problem without the use of labeled examples, which would require a robust set of ground truth labels.

For both methods, the case guidelines are the same instructions that are given to the SPs to portray the patient. The conversations are encoded as transcribed text and each question-response pair is appended up to the current question as context, noting the speaker of the text (i.e., student or SP).

5 Results

To evaluate the performance of the LLM-as-a-judge method relative to human annotations, we used a dataset comprising 2248 human-annotated question-response pairs drawn from 41 encounter transcripts. Table 3 displays the results for this comparison. Both F1 scores and accuracy metrics are reported to assess the degree of alignment between LLM-generated classifications and human reference annotations in the evaluated methods. Results are reported both in aggregate and disaggregated by individual cases: Cases 1 (n=814), 2 (n=382), 3 (n=571), and 4 (n=481). Across all cases and in the overall analysis, *Method 1* consistently outperforms *Method 2*. A baseline comparison, referred to as the *Majority* baseline, assigns the most frequent class label (which is always *Correct*) to all instances within each case and in the overall dataset. *Method 1* outperforms this baseline in terms of F1 scores, but its performance in terms of accuracy shows mixed results. Given the multi-class nature of the problem and the imbalanced label distributions, F1 score is a more informative metric.

Table 4 displays the results per label, including the distribution of predicted labels, their precision, recall, and F1 score for the entire human-annotated dataset.

6 Discussion

Overall, the findings of this study show positive results for both the human annotation process and the two proposed automated LLM-as-a-judge methods. Human annotation remains a very resource-intensive and cognitively demanding task that requires careful calibration and deliberation among annotators to ensure consistency and validity. LLMs offer a scalable and efficient alternative that can be used in conjuction with human annotations to reduce manual labor involved in annotating. One practical application of this hybrid approach is the selective delegation of annotations to LLMs for labels where model performance is demonstrably high (e.g., the *Correct* label). By pre-filtering such responses, human annotators can allocate their attention to more complex or ambiguous categories that require more nuanced judgment such as *Too Much Information* or *Too Little Information*. In addition, certain LLM-as-a-judge methods were shown to be effective in annotating responses that are *Not Applicable* (i.e., responses associated with non-clinical questions), providing another opportunity for filtering and streamlining the annotation process. Taken together, these findings suggest that LLM-as-a-judge approaches can serve as valuable tools for augmenting human annotation workflows, saving time and effort for human reviewers while preserving annotation quality.

Although human annotations served as the ground truth (reference standard) for this study, notable levels of disagreement were observed among annotators. For the subset that were doubly-annotated (n=1104), the Kappa value for inter-annotator agreement was 0.501. While this reflects only moderate agreement, the task of labeling responses with one of five possible subjective categories can lead to poor agreement. Ironically, the lowest annotator agreement was observed in Case 1, which was also the case for which the initial annotator calibration was performed. This finding underscores the influence of *case-specific* features on both human annotation and LLM labeling. The variability observed suggests the need to standardize the annotation guideline development process to promote consistency across clinical scenarios. Many of the disagreements among annotators were along the threshold of *Correct* and partially correct responses (e.g., *Too Much Information*). These are often very nuanced phrases and require carefully crafted definitions during case generation.

Across all individual cases and in the aggregate analysis, *Method 1* consistently outperformed both the *Majority* baseline and *Method 2*. With the advancement of LLMs and continual refinement of prompting techniques, it is not surprising that an LLM-as-a-judge method can outperform a majority class baseline. What is surprising and noteworthy is the difference in performance between the two LLM-based methods. *Method 1* incorporates the case-specific guidelines directly into the LLM request that conducts the evaluation of each SP response, while *Method 2* references the guidelines only during the initial generation of LLM-recommended patient responses and not during the LLM request that conducts the evaluation. As a result, *Method 2* loses information when making the actual evaluation and classification of the SP response, only comparing it to another response. Our team had hypothesized that this would improve performance by allowing the LLM to split the evaluation task into multiple steps. This result underscores the power of current LLMs in navigating large context windows (e.g., long conversations and long reference documents, simultaneously). It may also suggest that the LLM-generated responses may not be informative enough to serve as a ground-truth response. Despite its superior overall performance, *Method 1* exhibited uneven classification accuracy across labels, with the majority of predicted labels falling under the *Correct* category. Among the minority classes, only *Not*

Method	Overall F1 (Accuracy)	Case 1 F1 (Accuracy)	Case 2 F1 (Accuracy)	Case 3 F1 (Accuracy)	Case 4 F1 (Accuracy)
Majority	0.60 (0.72)	0.56 (0.69)	0.75 (0.83)	0.57 (0.70)	0.59 (0.71)
1	0.72 (0.67)	0.71 (0.70)	0.79 (0.75)	0.69 (0.63)	0.69 (0.64)
2	0.48 (0.42)	0.47 (0.44)	0.59 (0.49)	0.45 (0.38)	0.44 (0.38)

Table 3: LLM-as-a-Judge performance.

Method	Label	Predicted Count	Precision	Recall	F1
1	*Correct*	1396	0.86	0.74	0.80
1	*TMI*	278	0.09	0.33	0.14
1	*TLI*	93	0.02	0.4	0.03
1	*Incorrect*	92	0.09	0.17	0.12
1	*NA*	389	0.74	0.61	0.67
2	*Correct*	1037	0.83	0.53	0.65
2	*TMI*	555	0.11	0.81	0.19
2	*TLI*	501	0.06	0.60	0.10
2	*Incorrect*	151	0.03	0.09	0.04
2	*NA*	4	1.0	0.01	0.02

Table 4: LLM-as-a-Judge performance by label. *TMI=Too Much Information*, *TLI=Too Little Information*, and *NA=Not Applicable*.

Applicable achieved comparable performance. By contrast, *Method 2* demonstrated greater sensitivity to minority classes (i.e., *Too Much Information* and *Too Little Information*), suggesting a stricter approach and potentially greater attention to subtle linguistic nuances. These findings indicate that a hybrid approach, combining the contextual breadth of *Method 1* with the sensitivity of *Method 2*, may yield further improvements in performance.

This study focused specifically on annotation of SP responses to student-initiated questions. The ultimate goal of this work is to develop an automated system for evaluating SSP responses, with the dual goals of monitoring and improving system performance. Importantly, the proposed evaluation framework is equally applicable to human SP responses. Additionally, SP responses offer more natural conversation, more variability in phrasing, and more room for the student to ask variable questions compared to SSP responses. As a result, building an evaluation system by first using SP responses enables a finer-grained view of the types of responses that constitute a realistic patient encounter. This approach has the potential to enhance the formative utility of SSP systems, ultimately supporting more effective development of clinical reasoning skills in students. Future evaluations of SSP responses may find differences in response characteristics compared to SP responses, potentially leading to the adjustment of this evaluation framework.

Limitations

This study represents an initial investigation of both human annotation and automated LLM annotation of SP responses in physician-patient interactions. Due to resource constraints, the annotation process was limited to two annotators. With more annotators, an even larger pool of labeled responses could be annotated. Independent of the number of responses, ensuring high-quality annotations with high agreement is critical. This is especially important for labels that are actionable, such as when the SP provides too much or too little information in response to the student. A larger pool of annotators, with a more rigorous calibration and preparatory period would likely yield improved results.

7 Conclusion

The automatic annotation of SP responses has significant potential for advancing the development of more accurate and effective formative assessments of clinical reasoning. Enhancing the performance of either SPs or SSPs can contribute to more meaningful student-patient interactions. Building such an evaluation system requires high-quality human annotations to serve as the ground truth for what constitutes an effective (e.g., accurate) patient response. This study reported the results of a human annotation effort involving SP responses, guided by a structured rubric comprising five response categories. Building on this foundation, two LLM-as-a-judge methods were introduced as automated approaches to replicate the human annotation process. Both methods showed promising agreement with human judgments. Future research should focus on integrating the strengths of each LLM-as-a-judge method into a unified automated annotation pipeline. Ultimately, these methods will be applied to SSP-generated responses, enabling systematic evaluation of both the evaluation engine and the underlying SSP system, and thereby contributing to the iterative improvement of AI-supported clinical reasoning learning tools.

References

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

David A. Cook. 2024. Creating virtual patients using large language models: Scalable, global, and low cost. *Medical Teacher*, 3:1–3.

Lori A. Erby, Debra L. Roter, and Barbara B. Biesecker. 2011. Examination of standardized patient performance: accuracy and consistency of six standardized patients over time. *Patient Education and Counseling*, 85(2):194–200. Epub 2010 Nov 20.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Justin Sewell, Susannah Cornes, Rene F. Kizilcec, and Dennis Shung. 2025. Benchmarking generative ai for scoring medical student interviews in objective structured clinical examinations (osces). *Preprint*, arXiv:2501.13957.

Ipek Gonullu, Cansu Derya Doğan, Serap Erden, and Derya Gökmen. 2023. A study on the standard setting, validity, and reliability of a standardized patient performance rating scale - student version. *Annals of Medicine*, 55(1):490–501.

Megan Gray, Austin Baird, Taylor Sawyer, Jasmine James, Thea DeBroux, Michelle Bartlett, Jeanne Krick, and Rachel Umoren. 2024. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of chatgpt: Exploratory observational study. *JMIR Medical Education*, 10:e50705.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

R.M. Harden. 1988. What is an OSCE? *Medical Teacher*, 10(1):19–22.

Florian Holderried, Carolin Stegemann-Philipps, Lisa Herschbach, Jan-Alexander Moldt, Alexander Nevins, Jan Griewatz, Marc Holderried, Anne Herrmann-Werner, Thomas Festl-Wietek, and Martin Mahling. 2024. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, 10:e53961.

S. Laschinger, J. Medves, C. Pulling, R. McGraw, B. Waytuck, M.B. Harrison, and K. Gambeta. 2008. Effectiveness of simulation on health profession students' knowledge, skills, confidence and satisfaction. *International Journal of Evidence-Based Healthcare*, 24:278–302.

Kelly L. Lewis, Carrie A. Bohnert, William L. Gammon, Henrike Hölzer, Layla Lyman, Cheryl Smith, Tara M. Thompson, and Georgia Gliva-McConvey. 2017. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Advances in Simulation*, 2:10.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

T. Rau, J. Fegert, and H. Liebhardt. 2011. How high are the personnel costs for OSCE? A financial report on management aspects. *GMS Journal for Medical Education*, 28(1):16.

Neil Sardesai, Peter Russo, James Martin, and Anjali Sardesai. 2024. Utilizing generative conversational artificial intelligence to create simulated patient encounters: A pilot study for anaesthesia training. *Postgraduate Medical Journal*, 100(1182):237–241.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Naoki Shindo and Masaki Uto. 2024. ChatGPT-based virtual standardized patient that amends overly detailed responses in objective structured clinical examinations. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 263–269, Cham. Springer Nature Switzerland.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards Conversational Diagnostic AI. *arXiv preprint arXiv:2401.05654*.

Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, and Hideo Ino. 2024. Enhancing medical interview skills through AI-simulated patient interactions: Non-randomized controlled trial. *JMIR Medical Education*, 10:e58753.

A Examples of Human-Labeled SP Responses

A.1 Correct

Student Question: "Consistently how long has your breathing difficulty been going on? Has it gotten worse recently?"
SP Response: "Yeah, uh, it's only been going on for about two years. Um, it's been the same from the beginning."
Explanation: The SP answered both questions correctly and did not provide additional detail.

A.2 Too Much Information

Student Question: "So, when you're having a really bad shortness of breath, do you have feelings like you're going to die or like it's like impending doom?"
SP Response: "No, but I'll just have that pounding in my chest and neck kind of on and off sometimes when I feel like I can't breathe."
Explanation: The student is asking only about the SP's thoughts or feelings during shortness of breath. The SP provided "pounding in my chest and neck" in addition to denying feelings of impending doom.

A.3 Too Little Information

Student Question: "So, I'm just going to go through a bunch of questions, and just tell me if you've had any of them in the last two weeks, okay? Have you had any headaches?"
SP Response: "Well, I do get stress headaches at work sometimes."
Explanation: The complete response for headaches is "When I'm at work, I get this feeling like a band around my head."

A.4 Incorrect

Student Question: "Can you tell me a bit more about how your periods are, like how many pads you use, kind of like what the experience is with those?"
SP Response: "I'll go through maybe four to five pads each day while I'm on my period."
Explanation: Her periods are heavy only during the first three days, not throughout her period.

A.5 Not Applicable

Student Question: "Well, thank you so much for talking to me and coming in when you did. I will be with you to support you throughout this process."
SP Response: "All right, thank you. I appreciate everything."
Explanation: The question and the response are not related to the case content.

LLM-Human Alignment in Evaluating Teacher Questioning Practices: Beyond Ratings to Explanation

Ruikun Hou[1,2], Tim Fütterer[2], Babette Bühler[1], Patrick Schreyer[3], Peter Gerjets[4], Ulrich Trautwein[2], Enkelejda Kasneci[1],

[1]Technical University of Munich, [2]University of Tübingen,
[3]University of Kassel, [4]Leibniz-Institut für Wissensmedien

{ruikun.hou, babette.buehler, enkelejda.kasneci}@tum.de, {tim.fuetterer, ulrich.trautwein}@uni-tuebingen.de
schreyer@uni-kassel.de, p.gerjets@iwm-tuebingen.de

Abstract

The systematic assessment of teaching quality through classroom observation is a critical yet challenging task in educational research. Traditionally, trained raters evaluate instructional practices by analyzing classroom interactions (e.g., watching videos and annotating transcripts) based on structured protocols. Whereas the potential of using Large Language Models (LLMs) to automate teaching quality assessment is increasingly being explored, few studies have examined the underlying reasoning behind those generated holistic scores, which could provide specific feedback for raters and teachers. In this study, we investigate the alignment between LLM- and human-generated assessments of teacher questioning practices, focusing on both quality rating agreement and evidence selection overlap. Specifically, advanced GPT models (GPT-4o and o1) were prompted using Chain-of-Thought (CoT) reasoning to analyze transcripts sequentially by extracting textual evidence, classifying question types, and assigning ratings. Analyzing 28 lesson transcript segments from the Global Teaching Insights study, each carefully annotated with highlights related to questioning practices, we found that CoT prompting generally improved both rating and evidence alignment compared to basic instructions. Under CoT reasoning, GPT-4o achieved the highest Quadratic Weighted Kappa score of 0.33 for quality rating agreement, whereas o1-extracted evidence yielded the highest character-level Intersection over Union of 0.14 with human transcript annotations. Qualitative analyses revealed that LLM and human annotations aligned in identifying explicit questioning forms, but they differed in annotation scope and granularity. Our study highlights LLMs' potential to enhance the explainability of rating decisions, assist manual assessment by highlighting relevant discourse evidence, and suggest possible approaches to offer teachers specific feedback that goes beyond numerical scores.

1 Introduction

Observing teaching practices in classrooms provides a crucial approach to assessing teaching quality and promoting teachers' professional development (Pianta and Hamre, 2009; Seidel and Shavelson, 2007; Praetorius et al., 2025). To systematically evaluate the quality of teacher-student interactions, multiple classroom observation protocols have been developed over the past decades, such as the Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) and the Protocol for Language Arts Teaching Observations (PLATO) (Grossman et al., 2013). These structured protocols typically assess multiple facets of teaching dynamics, e.g., instructional practices, emotional support, and classroom management. To code such protocols, trained raters capture important events of classroom interactions from videotaped lessons and assign scores to pre-defined teaching quality dimensions based on the observed evidence within a lesson segment. Due to the complex nature of classroom interactions, this manual observation process often requires substantial human effort and time. In addition, raters typically undergo intensive training and pass quality control checks to ensure rating reliability. Given these resource-intensive demands, automated assessment approaches using artificial intelligence (AI) techniques could enhance both the scalability of classroom observation studies and the frequency of feedback provided to teachers.

Rapid advancements in Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and analyzing natural language, leading to their expanding applications across diverse fields (Yang et al., 2024). In educational contexts, LLMs have been explored for various tasks to foster teaching and learning (Kasneci et al., 2023), including essay writing assessment (Seßler et al., 2023, 2024), teachable agents for programming (Ma et al., 2024), and instructional plan genera-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 239–249

tion (Hu et al., 2024). Meanwhile, the emerging paradigm of "LLMs-as-Judges" has gained increasing attention for leveraging LLMs as evaluators in complex tasks (Gu et al., 2024), presenting new possibilities for supporting classroom observation procedures. Recent studies have explored the use of LLMs for assessing teaching practices through the analysis of lesson transcripts (Wang and Demszky, 2023; Tran et al., 2024). Whereas in these studies, LLMs are prompted to provide supporting evidence before generating ratings, their evaluations focus primarily on the consistency between model-generated and human-assigned scores, leaving the alignment of evidence-based rationales unexplored. As classroom discourse often involves complex interactions, understanding how LLMs analyze teaching dynamics and justify their rating decisions is critical for validating their assessment capabilities and ensuring the interpretability and trustworthiness of automated evaluations.

In this study, we investigate rating agreement and the correspondence between LLM-identified and human-documented evidence in assessing teacher questioning practices. The analysis is based on lesson transcript data from the German subset of Global Teaching Insights (GTI) (OECD, 2020), a large-scale international classroom observation study. These transcripts document authentic mathematics instruction dialogues and contain detailed annotations from trained raters, including highlighted text spans that reflect specific teaching practices. We focus on the Questioning component within the GTI observation protocol (Bell et al., 2018a), as questioning practices can be effectively analyzed using text transcripts alone, whereas other dimensions (e.g., social-emotional support) may rely more on non-verbal cues such as visual behaviors. Furthermore, teacher questioning serves as a prominent feature of classroom discourse, and effective questioning practices play a pivotal role in student cognitive engagement and learning outcomes (Redfield and Rousseau, 1981; Chin, 2007). The GTI Questioning component reflects the cognitive demand of teacher questions, emphasizing how they engage students at different levels of cognitive complexity and depth of processing information. To automate assessment, we leverage the zero-shot capabilities of advanced GPT models (GPT-4o and o1), using Chain-of-Thought (CoT) reasoning (Wei et al., 2022) to guide them in identifying key evidence, categorizing question types, and assigning ratings. Afterward, we analyze human and LLM alignment in holistic ratings, evidence overlap at both character and span levels, and questioning classification. This comprehensive analysis reveals how automated assessment approaches correspond to authentic human rating practices, providing insights into LLMs' potential to support manual observation processes and offer teachers concrete feedback.

2 Related Work

2.1 Automated Questioning Identification

Early research in automated identification of teacher questions focused on analyzing classroom discourse transcripts and audio recordings. (Donnelly et al., 2017) trained machine learning (ML) models that combined linguistic, acoustic, and context features to identify whether a teacher utterance constitutes a question. (Kelly et al., 2018) developed automated methods to estimate the proportion of authentic questioning in a class period. Recent studies have advanced beyond binary question detection to analyze specific questioning strategies. (Alic et al., 2022) employed both supervised and unsupervised ML approaches to distinguish between "funneling" questions that guide students toward normative answers and "focusing" questions that encourage deeper thinking. Similarly, (Datta et al., 2023) leveraged pre-trained language models to classify teacher questions into four categories: probing, procedural, expository, and others. Moreover, (Kupor et al., 2023) fine-tuned GPT models to identify various instructional talk moves, including questioning strategies such as eliciting and probing student ideas. Whereas these studies demonstrated advances in automated questioning identification, they primarily treated questions as isolated instances rather than analyzing patterns of questioning practices within the broader context of classroom discourse.

2.2 LLMs in Teaching Practice Assessment

With recent progress in LLMs, researchers have explored their potential for automated assessment of teaching practices by analyzing classroom transcripts. One of the earliest studies in this area was conducted by (Wang and Demszky, 2023), who leveraged GPT-3.5's zero-shot capability with different prompting methods to evaluate several aspects of teaching practices based on the CLASS and Mathematical Quality Instruction (MQI) (Hill et al., 2008) frameworks. In addition to score prediction, they prompted GPT-3.5 to identify exemplary and problematic examples for each assessed dimension and to generate suggestions for how teachers could enhance student reasoning within the given classroom discourse. Analyzing a dataset of 100 transcript segments, they found that model-predicted scores showed generally poor alignment with human-assigned ratings and that employing CoT reasoning did not improve performance. Following this direction, (Hou et al., 2024) demonstrated that the more advanced GPT-4 model achieved superior zero-shot performance compared to GPT-3.5 in assessing classroom social-emotional support levels. Moreover, (Tran et al., 2024) examined how different task formulations affect LLMs' performance in evaluating classroom discussion quality, finding that explicitly guiding LLMs to extract relevant dialogue turns improved rating accuracy. Additionally, (Whitehill and LoCasale-Crouch, 2024) proposed a novel approach to assessing CLASS scores by prompting Llama 2 (Touvron et al., 2023) to identify behavioral indicators in individual teacher utterances, then aggregating these to predict holistic scores via linear regression. Their results showed that this automated approach could approach human inter-rater reliability while providing explanations at the utterance level.

Whereas existing studies explored prompting LLMs to extract relevant utterances beyond rating scores, they either focused solely on rating performance (Hou et al., 2024; Tran et al., 2024) or validated the extracted evidence through external reviewers (such as recruited teachers in (Wang and Demszky, 2023) and authors themselves in (Whitehill and LoCasale-Crouch, 2024)) rather than examining their alignment with trained raters directly involved in the protocol coding process. In this context, our study contributes to this line of research by (1) conducting a comprehensive analysis of both rating alignment and evidence selection overlap between LLM outputs and human annotations, (2) examining whether CoT reasoning enhances LLMs' capability to replicate human assessment procedures, and (3) providing insights into the similarities and differences between LLM and human approaches in extracting discourse evidence to reason their rating decisions.

3 Methodology

As illustrated in Figure 1, our study investigates the extent to which LLM reasoning aligns with manual annotations in assessing teacher questioning practices, comparing both rating assignments and the evidence selected from lesson transcripts to justify these decisions. This fosters the comprehension of LLM explainability, which is critical for the trustworthy use of AI in educational contexts.

3.1 Dataset

Our analysis utilized data from Germany, one of the participating countries in the Global Teaching Insights (GTI) study (formerly known as Teaching and Learning International Survey–Video (TALIS–Video)) (OECD, 2020). The GTI study systematically collected extensive classroom data on the teaching of quadratic equations and conducted a comprehensive global analysis of effective instructional practices. The German data includes 100 lesson recordings from 50 classrooms with 1,140 students across 38 schools. Anonymized lesson transcripts were created manually from the videos, with timestamps and speaker identifiers (e.g., "L" for teachers, "S01", "S02" for students).

Grounded in the GTI video observation protocol (Bell et al., 2018a), each lesson was divided into 16-minute segments, with each segment rated by intensively trained raters on a 1-4 scale across six domains. Each domain contains three components of instructional practice. Our study focused on the Questioning component, an important element of the Discourse domain. Effective questioning practices that facilitate student learning require students to engage across multiple levels of cognitive reasoning, with particular emphasis on higher-order thinking skills (Henningsen and Stein, 1997). To this end, the GTI Questioning component evaluates the cognitive demands of teacher questions, categorizing them into three types: (1) questions that prompt students to recall information, report answers, provide yes/no responses, or define terms; (2) questions that require students to summarize, explain, classify, or apply rules, processes, or formulas; and (3) questions that challenge students to analyze, synthesize, justify, or conjecture. Table 1 presents examples of each question type. The four-point rating scale reflects the relative emphasis and distribution of these question types throughout a lesson segment, with higher ratings indicating a greater proportion of more cognitively demanding questions (see Figure 3).

During video observation, raters were instructed to use the lesson transcripts as auxiliary tools, in which they highlighted relevant text spans and

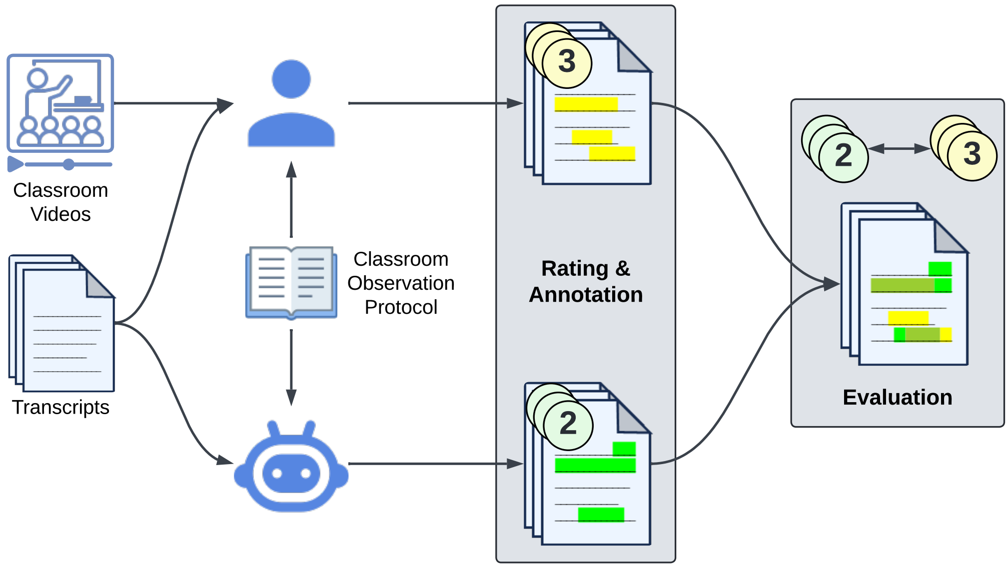


Figure 1. An overview of our study. LLMs are prompted to evaluate instructional questioning practices on a four-point scale and identify relevant evidence excerpts from transcripts. We then analyze the alignment between LLMs' assigned ratings and text selections with those provided by a trained rater.

Questions that request students to...	**Examples**
recall, report an answer, provide yes/no answers, and/or define terms	(1) What did you get Patrick? (2) What is the equivalence principle? (3) What is A? B? C? (4) Did you understand that explanation? (5) Do you remember what we did yesterday?
summarize, explain, classify, or apply rules, processes, or formulas	(1) Can you tell me how did you get this answer? (2) Let's see if substituting 4 and 8 each into x in equation 2 would work. Why do we substitute 4 and 8? (3) How many conditions do the roots of quadratic equations with one unknown have? What are they?
analyze, synthesize, justify, or conjecture	(1) The perimeter of a rectangle is 20. What is the area of the rectangle? (2) What is the pattern you notice across the three problems we just solved? Look carefully. (3) Can you explain why you disagree? Why do you think completing the square is a more efficient approach than just using the quadratic equation for number 4 on the board?

Table 1. Three questioning types and their examples (Bell et al., 2018b).

```
0:07:40  T:  yes exactly no that is thus the case where we have no
             solution when do we get exactly one solution from our
             quadratic equation DC_QUEST(recall) S05
```

Figure 2. Visualization of an authentic manual annotation example (translated word-for-word from German to English for readability). DC_QUEST means the Questioning component in the Discourse domain. T: Teacher; S05: Student.

specified the corresponding components. As these transcript annotations served primarily as working notes to support the rating process rather than as systematically standardized documentation, individual raters could vary in their annotation styles. Some raters provided detailed evidence highlights for certain components while annotating others sparsely, reflecting individual documentation preferences. This authentic variation, although valuable for understanding real-world assessment practices, presented challenges for conducting a unified analysis across all raters. Therefore, to enable a focused analysis of questioning practices, we selected data from one rater who provided detailed textual evidence for the Questioning component, including fine-grained annotations that categorized questioning strategies. This resulted in a dataset of 8 lessons containing 28 segments and 149 highlighted text spans. Figure 2 illustrates an annotation example. Additionally, as each segment was assessed by two randomly assigned raters, we calculated the selected rater's agreement with others who evaluated the same set of segments for the Questioning component, resulting in a Quadratic Weighted Kappa (QWK) score of 0.48. To contextualize this agreement level, we extended the leave-one-rater-out analysis to all 14 raters over the whole GTI Germany dataset. The average QWK score across all raters was 0.23, indicating that the chosen rater exhibited relatively high consistency with others in evaluating questioning practices. Moreover, we processed the annotated transcripts using INCEpTION (Klie et al., 2018), an open-source text annotation platform, to convert the original annotations into a structured dataset by pairing highlighted text spans with corresponding question categories for subsequent comparison with LLM annotations.

3.2 ChatGPT Zero-Shot Annotation

We employed two state-of-the-art GPT models[1], GPT-4o (*gpt-4o-2024-11-20*) and o1 (*o1-2024-12-17*), to automatically assess questioning practices while extracting supporting evidence from classroom dialogues. For each model, we investigated two zero-shot prompting strategies: a basic prompt and Chain-of-Thought (CoT) (Wei et al., 2022) reasoning. The basic prompt (see Figure 3) positioned LLMs as expert raters in evaluating teaching practices and outlined multiple rating guidelines, including examples of the three questioning types (see Table 1) and detailed scoring criteria for each rating level. Additionally, the prompt incorporated a 16-minute segment transcript for evaluation and instructed LLMs to provide a rating score along with a list of supporting evidence in the form of verbatim excerpts. Model responses were required in JSON format to ensure consistency and facilitate systematic analysis. The CoT prompt maintained all elements of the basic prompt but introduced explicit procedural analysis steps shown in Figure 4. Besides rating assignment and evidence extraction, the model was required to classify each identified evidence excerpt into one of three predefined question categories in its JSON response. This additional categorization enabled a more granular analysis of instructional questioning practices.

We accessed both GPT models via the OpenAI API and used default hyperparameters for inference. Due to the variability in LLM outputs, we conducted three independent runs for each experimental setting and averaged their evaluation results.

3.3 Evaluation Metrics

To understand how closely automated assessments correspond to authentic human annotations in evaluating teacher questioning practices, we examined three aspects: (1) rating alignment, (2) textual evidence overlap, and (3) questioning categorization consistency. The used metrics are described below.

(1) First, we utilized Quadratic Weighted Kappa (QWK) to measure agreement between model-generated and human-assigned scores. QWK accounts for the ordinal nature of rating levels by penalizing larger disagreements more heavily than smaller ones. (2) Second, we explored the degree of overlap between model-extracted and human-highlighted evidence spans at both character and span levels. At the character level, for a given transcript segment, we mapped each evidence excerpt (from either the model or the human rater) to its respective position in the transcript by identifying its start and end character indices. Afterward, we calculated the Intersection over Union (IoU) between model-extracted and human-annotated evidence sets, defined as the ratio of overlapping characters relative to the total number of characters covered by either set. However, interactive classroom discourse differs from structured written text, often involving fragmented, overlapping, and context-dependent utterances. As a result, annotators could include varying amounts of surrounding context when marking the same evidence. This in-

[1] https://platform.openai.com/docs/models

Task
You are an expert in evaluating the quality of classroom interactions based on lesson transcripts. You will be provided with a German transcript of a mathematics lesson segment focusing on quadratic equations. The transcript includes timestamps and speaker annotations, with 'L' indicating the teacher and 'S' followed by an ID number identifying anonymous students. Your task is to assess the teaching quality dimension of 'Questioning', which evaluates the nature of the questions asked by teachers that request students engage in a range of types of cognitive reasoning.

Important Note
* Rhetorical questions (i.e., questions the teacher poses and either does not answer or answers him or herself) should not be counted during rating.
* The rater should focus on what kinds of questions characterize the segment.
* Here are three types of questions with examples:
    {*Examples of three question types*}

Rating Scale (1-4, low to high)
* Score 1: Questions generally request students recall, report an answer, provide yes/no answers, and/or define terms.
* Score 2: Questions generally request students recall, report an answer, provide yes/no answers, and/or define terms, although there are some questions that request students summarize, explain, classify, or apply rules, processes, or formulas.
* Score 3: Despite a few questions that request students recall, report, and/or define, most questions request that students summarize, explain, classify, or apply rules, processes, or formulas. There may be a small number of questions that request students analyze, synthesize, justify, or conjecture.
* Score 4: Questions request a mixture of recall, reporting, defining, summarizing, explaining, classifying, applying rules, processes, or formulas, analyzing, synthesizing, justifying, and/or conjecturing, but the emphasis is on questions that request students analyze, synthesize, justify, or conjecture.

Instructions
Provide a JSON response with a rating and a list of key evidence supporting your rating:
{
  "rating": <integer score>,
  "evidence": ["<exact quote 1>", "<exact quote 2>", ...]
}
Note: Use exact character-for-character text spans from the provided transcript as complete evidence. Do NOT modify, paraphrase, abbreviate, or omit any content (including punctuation and formatting).

Transcript
Below is the transcript to be rated, enclosed in triple backticks:
```{*Transcript*}```

Figure 3. Basic prompt including comprehensive coding rubrics and response instructions. {*Examples of three question types*} can be found in Table 1.

...
Instructions
Analyze the transcript following these steps:
1. Read the transcript carefully,
2. Identify teacher questions maching the three aforementioned types,
3. Rate the transcript based on the coding rubrics.
Provide a JSON response with a rating and a list of relevant teacher questions along with their types:
{
  "rating": <integer score>,
  "evidence":
    [{"question": <exact quote 1>, "type": <1, 2, or 3>}, ...]
}
Note: Use exact character-for-character text spans from the provided transcript as complete evidence. Do NOT modify, paraphrase, abbreviate, or omit any content (including punctuation and formatting).
...

Figure 4. Chain-of-Thought prompt (sharing identical rating guidelines with the basic prompt).
```

herent variation in annotation granularity made exact character-level alignment overly restrictive. To address this, we adopted a more flexible matching criterion at the span level. A model-extracted span was considered a match with a manual annotation if they overlapped within the same general region of the transcript. Based on these counts, we calculated Precision (i.e., proportion of matches among model-extracted spans), Recall (i.e., proportion of matches among human-annotated spans), and F1-Score (i.e., harmonic mean of Precision and Recall). The resulting metrics were averaged across all transcript segments. (3) Finally, for matched evidence spans, we evaluated the classification of question types utilizing weighted Precision (i.e., how many of model-predicted question types were correct), Recall (i.e., how many of human-labeled question types were identified), and F1-Score.

4 Results

Table 2 presents the results on the consistency between LLM- and human-assigned ratings. GPT-4o with the CoT prompt achieved the highest agreement with human ratings (QWK=0.33), yielding notable improvement over its basic prompt counterpart. In contrast, the o1 model showed similar QWK scores for both basic and CoT prompts. Subsequently, we analyzed the overlap between model-extracted and human-annotated evidence at both character and span levels, along with their question type categorization. The results are summarized in Table 3. Regarding the selection of textual evidence, CoT prompting generally yielded higher character-level overlap than the basic prompt across both models, with the o1 CoT variant achieving the highest IoU of 0.14. At the span level, o1 resulted in stronger alignment across all metrics compared to GPT-4o under both prompting conditions. Similar to the rating results, the o1 model maintained a consistent F1 score of 0.38 between basic and CoT prompts. For question type categorization, under CoT prompting, the o1 model achieved a higher F1-Score than GPT-4o in distinguishing questioning evidence across three levels of cognitive demand.

In addition to the quantitative analysis, Figure 5 illustrates an exemplary comparison between human and LLM (o1-CoT) evidence selections from a transcript excerpt. The excerpt presents classroom discourse at the start of a math lesson, where the teacher reviews quadratic equations through a series of questions. The visualization captures both areas of convergence and divergence between human and model annotations. The overlapping regions (lime green) indicate alignment in identifying explicit instructional questioning while also revealing variations in the amount of surrounding context included by annotators when highlighting the same evidence. Whereas model-specific annotations (pure green) frequently include direct references to individual students (e.g., "S18") and focus on explicit computational prompts such as "yes and q," human-only annotations (yellow), such as "now we substitute this into...," tend to reflect a more contextualized and indirect questioning approach directed at students. Moreover, this comparison reveals that both model and human annotations occasionally overlook certain questions, leaving them unmarked and thus excluded from the assessment (e.g., "-6/2 is how much").

5 Discussion

In this study, we explored the zero-shot capabilities of advanced LLMs in evaluating instructional questioning practices from classroom transcripts. Our comparative analysis between LLM judgments and authentic human annotations revealed both areas of alignment and notable discrepancies in assessment approaches. The overall inter-rater agreement (QWK=0.23, see Sect. 3.1) in the GTI Germany dataset for the Questioning component underscores the inherent subjectivity of classroom observation over 16-minute instruction segments, suggesting that even trained human raters may differ in their interpretations of questioning practices. Within this context, the alignment between LLMs and the selected rater (QWK up to 0.33), whereas falling below the selected rater's agreement with other raters (QWK=0.48) on the same segment set, indicates the potential of LLMs to serve as assistive tools in classroom observation. Our results showed a moderate overlap in evidence selections (IoU up to 0.14, F1 up to 0.38) between LLM outputs and human annotations. Through analysis of multiple transcript examples, we observed that, whereas LLMs and the human rater aligned in identifying explicit instructional questions, they differed in annotation scope and granularity. LLMs tended to provide more comprehensive coverage of questioning instances throughout the discourse and include more surrounding dialogue around each question. It is important to consider the nature of human-annotated transcripts, which were created as prac-

Model	Prompt Type	QWK
Human Raters	-	0.48
GPT-4o	Basic	0.17
	CoT	0.33
o1	Basic	0.22
	CoT	0.23

Table 2. Results of agreement between LLM-generated scores and human-assigned ratings. For human raters, QWK is computed by comparing the chosen rater's scores with those of other raters who evaluated the same set of segments.

Model	Prompt Type	Evidence Overlap				Question Categorization		
		IoU	Recall	Precision	F1	Recall	Precision	F1
GPT-4o	Basic	0.06	0.14	0.23	0.18	-	-	-
	CoT	0.10	0.21	0.29	0.24	0.44	**0.67**	0.50
o1	Basic	0.09	0.33	**0.45**	**0.38**	-	-	-
	CoT	**0.14**	**0.52**	0.29	**0.38**	**0.61**	0.55	**0.58**

Table 3. Results of LLM and human alignment in evidence selections and question categorization. Bold values indicate the highest alignment for each metric.

0:00:00 T: I wish you all a very good morning (C: good morning) so we continue with quadratic equations we started in the last lesson solving these equations using the solution formula (.) do you still remember what the solution formula is if not take a look in your folder in the notebook S18 (S18: x1 x2 = uhm =-p/2²+/- uhm √(p/2)²-q) exactly you have already solved some exercises that were in normal form let‘s do one more example x²+6x+8=0 (.) what is p in this equation (.) S06 (S06: 6) yes and q S06 (S06: 8) yes (.) now we substitute this into the solution formula who will dictate the equation for me S15 (S15: uhm x1,2 uhm =- uhm 6/2) mhm (S15: +/- uhm √ uhm (uhm 6/2)²-8) exactly -6/2 is how much S17 (S17: - 3) -3 (.) how large is this term use a calculator if necessary (.) 6/2²-8 S19 (S19: 1) yes so -3+/-1 (.) what is then x1 S17 (S17: -2) say the full calculation (S17: uhm -3+1) exactly (S17: equals minu-) and x2 (S17: -3-1 (.) =-4) correct so our solution set is -2 and -4

Figure 5. Visualization of evidence overlap between human and model (o1-CoT) annotations in a transcript excerpt (translated from German to English). Yellow: human-only annotations; Pure green: model-only annotations; Lime green: overlapping selections. T: Teacher; C: Class; S06, S18, ...: Student.

tical notes rather than exhaustive documentation. Given that human raters assessed multiple components simultaneously, they may prioritize salient evidence that most effectively supports their evaluations, potentially omitting trivial or redundant instances. For example, in Figure 5, whereas the model identified some short questions, human annotations tended to focus on more substantive spans that warranted explicit documentation. This selective nature reflects authentic annotation practices in real-world classroom observation settings.

Moreover, our findings revealed distinct patterns in how GPT-4o and o1 performed under basic and CoT prompting. For GPT-4o, CoT reasoning resulted in higher agreement with human annotations across all metrics compared to the basic instruction, indicating that explicit guidance through structured reasoning steps helps LLMs better approximate human assessment practices. In contrast, o1 yielded comparable alignment levels between basic and CoT prompts, possibly attributed to o1's intrinsic reasoning mechanism that enables it to infer and apply implicit procedural analysis from basic instructions paired with rating guidelines. Further, whereas GPT-4o with CoT achieved a higher rating agreement than o1, o1 led to greater consistency with human annotations in evidence selections and question type classification. This discrepancy suggests potential differences in how human raters and LLMs translate identified evidence into holistic scores, indicating the complexity of teaching quality assessment. Additionally, o1 generates numerous intermediate tokens during its internal reasoning process before producing a final response, resulting in longer inference time and higher API costs. Thus, GPT-4o with CoT prompting might present a more balanced solution, offering a trade-off between accuracy and efficiency.

Our findings suggest promising implications for integrating LLMs into classroom observation. Unlike traditional automated methods (Ramakrishnan et al., 2021; James et al., 2018) that often lack explainability, LLMs can explicitly identify relevant discourse evidence to support their ratings. Thus, they could serve as complementary raters by suggesting highlights in transcripts for human raters to validate or refine. This human-in-the-loop approach would reduce the manual workload of sifting through lengthy transcripts while maintaining expert-level judgment. Beyond assisting manual observation, LLM-generated evidence-based assessments could form the basis for systems that provide teachers with timely feedback, offering both holistic ratings and representative examples of teaching practices (e.g., high-quality questions that promote students' high-order thinking). Moreover, these automated analyses could be valuable as training materials for novice raters by providing annotated instances of different questioning types and their impact on instruction. However, given the high-stakes nature of teaching quality assessment and potential algorithmic biases, it is crucial to recognize that LLMs should complement, rather than replace, manual coding or professional development resources. This allows raters and teachers to choose how to utilize these automated annotations to refine professional skills at their own pace.

One limitation of our study is its focus solely on the GTI Questioning component. While this verbally-oriented practice suits transcript analysis, the generalizability of our findings to other instructional components (e.g., teacher feedback) remains to be explored. Moreover, although the selected rater exhibited above-average agreement with peers, expanding the dataset to include more raters would allow for a more comprehensive understanding of human annotation variability. Further, while CoT reasoning showed promising results, future research could benefit from employing more sophisticated promoting engineering strategies, like in-context learning. Additionally, with the rise of open-source models such as Llama (Touvron et al., 2023), future work could explore their applications to classroom observation tasks, offering potential alternatives to closed-source models and enabling large-scale studies at reduced operational costs.

6 Conclusion

This study examines the alignment between LLM- and human-generated assessments of teacher questioning practices, involving both rating assignments and evidence extraction. CoT prompting proves effective in guiding LLMs to approximate human assessment procedures. Although LLM and human annotations exhibit different patterns in granularity and context inclusion, these variations highlight complementary approaches to identifying instructional events. Our findings suggest LLMs' potential to enhance the explainability and trustworthiness of rating decisions and to facilitate manual observation and foster teacher professional growth in future applications.

References

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. *arXiv preprint arXiv:2208.04715*.

Courtney Bell, Yi Qi, Margaret Witherspoon, Mariana Barragan, and Heather Howell. 2018a. Annex a: Talis video observation codes: Holistic domain ratings and components. In *Global Teaching Insights: Technical Report*. OECD.

Courtney Bell, Yi Qi, Margaret Witherspoon, Mariana Barragan, and Heather Howell. 2018b. Annex a: Talis video training notes: Holistic domain ratings and components. In *Global Teaching Insights: Technical Report*. OECD.

Christine Chin. 2007. Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 44(6):815–843.

Debajyoti Datta, James P Bywater, Maria Phillips, Sarah Lilly, Jennifer L Chiu, Ginger S Watson, and Donald E Brown. 2023. Classifying mathematics teacher questions to support mathematical discourse. In *International Conference on Artificial Intelligence in Education*, pages 372–377. Springer.

Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 218–227.

Pam Grossman, Susanna Loeb, Julie Cohen, and James Wyckoff. 2013. Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*, 119(3):445–470.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Marjorie Henningsen and Mary Kay Stein. 1997. Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for research in mathematics education*, 28(5):524–549.

Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In *International Conference on Artificial Intelligence in Education*, pages 60–74. Springer.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.

Anusha James, Mohan Kashyap, Yi Han Victoria Chua, Tomasz Maszczyk, Ana Moreno Núñez, Rebecca Bull, and Justin Dauwels. 2018. Inferring the climate in classrooms from audio and video recordings: a machine learning approach. In *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, pages 983–988.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations*, pages 5–9.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*, pages 265–279. Springer.

OECD. 2020. *Global Teaching InSights: A Video Study of Teaching*. OECD, Paris.

Robert C Pianta and Bridget K Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, 38(2):109–119.

Robert C Pianta, Karen M La Paro, and Bridget K Hamre. 2008. *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing.

Anna-Katharina Praetorius, Charalambos Y Charalambous, Svenja Vieluf, Mirjam Steffensky, Richard Göllner, and Benjamin Fauth. 2025. Rethinking teaching-quality research: a reflection on the role of core working assumptions and possible pathways for future research. *School Effectiveness and School Improvement*, 36(2):314–334.

Anand Ramakrishnan, Brian Zylich, Erin Ottmar, Jennifer LoCasale-Crouch, and Jacob Whitehill. 2021. Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing*.

Doris L Redfield and Elaine Waldman Rousseau. 1981. A meta-analysis of experimental research on teacher questioning behavior. *Review of educational research*, 51(2):237–245.

Tina Seidel and Richard J Shavelson. 2007. Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of educational research*, 77(4):454–499.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *arXiv preprint arXiv:2411.16337*.

Kathrin Seßler, Tao Xiang, Lukas Bogenrieder, and Enkelejda Kasneci. 2023. Peer: Empowering writing with large language models. In *European Conference on Technology Enhanced Learning*, pages 755–761.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment. *arXiv preprint arXiv:2406.08680*.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Leveraging Fine-tuned Large Language Models in Item Parameter Prediction

Suhwa Han[1], Frank Rijmen[1], Allison Ames Boykin[2*], Susan Lottridge[1]

[1]Cambium Assessment, [2]National Board of Medical Examiners

Correspondence: suhwa.han@cambiumassessment.com

Abstract

Accurate prediction of item parameters using item characteristics has been a long-standing objective in educational measurement, and recent advances in natural language processing (NLP) and large language models (LLMs) have opened new possibilities for modeling item parameters directly from item text. In this study, we introduce novel fine-tuning approaches that leverage item text as well as structured item attribute variables for enhanced prediction. For benchmarking, we compare suggested approaches with a traditional tree-based machine learning model that uses item attributes as primary inputs. The proposed methods are evaluated on a dataset of over 1,000 operational English Language Arts (ELA) items, with both dichotomous and polytomous scoring. Our work offers a unique opportunity to evaluate the prediction of item difficulty for polytomous items as well as item discrimination—areas that have received limited attention in prior research.

1 Introduction

Item parameter prediction in educational measurement refers to the modeling of item response theory (IRT) model parameters such as difficulty by using item-level features inherent in the items (AlKhuzaey et al., 2024). Accurately and reliably predicted item parameters offer multiple benefits. First, it can reduce the heavy reliance on field testing in evaluating new items, which is costly and increases the risk of security breaches due to item exposure (Ulitzsch et al., 2025). If non-functional items, such as those that are too easy or not discriminative, can be identified through predictive modeling, test developers can save resources in vetting new items. In addition, item parameter prediction has broader implications beyond large-scale assessments. If the methodology is sufficiently validated, it can be applied to classroom settings, where educators evaluate how their own items align with summative scales and make data-informed adjustments to instructions.

Given its potential to support a wide range of assessment activities, item parameter prediction has been a long-standing objective in the field (Fischer, 1995; Ferrara et al., 2022). In particular, recent advances in natural language processing (NLP) techniques has enabled researchers to leverage textual information in items in predicting item parameters (AlKhuzaey et al., 2024; Benedetto et al., 2023). Researchers have utilized language models to extract surface-level linguistic features and/or to derive embeddings to capture deeper semantic meanings, which are then used as features in statistical models or machine-learning (ML) algorithms (Xue et al., 2020; Yaneva et al., 2019, 2023). More recently, fine-tuning large language models (LLMs) on item texts has shown improved predictive performance compared to feature-based approaches (Benedetto et al., 2021; Yaneva et al., 2024; Zu and Choi, 2023). Given these recent findings and ever-improving capabilities of LLMs, further investigation into the use of fine-tuned LLMs for the item parameter modeling appears warranted and timely.

1.1 Study Purpose and Contributions

The purpose of the current study is to examine the performance of fine-tuned LLMs and explore how they can be more effectively leveraged for the item parameter prediction. To this end, the study addresses several research questions that have received limited attention in the existing literature. First, it investigates ways to integrate textual information from items with additional item attributes—such as content classification variables—within a fine-tuned LLM architecture. The study also compares the performance of fine-tuned LLMs with traditional machine learning algorithms to evalu-

[*]Work conducted while the author was at Cambium Assessment.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 250–264
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

ate their relative performance. Third, the study suggests methodologies for applying LLMs in the prediction of parameters for polytomously scored items, an area that has been underexplored. Finally, the study investigates the capacity of fine-tuned LLMs to predict discrimination parameters, which has historically received less attention in item parameter modeling.

2 Prior Research on IRT Parameter Prediction Using Fine-tuned LLMs

One of the key advantages of using LLMs is that they can be further fine-tuned for a specific task on top of its general linguistic capabilities obtained from pre-training. Through fine-tuning, the model parameters are optimized for a given task, enabling improved performance on downstream applications. Due to this flexibility and ability to directly model textual input, fine-tuned LLMs have been increasingly used to predict IRT parameters in the context of educational assessments.

Benedetto et al. (2021) demonstrated that a fine-tuned BERT model (Devlin et al., 2019) could effectively estimate difficulty parameters of Rasch model (Rasch, 1993) using items from e-learning platforms. They found that the fine-tuning approach reduced estimation error by 6.5% compared to traditional feature-based ML approaches using TF-IDF and embeddings. Zu and Choi (2023) also examined performance of fine-tuned RoBERTa model (Liu et al., 2019) in predicting item difficulty parameters of autogenerated multiple-choice items for English-language proficiency tests. By first fine-tuning RoBERTa on a key classification task subsequently adapting it for difficulty prediction, they achieved stronger correlations—$r = .733$ for listening and $r = .684$ for reading—compared to traditional methods based on hand-crafted features and embeddings.

Building on this line of work, Gombert et al. (2024) explored fine-tuning various transformer-based models to jointly predict both item difficulty and response time for multiplice-choice items in a medical licensure exam. They introduced architectural enhancements to LLMs by incorporating scalar mixing and a custom regression head. While their approach ranked first in a share task competition, their predictive power was relatively modest, yielding a maximum correlation of .27. Using a different dataset—math proficiency test data set for adults, Feng et al. (2024) found that fine-tuned RoBERTa achieved the best prediction, outperforming linear regression and zero-shot prompting approaches in terms of minimizing mean squared error, while explaining approximately 43% of the variance in the difficulty.

3 Methods

While several studies have successfully fine-tuned LLMs for item parameter prediction, to the best of the authors' knowledge, none have explored the integration of item attribute variables—such as content-wise classifications—directly within the LLM fine-tuning process. Given that most operationally maintained items are accompanied by such metadata, leveraging these additional features may enhance the predictive performance of LLMs.

In addition, the dataset used in this study is notable for its diversity, encompassing a range of item types that are currently operationally used in a large scale assessment. As such, evaluating the performance of the proposed methods on this dataset can provide insights that are both methodologically novel and practically relevant.

3.1 Dataset

The dataset used in this study consists of 1,119 items to assess English Language Art (ELA) proficiency for Grade 6 students. These selected items were drawn from the operational pool for the 2024-2025 Smarter Balanced assessment administration. The authors gratefully acknowledge the collaboration and support of Smarter Balanced in providing access to this high quality dataset.

These items span seven distinct item types, including five machine-scorable types (EBSR, HT, MC, MI, and MS) and two constructed response types (SA, WER) (see Appendix A for the description of the item types). In this dataset, the machine-scorable items were scored dichotomously, and constructed response items were all scored polytomously. The items were field-tested across 8 years (2014, 2015, 2016, 2017, 2018, 2019, 2020 and 2022), and include 935 summative items and 184 interim ones. The actual counts of the item types across the field-tested years can be found in Appendix B.

Two Sources of Information: Texts and Item Attributes. Each item in the dataset was associated with two types of texts: a stimulus text and as item text. Since these items were ELA items, stimulus texts typically consisted of a reading passage designed to provide necessary information needed to answer questions. Item texts contained the actual question or prompt. To optimize input construction for the modeling, we concatenated the item text followed by the stimulus text. This ordering was to ensure inclusion of the item prompt within limited sequence length in the LLM modeling process. For polytomously scored constructed-response items, an additional piece of textual information was incorporated: rubric texts. The rubric texts were needed to provide unique information to model multiple difficulty parameters for the polytomous items. To ensure this critical information was retained in the modeling, rubric texts were prepended to the item text, followed by the stimulus text.[1]

In addition to textual data, this study extracted a set of item attribute variables to evaluate their contribution to the prediction. In total, 152 attribute variables were compiled: 59 content-based specification variables and 93 hand-crafted linguistic features extracted from both item and stimulus texts based on Baldwin et al. (2021) (see Appendix C for examples of item attribute variables used in this study.)

Target Variable: Banked IRT parameters. The target variables in this study were IRT parameters—both item difficulty and discrimination parameters—from the operational Smarter Balanced item bank. In the bank, the dichotomous items were calibrated using two-parameter logistic (2PL) model (Birnbaum, 1968), while the polytomous items were calibrated using generalized partial credit model (GPCM) (Muraki, 1992).

3.2 Item Response Theory Model: Generalized Partial Credit Model

Because the 2PL model is a special case of GPCM, this study treated the 2PL-calibrated parameters as a simplified instance of GPCM. GPCM describes the probability of an examinee with a latent trait level θ to obtain a score of $v \in \{0, 1, \dots, m_i\}$ for item i as:

$$p_{iv} = \frac{\exp(\sum_{r=0}^{v} Da_i(\theta - b_i + d_{ir}))}{\sum_{c=0}^{m_i} \exp(\sum_{r=0}^{c} Da_i(\theta - b_i + d_{ir}))},$$

where a_i and b_i respectively denote the overall discrimination and difficulty parameters for item i, and d_{ir} represents the step parameter for the category r for the item. The GPCM parameters used in this study were estimated under the constraints $d_{i0} = 0$ and $\sum_r d_{ir} = 0$. For the difficulty modeling, the target variable was defined as the item category threshold $b_i - d_{ir}$ for polytomously scored items, where $d_{ir} = 0$ for the dichotomous cases.[2] The discrimination parameter a_i was used as the target variable for modeling item discrimination.

3.3 Sampling

This study adopted an 80%:10%:10% split approach to create training, development, and test sets, respectively. The training set was used to train the models, while the development set was used to guide hyperparameter tuning and modeling decisions. The test set was held out to ensure the generalized performance of the trained models. For the sampling, items were stratified by item types to equally distribute all item types across the sets. A detailed breakdown of item type counts across the three sets is provided in Appendix D.

3.4 Fine-tuning LLMs Using Texts as Primary Input

To evaluate how item texts can be fine-tuned for predicting IRT parameters, we implemented two distinct LLM architectures as shown in Appendix E: a baseline model that uses only item texts as input, and an experimental model that incorporates both item texts and attributes as input. In both architectures, the model started by encoding item text into static token-level embeddings of 768 dimensions using a pre-trained LLM. These token embeddings were then aggregated to a single 768-dimensional vector using mean pooling. Subsequently, the pooled vector was passed through three consecutive hidden layers, with each normalized by batch normalization, followed by Leaky ReLU activations (Maas et al., 2013). Finally, a regression head was attached to the last hidden layer to produce a continuous output for the target IRT parameters.

Model with Item Attributes. As shown on the right side of Figure 1, the experimental model with item attributes was implemented by concatenating item attribute variables with the pooled embeddings before passing it through the hidden layers. This design allowed the model to leverage both item text and attributes seamlessly within the fine-tuning process.

[1] Although the text input consists of rubric, item and stimulus texts, we refer to this combined input as "item text" for brevity throughout the remainder of this paper.

[2] These category threshold parameters are referred to as *difficulty* parameters throughout the remainder of this paper for simplicity.

Within this architecture, we explored two model variants: one that uses the raw item attribute variables directly for the concatenation and the other that concatenates predicted values from another predictive ML model based on item attributes. We refer to the first variant as the *feature augmented* model, where the raw variables are used to augment the LLM feature space. The other variant is referred to as the *transfer learning* model, as it transfers predictive outputs from a separate model into the LLM. Note that the inclusion of two additional model variants resulted in a total of three fine-tuned LLM approaches: (i) *baseline* models using item texts as the sole input, (ii) *feature augmented* models using raw item attributes alongside the texts, and (iii) *transfer learning* models that used LLM-predicted values as an additional input.

Selected Pre-trained LLMs. Four different LLMs were experimented in this study to evaluate differences in performance: RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), XLNet (Yang et al., 2019), and Longformer (Beltagy et al., 2020). Three models except for XLNet were encoder-based model. These encoder models were chosen as they are designed to transform texts into contextualized embeddings, which can be seamlessly adjusted for regression tasks. XLNet, while not an encoder-only model, was chosen as it often outperformed other transformer-based models due to its recurrent mechanism that can accommodate long-term dependencies (Ormerod et al., 2023).

Hyperparmeters. The following hyperparameter settings were used throughout this study:

- **Batch size:** 16 for most models; reduced to 8 for DeBERTa due to GPU memory limitations
- **Number of epochs:** 40 for most models; increased to 100 for DeBERTa to ensure sufficient updates to mitigate noisier gradients due to the smaller batch size [3]
- **Sequence length:** 512 [4]
- **Learning rate:** $1e^{-5}$ for pre-trained LLM parameters and $1e^{-2}$ for other model parameters

3.5 Traditional ML Approach Using Item Attributes as Primary Input: CatBoost

In addition to fine-tuned LLMs, this study implemented a traditional ML approach to predict IRT parameters using item attribute variables as primary input. Specifically, CatBoost (Dorogush et al., 2018)—a gradient boosting algorithm based on decision trees—was chosen due to its ability to natively handle categorical variables without requiring dummy coding [5]. Following the structure used in the LLM-based modeling, three variants of the CatBoost approach were developed: (i) *baseline* models, (ii) *feature augmented* models, and (iii) *transfer learning* models.

In the *baseline* model, only the item attribute variables were used as input features. For the *feature augmented* model, embeddings extracted from each of the fine-tuned baseline LLMs were appended to the item attribute feature set. In the *transfer learning* model, the predicted values generated by the fine-tuned baseline LLMs were appended to the item attribute feature set as an additional predictor.

4 Results

The predictive performance of the models was evaluated using two metrics: Pearson correlation (COR) and root mean squared error (RMSE). These metrics were calculated by treating the banked IRT parameters as the ground truth for the development and test sets. Table 1 displays descriptive statistics of the banked IRT parameters across the subsets.

Parameter	Statistic	Train	Dev	Test
Discrimination (a_i)	Min	0.110	0.166	0.203
	1st Qu.	0.448	0.437	0.466
	Median	0.579	0.548	0.633
	Mean	0.586	0.553	0.610
	3rd Qu.	0.718	0.681	0.743
	Max	1.354	1.043	1.075
	SD	0.199	0.179	0.209
Difficulty ($b_i - d_{ir}$)	Min	-2.719	-1.770	-1.631
	1st Qu.	-0.175	-0.024	-0.086
	Median	0.798	0.863	0.648
	Mean	0.891	0.978	0.805
	3rd Qu.	1.766	1.770	1.681
	Max	9.068	6.251	4.607
	SD	1.379	1.391	1.257

Table 1: Summary statistics for the banked IRT parameters across the sets.

The prediction results presented in the following are all based on the held-out test set.

[3]The study used the model state after completing all training epochs as the final model.

[4]While Longformer and XLNet can process inputs longer than 512 tokens, the sequence length was fixed at 512 based on preliminary analysis showing no performance advantage from longer inputs.

[5]The study used CatBoost v1.2.7 with default hyperparameter settings.

4.1 Item Difficulty Prediction Results

Table 2 presents the performance of fine-tuned LLM and CatBoost models on the item difficulty prediction task.

Positive Impact of Item Attribute Integration in LLM Fine-Tuning. The results demonstrate promising prediction accuracy for baseline fine-tuned LLMs, achieving correlations close to 0.7 with Longformer and DeBERTa. These findings suggest that item text alone can contribute substantial information relevant to predicting difficulty parameters when leveraged through LLM fine-tuning. Further improvements were observed with the *feature augmented* models, where raw item attribute variables were integrated into the LLM fine-tuning. This method consistently (albeit marginally) outperformed the baseline across all four LLMs, yielding the highest correlations and lowest RMSEs in most of LLMs.[6] These results indicate that item attributes can provide additional information in prediting difficulty parameters. In contrast, the *transfer Learning* LLM models—where predicted values from CatBoost model were appended to inputs—suffered reduced performance. This indicates that incorporating predicted values from an external model may have simply added additional noise rather than signal, particularly when the predictions themselves were only moderately accurate (e.g., a correlation of 0.492 in this case).

Improved CatBoost Performance via LLM-Based Augmentation. The CatBoost Baseline model, which used only item attribute variables for predicting item difficulty, showed limited predictive power, with a correlation less than 0.5. However, its performance improved substantially when augmented with embeddings from fine-tuned LLMs, regardless of the LLM type. For example, augmenting item attributes with fine-tuned embeddings increased the correlation from .492 to .706 in the case of Longformer. A similar positive effect was observed with *transfer learning* model; when predicted values from fine-tuned LLMs were added as additional inputs, performance markedly improved over baseline.

While both the *feature augmented* and *transfer learning* CatBoost models showed notable gains, their performance remained short of the best results achieved by the fine-tuned LLMs—those augmented with raw item attribute variables.

[6]The magnitude of improvement was more pronounced in the development set results; see Appendix F.

4.2 Item Discrimination Prediction Results

Table 3 presents the item discrimination prediction performance of fine-tuned LLM and CatBoost models. The discrimination prediction results showed distinctively different patterns from the difficulty prediction.

Strong Performance of CatBoost for Discrimination Prediction. In contrast to the pattern observed in difficulty prediction, the baseline CatBoost model yielded the strongest performance for discrimination prediction among all conditions. As shown in Table 3, this baseline model, which used only item attribute variables, achieved the highest correlation (0.537) and the lowest RMSE (0.174), consistently outperforming all other model variants.

In comparison, the baseline fine-tuned LLM models performed notably worse than they had in the difficulty prediction task. When fine-tuned solely on item text, the LLMs produced correlation values as low as 0.310. Likely due to this low baseline performance, augmenting CatBoost with fine-tuned embeddings or LLM-based predictions resulted in noticeable drops in prediction accuracy. For instance, with RoBERTa, the addition of fine-tuned embeddings reduced the correlation from 0.537 (CatBoost baseline) to 0.396. This degradation further highlights the limited capacity of fine-tuned LLMs for modeling item discrimination.

Conversely, the contribution of item attribute variables to the fine-tuned LLMs was notable, leading to consistent performance gains. For example, Longformer's correlation improved from 0.337 in the baseline LLM to 0.425 with the addition of raw item attributes, and further increased to 0.477 when predictions from the CatBoost model were appended. This trend was consistent across all LLMs: the *transfer learning* fine-tuned LLM models always outperformed the baseline LLMs, often by substantial margins.

5 Discussion

In this study, we presented novel approaches for fine-tuning LLMs using item text to predict IRT parameters. Beyond the baseline model that relied solely on item text, we introduced structured methods for incorporating item attribute variables into the fine-tuning process to further enhance predictive performance. We also examined the use

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.633	<u>0.691</u>	0.686	0.691	1.029	<u>0.910</u>	0.964	0.967
	F.A.	**0.707**	**0.712**	**0.699**	**0.717**	**0.890**	**0.889**	**0.908**	0.966
	T.L.	0.563	0.644	0.547	0.609	1.070	0.980	1.095	1.069
CatBoost	Baseline	0.492	0.492	0.492	0.492	1.133	1.133	1.133	1.133
	F.A.	<u>0.669</u>	0.686	0.559	<u>0.706</u>	<u>0.965</u>	0.936	1.054	**0.907**
	T.L.	0.646	0.678	<u>0.688</u>	0.696	0.997	0.936	<u>0.919</u>	<u>0.913</u>

Table 2: Item difficulty prediction results on the test set using fine-tuned LLM and CatBoost models across three model variants. Within each LLM, **bold** marks the best performance and <u>underline</u> marks the second best. F.A.=Feature Augmented; T.L.=Transfer Learning. Corresponding results to the development set can be found in Appendix F.

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.394	0.310	0.334	0.337	0.196	0.205	0.218	0.200
	F.A.	0.348	0.320	0.324	0.425	0.203	0.201	0.208	0.196
	T.L.	<u>0.465</u>	<u>0.495</u>	<u>0.470</u>	<u>0.477</u>	<u>0.186</u>	0.195	0.276	<u>0.183</u>
CatBoost	Baseline	**0.537**	**0.537**	**0.537**	**0.537**	**0.174**	**0.174**	**0.174**	**0.174**
	F.A.	0.396	0.428	0.426	0.409	0.196	<u>0.192</u>	<u>0.190</u>	0.197
	T.L.	0.414	0.352	0.376	0.335	0.194	0.200	0.200	0.202

Table 3: Item discrimination prediction results on the test set using fine-tuned LLM and CatBoost models across three modeling variants. Within each LLM, **bold** marks the best performance and <u>underline</u> marks the second best. F.A. = Feature Augmented; T.L. = Transfer Learning. Corresponding results to the development set can be found in Appendix G.

.

of a traditional ML algorithm—CatBoost—using item attribute variables as primary inputs, and further investigated whether combining CatBoost with information derived from fine-tuned LLMs could improve prediction accuracy.

Performance of the suggested methods was evaluated using a large dataset of Grade 6 ELA assessment items. The dataset included a mix of dichotomously and polytomously scored items, offering a valuable opportunity to assess model performance on predicting multiple difficulty parameters in polytomous items. In addition, we also fully investigated the prediction performance of the item discrimination parameters, which has received limited attention in prior IRT parameter modeling research.

Our results suggested that predicting item difficulty parameters was a relatively more amenable modeling task, with several models achieving moderately high correlations. In contrast, predicting item discrimination parameters were found to be more challenging, consistently yielding lower performance. In particular, we found that the fine-tuned LLMs performed well in the difficulty prediction, but were susbtantially less effective for discrimination. This disparity indicates that item texts contain meaningful signals for modeling difficulty, but offer limited information in capturing item discrimination.

Interestingly, the traditional CatBoost model using only item attribute variables showed relatively strong performance in predicting discrimination parameters, achieving highest correlations and lowest RMSEs. This finding highlights the potential value of using structured item attribute features in modeling discrimination parameters and may offer useful direction to researchers and practitioners.

The study also explored the integration of two information sources—item text and item attributes—as inputs into the prediction models. This strategy showed mixed results. When the added information came from a strong predictive source, such as fine-tuned LLM derived values in the difficulty modeling, it considerably enhanced model performance. However, when the appended information had limited predictive quality, it often introduced noise and reduced accuracies. These findings highlight both the promise and the risks of multi-source modeling: while combining signals can enhance prediction, it is crucial to assess the individual contribution of each source before integration.

Limitations

This study is not without limitations. First, although we partitioned the dataset into training, development, and test sets, we did not employ full cross-validation during hyperparameter tuning. As a result, model performance may have been somewhat sensitive to the specific data split that we used. Second, all hyperparameter settings were optimized based on the development set performance for the difficulty prediction task. These settings were then applied to the discrimination prediction without futher tuning. Given the distinct nature of the prediction targets, task-specific hyperparameter optimization—particularly for discrimination modeling using fine-tuned LLMs—could have yielded improved performance. Third, while several models achieved strong correlations for difficulty prediction, the overall predictive accuracy indicates considerable potential for future improvement. This reflects the inherent complexity of this task and highlights the need for continued research.

Future Work

In an effort to improve the alignment of predicted values with the true parameters, the authors conducted preliminary investigation of a sequential approach that incorporates predicted values as informative priors within a Bayesian estimation framework with small samples, as explored in Ulitzsch et al. (2025). Initial results suggest that incorporating a small response sample—as small as 50 examinees—can significantly improve estimation accuracy. However, a detailed discussion of this extension lies beyond the scope of the current study and will be addressed in future work. In addition, future research will explore the differential performance of the predictions across item types as a larger and more diverse sample of items becomes available. Such analysis is expected to provide practical insights for practitioners by illuminating conditions where fine-tuned LLMs are most effective in predicting IRT parameters.

References

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157, Online. Association for Computational Linguistics.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Allan Birnbaum. 1968. Some latent trait models. *Statistical theories of mental test scores*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Wangyong Feng, Peter Tran, Hunter McNichols, Steven Sireci, and Andrew Lan. 2024. Using artificial intelligence to scale multiple choice math items. Presentation delivered at the Annual Conference of the Northeastern Educational Research Association, Trumbull, CT.

Steve Ferrara, Jeffrey T Steedle, and Roger S Frantz. 2022. Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, 35(3):237–253.

Gerhard H Fischer. 1995. The linear logistic test model. In *Rasch models: Foundations, recent developments, and applications*, pages 131–155. Springer.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, and 1 others. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA.

Eiji Muraki. 1992. A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, 16(2):159–176.

Christopher Ormerod, Amy Burkhardt, Mackenzie Young, and Sue Lottridge. 2023. Argumentation element annotation modeling using xlnet. *arXiv preprint arXiv:2311.06239*.

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Esther Ulitzsch, Dmitry Belov, Oliver Lüdtke, and Alexander Robitzsch. 2025. Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *Journal of Educational Measurement*.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 193–197.

Victoria Yaneva, Peter Baldwin, Janet Mee, and 1 others. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 11–20.

Victoria Yaneva, Peter Baldwin, Christopher Runyon, and 1 others. 2023. Extracting linguistic signal from item text and its application to modeling item characteristics. In *Advancing natural language processing in educational assessment*, pages 167–182. Routledge.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jiyun Zu and Ikkyu Choi. 2023. Predicting the psychometric properties of automatically generated items. Presentation delivered at the 88th Annual Meeting of the Psychometric Society, College Park, MD.

A Descriptions of Item Types

Abbreviation	**Item Type**	**Description**
EBSR	Evidence-Based Selected Response	This item type has two parts: Part A asks examinees to select a correct response from four options, and Part B asks them to identify textual support for their answer
HT	Hot Text	This item type asks examinees to either select a correct word or rearrange words/phrases by clicking and dragging
MC	Multiple Choice	This item type asks examinees to choose one answer from multiple options
MI	Match Interaction	This item type requires examinees to match text or images in rows to values in columns by clicking cells
MS	Multi Select	This item type asks examinees to select one or more options
SA	Short Answer	This item type asks examinees to enter a response using alphanumeric characters via a keyboard
WER	Writing Extended Response	This item type asks examinees to provide a longer written response using keyboard entry of alphanumeric characters

Table 4: Descriptions of the item types used in this study

B Item Counts by Year and Type

Year	EBSR	HT	MC	MI	MS	SA	WER	Total
2014	23	75	174	1	119	42	6	440
2015	20	79	127	1	85	31		343
2016		1	7		4			12
2017	1	2	8		3	2	3	19
2018	2	6	9	5	3	19	4	48
2019	11	24	84	4	14	1	2	140
2020						5		5
2022	3	16	55	4	11	21	2	112
Total	60	203	464	15	239	116	22	1119

Table 5: Item counts by field test year and item type

C Examples of Item Attribute Variables

Attribute Type	Variable Type	Label	Description
Content Spec	Categorical	itemType	Item types
Content Spec	Categorical	WERdimension	Dimension of WER items
Content Spec	Categorical	claim	Four main claims in Smarter Balanced ELA
Content Spec	Categorical	lowestLevel	Content standards for ELA
Content Spec	Categorical	stimGenre	Genre of stimulus
Content Spec	Numeric	IAT.Tables	Number of tables embedded in the item
Content Spec	Numeric	IAT.Images	Number of images embedded in the item
Content Spec	Numeric	choiceInt	Number of choice-type interactions in the item
Content Spec	Numeric	hotTextInt	Number of hot-text-type interactions in the item
Content Spec	Numeric	FleschEase	The Flesch Reading readability level measuring easiness of text
Content Spec	Numeric	FleschKinc	The Flesch Kincaid Readability level measuring US grade level required to understand text
Linguistic	Numeric	numWords	Number of words in the text
Linguistic	Numeric	numContWords	Number of content words in the text
Linguistic	Numeric	numPolySem	Number of words that have multiple meanings
Linguistic	Numeric	numWSenseNoun	Number of word senses for nouns
Linguistic	Numeric	avgSynTreeDep	Average depth of syntax trees in sentences
Linguistic	Numeric	notCommon2000	Number of words that are not in common 2000 words in Reuter corpus
Linguistic	Numeric	avgImage	Average rating of words based on how easily and quickly a mental image can be evoked, according to the MRC Psycholinguistic Database

Table 6: Example of item attribute features. Content Spec=Content-based specification features; Linguistic=Handcrafted linguistic features

D Distribution of Item Types Across Subsets

Item Type	Training	Development	Test	Total
EBSR	48	6	6	60
HT	162	20	21	203
MC	371	46	47	464
MI	12	1	2	15
MS	191	24	24	239
SA	92	12	12	116
WER	17	2	3	22
Total	893	111	115	1119

Table 7: Counts of Item Types by Subset.

E Fine-tuned LLM Model Architecture

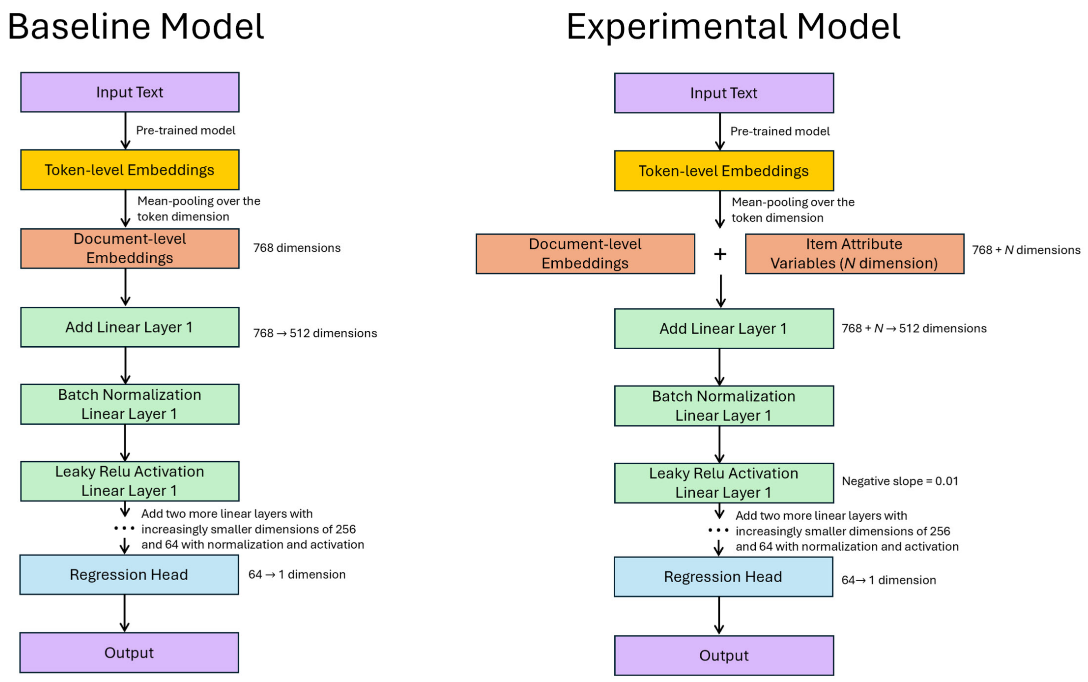


Figure 1: Fine-tuned LLM Model Architecture

F Item Difficulty Prediction Results on the Development Set

Approach	**Variant**	**Correlation**				**RMSE**			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.742	0.732	0.713	0.744	0.930	0.961	1.065	0.931
	F.A.	**0.762**	<u>0.777</u>	**0.777**	**0.785**	0.932	<u>0.886</u>	**0.884**	**0.882**
	T.L.	<u>0.759</u>	**0.792**	<u>0.752</u>	0.767	<u>0.924</u>	**0.867**	0.938	0.908
CatBoost	Baseline	0.692	0.692	0.692	0.692	1.008	1.008	1.008	1.008
	F.A.	0.729	0.728	0.749	<u>0.772</u>	0.953	0.970	<u>0.932</u>	<u>0.896</u>
	T.L.	0.754	0.742	0.741	0.751	**0.915**	0.944	0.952	0.927

Table 8: Item difficulty prediction results on the development set using fine-tuned LLM and CatBoost approaches across three model variants. Within each LLM, **bold** marks the best and <u>underline</u> marks the second best performance. F.A.=Feature Augmented; T.L.=Transfer Learning.

G Item Discrimination Prediction Results on the Development Set

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.258	0.216	0.190	0.188	0.187	0.192	0.204	0.250
	F.A.	0.374	0.241	0.292	0.371	<u>0.180</u>	0.186	0.191	<u>0.174</u>
	T.L.	<u>0.406</u>	<u>0.339</u>	<u>0.368</u>	<u>0.386</u>	0.183	<u>0.175</u>	0.316	0.178
CatBoost	Baseline	**0.447**	**0.447**	**0.447**	**0.447**	**0.167**	**0.167**	**0.167**	**0.167**
	F.A.	0.272	0.235	0.296	0.294	0.186	0.194	<u>0.190</u>	0.185
	T.L.	0.304	0.221	0.221	0.266	0.184	0.192	0.196	0.186

Table 9: Item discrimination prediction results on the development set using fine-tuned LLM and CatBoost approaches across three model variants. Within each LLM, **bold** marks the best and <u>underline</u> marks the second best performance. F.A.=Feature Augmented; T.L.=Transfer Learning.

How Model Size, Temperature, and Prompt Style Affect LLM-Human Assessment Score Alignment

Julie Jung[1], Max Lu *[1], Sina Chole Benker[2], Dogus Darici[2,3]
[1]Harvard Graduate School of Education, [2]Munster University,
[3]Institute of Anatomy and Neurobiology, University of Münster
* Joint First Authors
Correspondence: maxlu@fas.harvard.edu

Abstract

We examined how model size, temperature, and prompt style affect Large Language Models' (LLMs) alignment within itself, between models, and with human in assessing clinical reasoning skills. Model size emerged as a key factor in LLM-human score alignment. Study highlights the importance of checking alignments across multiple levels.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has introduced new possibilities for evaluating text-based and conversational assessments, offering a scalable and cost-effective alternative and complement to traditional human scoring. While earlier approaches relied on statistical models, recent studies have demonstrated that LLMs can produce accurate, consistent, and personalized evaluations, which are particularly valuable in settings with limited access to expert raters and a need for timely feedback (e.g., Pack et al., 2024; Xiao et al., 2024). In educational contexts, assessments leveraging automated scoring have enabled more frequent formative and summative evaluations by reducing teacher workload and accelerating feedback cycles (Bailey et al., 2025). These developments raise critical questions about how such technologies compare to current assessment practices and what implications they hold.

Before LLM-generated scores can be meaningfully deployed, it is essential to consider the impact of their opaque scoring processes and how different settings influence their behavior. Although LLMs may approximate human ratings at the surface level, there is often a misalignment between the cognitive processes humans rely on and the pattern-based inferences LLMs make (Baldwin et al., 2025). Unlike human raters who may draw on domain-specific reasoning, LLMs operate largely as "black boxes," making it difficult to trace how inputs lead to particular scoring outcomes (Bathaee, 2017). This lack of transparency may be problematic in domains where judgments need to be informed by specific expertise. Additionally, LLMs are not a single, homogeneous system. Different models behave differently, and even the same model can produce varying results depending on its settings. As a result, establishing a strong body of validity evidence is critical before integrating LLMs into assessment practices, particularly as a complement to human scoring.

In this study, we investigate how LLM-generated ratings, with varying model size, temperature, and prompt style, compare to current practices in a specific case study of clinical reasoning skill assessments in medical education. Clinical reasoning, the ability to recognize and interpret a patient's needs and health condition, is an essential skill for all health professionals (Tanner, 2006). It requires problem-solving abilities, experiential knowledge, and deliberate decision-making. Pattern recognition, in particular, is strongly linked to diagnostic success (Coderre et al., 2003), yet many medical students lack the clinical exposure needed to develop this skill. Thus, there is a need to provide students with opportunities to practice encounters with patients and receive structured, targeted feedback to develop their clinical reasoning skills (Haring et al., 2017). However, formative assessment of clinical reasoning is resource-intensive, typically requiring advanced medical experts to review and score student-patient dialogue. While LLMs have been used to simulate patients in such dialogues, further research is needed to determine whether they can be reliably used for scoring, and how different configurations of LLMs affect their scoring behavior and reliability (Brügge et al., 2024). This study explores how LLM-based evaluations align with or diverge from current human-based assessments and what those findings imply for the valid-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 265–273
October 27-29, 2025

ity, fairness, and practical use of LLMs in medical education.

2 Literature

LLMs have shown great promise in evaluating complex, language-based assessments. Recent studies comparing various LLMs have demonstrated that some models, such as ChatGPT using GPT4, have high alignment with human scorers for tasks such as essay assessment, although they tend to exhibit a tendency toward inflated scores (Seßler et al., 2025). LLMs are also capable of extracting nuanced insights from text, aligning reasonably well with human ratings in tasks such as discourse coherence analysis (Naismith et al., 2023) and evaluating short-answer assessments in an undergraduate medical program (Morjaria et al., 2024) with similar performance to human expert assessors. However, most studies have focused on scoring outcomes rather than examining the processes through which LLMs generate their ratings. Given the "black box" nature of LLM reasoning (Bathaee, 2017), further research is necessary to better understand how these models arrive at their judgments.

Although directly understanding the internal processes of LLM scoring remains challenging due to their nature, insights can be indirectly gained by systematically exploring how prompt style and model parameters influence scoring outcomes. Morjaria and colleagues (2024) found that the absence of a rubric in the prompt given to the LLM improved their alignments with human scoring, suggesting prompt style significantly shapes how these models evaluate responses. Yet, existing studies have largely overlooked how prompting the LLM to adopt a certain persona, such as a clinical expert, may further improve its alignment with human expert judgments by capturing nuanced reasoning processes underlying expert scoring. Additionally, model parameters, such as temperature settings, influence the randomness of and variability of generated responses and thus may affect scoring consistency and accuracy. Official guide on OpenAI suggests that lower temperature leads to more deterministic and less random outputs, while higher temperature leads to more random outputs (OpenAI, 2023). Many studies have linked the randomness to "creativity" or variability in the response (Peeperkorn et al., 2024). While the specific temperature setting for public-accessible ChatGPT is undisclosed, it is commonly believed to be around 0.7 (University of Victoria Libraries, 2024). Lastly, the choice of model type or size could also substantially influence scoring alignment due to variations in reasoning capabilities (Seßler et al., 2025). As psychometrician Andrew Ho emphasized, it is essential to ask, "Who is using what score for what purpose?" Different types of assessments, tailored to different use cases, demand varying standards for scoring, such as consistency and inter-rater reliability (Ho, 2025). Accordingly, as LLMs are increasingly employed as assessors, it becomes critical to systematically examine how variations in prompts, temperature settings, and model size affect their performance across diverse contexts. Such investigation is a necessary step toward the effective and trustworthy adoption of LLM-generated assessments.

When it comes to using LLMs to evaluate student performance for formative feedback, it is essential to understand how reliable are their ratings—how does LLM ratings compare itself to other LLMs and, importantly, to that of human—and what these comparisons imply for validity and practical use. Clinical reasoning provides a challenging test case in that it encompasses a complex cognitive process that relies on both formal and informal reasoning strategies, deeply informed by domain-specific knowledge (Simmons, 2010). It involves collecting patient information, evaluating its clinical significance, forming inferences, and generating diagnostic hypotheses. To facilitate formative assessment of these skills, the Clinical Reasoning Indicators-History Taking-Scale (CRI-HT-S) was developed, grounded in clinical reasoning indicators identified through qualitative research (Haring et al., 2017). This scale assesses dimensions such as a medical student's capacity to lead patient conversations and ask questions in a logical sequence. An initial validation has shown that the CRI-HT-S demonstrates acceptable reliability and internal consistency for assessing undergraduate medical students' clinical reasoning skills (Fürstenberg et al., 2020). Given the complexity inherent to clinical reasoning, it is particularly important to explore how LLMs evaluate this nuanced and multidimensional construct.

Despite growing interest in leveraging LLMs for formative evaluations, existing research has not systematically investigated how interactions between critical LLM parameters (model size, temperature, and prompt style) influence their scoring alignment with expert human assessors. Our study addresses this gap by comprehensively evaluating the impact of these parameters within clinical reasoning assessments, providing guidance for valid and practical implementation of LLMs in educational contexts. Specifically, we address the following research questions:

RQ1:How consistent are LLM raters when re-rating the same student in the clinical reasoning assessment?

RQ2: How do LLM design parameters (model size, temperature, and prompt style) affect the inter-rater reliability and alignment between LLM raters in the clinical reasoning assessment?

RQ3: How do LLM design parameters (model size, temperature, and prompt style) affect the inter-rater reliability and alignment with human raters in the clinical reasoning assessment?

RQ4: How do LLM design parameters (model size, temperature, and prompt style) and their interactions affect the average score levels in the clinical reasoning assessment compared to human raters?

3 Data and Sample

Our dataset consists of the transcripts of 21 third-year medical students in Germany who were each engaged in conversations with four fictional patients with differing diagnoses to assess the students' clinical reasoning skills. Two human raters with medical expertise rated each transcript with the Clinical Reasoning Indictors-History Taking-Scale, which consists of eight items measuring the quality of the medical student's clinical decision making on a Likert scale from 1 to 5 (see Appendix A). We systematically varied three LLM parameters: model size (GPT-4o as the large model vs. GPT-4o-mini as the small model), temperature (low: 0.2 vs. high: 0.7), and prompt style (regular vs. expert prompt; see Appendix A & B).These variations resulted in eight distinct LLM configurations. Each model rated every student dialogue, producing eight item-level scores per student. All transcripts were in German. One conversation (student 16 with patient 1) was excluded due to missing dialogue. Students were informed that their interactions would be assessed for clinical reasoning.

4 Methods

To address the first research question, each student's responses were rated twice by the LLM raters. Each student received scores on eight items across four rounds of conversation. For each rater, we averaged a student's item scores across all rounds. We then calculated pairwise intraclass correlation coefficients (ICCs) using the icc() function from the irr package in R (Gamer et al., 2019) across all raters. A two-way model was used to appropriately assess absolute agreement across the two trials of the same rater when students are considered random effects. According to the classical test theory, a high ICC suggests that the LLM tends to agree with its own ratings, a sign of reliability but not necessarily validity. However, a low ICC would suggest poor agreement of the same LLM across trials, a sign of a lack of reliability and validity. Conventionally, ICC below 0.5, between 0.5 and 0.75, and above 0.75 suggests poor, moderate, good or excellent reliability, respectively (Koo and Li, 2016).

Rater	Model	Temperature	Prompt Style
LLM1	Small (gpt-4o-mini)	Low (0.2)	Default
LLM2	Small (gpt-4o-mini)	Low (0.2)	Expert Persona
LLM3	Small (gpt-4o-mini)	High (0.7)	Default
LLM4	Small (gpt-4o-mini)	High (0.7)	Expert Persona
LLM5	Large (gpt-4o)	Low (0.2)	Default
LLM6	Large (gpt-4o)	Low (0.2)	Expert Persona
LLM7	Large (gpt-4o)	High (0.7)	Default
LLM8	Large (gpt-4o)	High (0.7)	Expert Persona

Table 1: Color-coded settings of LLM models by Rater ID. Background colors indicate model size and temperature, with red boxes around prompt style when set to "Pretend to be expert."

To answer the second and third research questions, we evaluated interrater agreement between LLMs and across human and LLM ratings. To facilitate pairwise comparisons between LLM and human raters, we also created a composite "averaged human" rater by averaging the two human scores. We used a two-way model to compute the ICC across all pairs of raters (8 LLM raters, 2 human raters, and 1 "averaged human" rater). Comparing ICC values between rater pairs allowed us to assess the degree of alignment between them. A high ICC between LLM and human indicate that the LLM rater may be scoring similarly to humans. Meanwhile, high ICC between LLM raters alone reflects consistency between LLMs but not necessarily alignment with human judgment.

To address the fourth research question, we further aggregated the item-level scores into a single person-level mean score per student per rater. This approach minimizes the influence of within-person variability and allows us to focus on between-student differences. However, this simplification

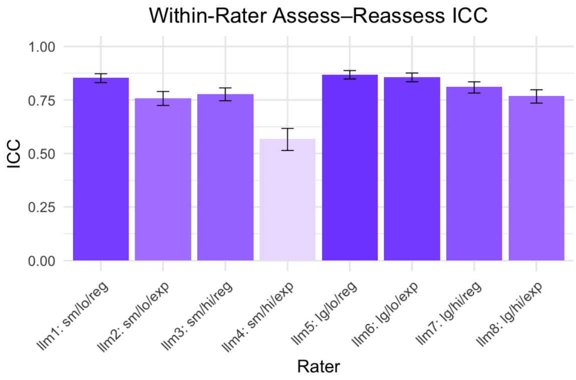


Figure 1: Bar plot of the intraclass correlations (ICCs) between two repeated assessments of the same student for each type of LLM rater. 95% confidence intervals are included.

assumes that the eight items reflect a unidimensional latent construct of performance, which is a limitation of this method.

To examine how different LLM design parameters affect LLM-generated ratings, we constructed a model including main effects and interactions among model size ($large$), temperature setting (hi_temp), and prompt style ($expert$) to reflect the eigh different LLM rater configurations. Each of these factors was coded as a binary indicator. For student i rated by rater j, we fit the following model using heteroskedasticity-consistent standard errors:

$$\begin{aligned} score_{ij} = {} & \beta_0 + \beta_1 large_j + \beta_2 hi\_temp_j + \beta_3 expert_j \\ & + \beta_4 large_j \cdot hi\_temp_j \\ & + \beta_5 large_j \cdot expert_j \\ & + \beta_6 hi\_temp_j \cdot expert_j \\ & + \beta_7 large_j \cdot hi\_temp_j \cdot expert_j + \varepsilon_{ij} \end{aligned} \tag{1}$$

This model estimates not only the average effects of each design parameter but also their two-way and three-way interaction effects, allowing us to explore whether the impact of one parameter depends on the levels of the others.

5 Results

For the within model consistency (RQ1), we compared the test–retest ICCs across LLM raters (see Figure 1). Most models showed high reliability (ICC≈.76–.87). The exception was GPT-4o mini with high temperature and the expert prompt, which exhibited lower consistency. In contrast, GPT-4o with low temperature and the regular prompt achieved the highest reliability (ICC = .87).

For consistency between models and between human and model, figure 2 shows the pairwise intraclass correlation coefficients (ICCs) among all raters. As a reference, the two human raters (r1 and r2) showed moderate agreement (ICC = 0.45). To answer RQ2, which focused on consistency between models, the four small-model LLM raters (llm1–llm4) showed good to very good internal agreement (ICC range: 0.60–0.81), and the large-model LLM raters (llm5–llm8) showed even stronger internal agreement (ICC range: 0.77–0.91).

However, further inspection revealed discrepancies between human and LLM ratings (RQ3). There appeared to be poor agreement between human raters and smaller LLMs (ICC < 0.20). In contrast, there appeared to be moderate agreement between humans and larger LLMs (ICC = 0.41–0.57), comparable to or slightly exceeding human–human agreement. These results indicated that while language models, regardless of size, could be highly agreeable with each other, this does not guarantee alignment with human evaluations. Therefore, if human ratings are considered the benchmark, alignment should be assessed based on direct agreement with human raters, rather than relying solely on internal consistency among the models.

Examining the ICCs across individual items to identify discrepancies between LLM-generated and human scores revealed interesting patterns. For example, item 4, assessing whether the student's questions suggested specific causes of symptoms, showed high consistency among LLM raters but low alignment between LLM-generated and human scores. This suggested that LLMs and human raters may interpret the criterion of "suggesting specific causes" differently, emphasizing the need for improved prompts that more effectively capture human evaluative reasoning (see Appendix C).

For RQ4, we estimated a linear model with robust standard errors to examine how model size, temperature, and prompt style, as well as their interactions, affected the average clinical reasoning scores produced by different LLM raters (see Appendix D). As shown in Figure 3, all LLM raters, regardless of configuration, produced higher average student scores than the human rater average (M = 3.19, SE = 0.054). The estimated intercept represented the expected score under the baseline

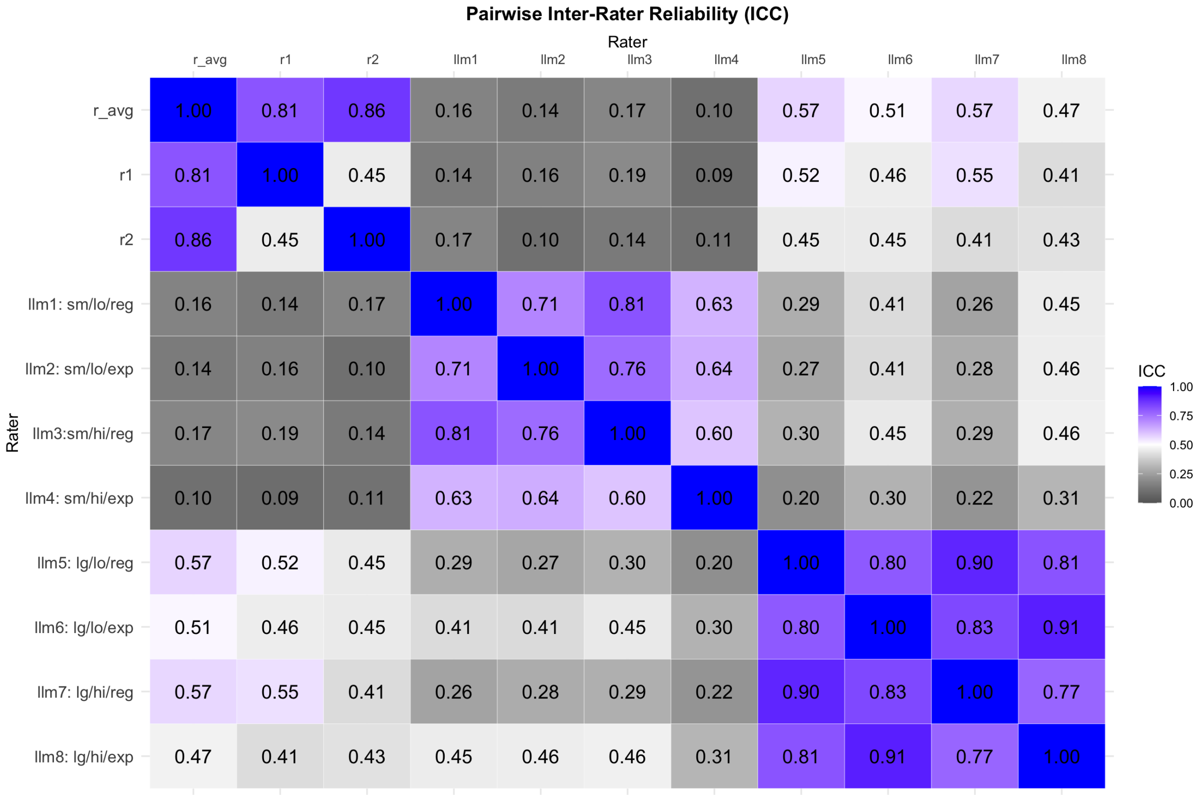


Figure 2: Pairwise intraclass correlation coefficients (ICC) between human and LLM raters across rating conditions.

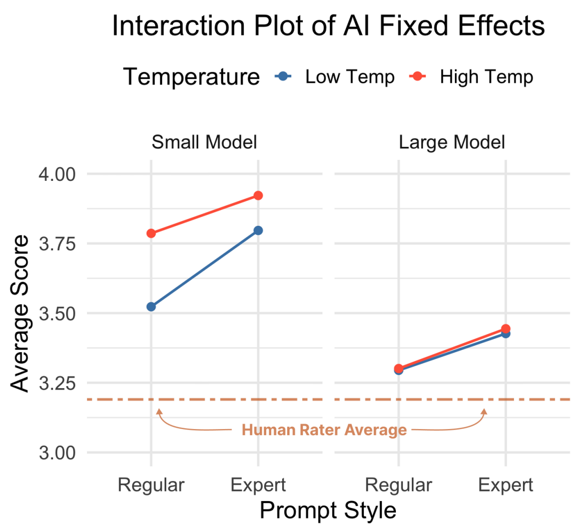


Figure 3: Interaction plot showing how model size, temperature, and prompt style influence AI-generated average scores compared to human rater average.

configuration (small model, low temperature, regular prompt) and was significantly above the human rater average (β = 3.52, p < .001), indicating that this model configuration was less stringent than the human raters.

Both high temperature and expert prompting were associated with significantly higher scores overall (β = 0.26 and β = 0.27, respectively; both p < .01). The interaction between model size and temperature was negative and statistically significant (β = –0.26, p < .05), suggesting that higher temperature inflated the score more for the small model than the large one, holding prompt style constant. On the other hand, the interaction between model size and expert prompting was nonsignificant (β = –0.14, p > .1), indicating that the effect of expert prompting on average scores did not significantly differ between the large and small models, holding temperature constant. The interaction among all three factors was also nonsignificant (β = 0.15, p > .1), suggesting that the combined effect of temperature and prompt style did not differ significantly between large and small models.

For the small model, expert prompt and high temperature each led to increases in average scores, up to almost 3.9. In contrast, the large model showed a smaller range of scores across conditions, around 3.3 to 3.4, regardless of prompt style or temperature. Notably, if the human rater average is taken as the gold standard, the configuration most closely aligned with human scoring norms was the large model with regular prompt and low temperature (M = 3.29, SE = 0.065).

6 Discussion

The purpose of this study was to explore how variations in model settings influence LLM-generated ratings and how these scores align within themselves, with each other, and with human scores, given the intended formative use of this assessment. We examined the impact of model size, temperature, and prompt style. Overall, the results showed that certain combinations of these parameters may be more suitable for specific assessment purposes. In evaluating medical students' clinical reasoning skills, findings revealed that model size influenced both alignment with human raters and internal consistency, with GPT-4o consistently outperforming GPT-4o-mini. Temperature and prompt style had a relatively minor effect when using the larger model compared to the smaller one. Thus, for formative assessment of clinical reasoning, employing a larger model such as GPT-4o is recommended to achieve greater consistency and human alignment.

Our results revealed that LLMs showed high ICC within the same model and between models of the same size, but can systematically differ from human ratings. Across almost all models, the assess–reassess reliability was fairly high, indicating fairly high reliability in reproducing their ratings across different trials. The four smaller models generally have ICC higher than 0.6 and the four larger models have ICC higher than 0.77, suggesting high internal consistency between models with the same size. However, when it comes to agreement with human, their performance vary. The smaller model (GPT-4o mini) showed very poor agreement with human raters, while the larger model showed ICC levels that are comparable to human-human ICC. Combining these insights highlight a key feature or limitation of LLM raters: high internal consistency within or across LLM raters does not imply alignment with human judgments. Simply deploying a group of LLM raters and observing agreement among them is insufficient for validating their use in scoring. Direct comparison with human ratings is necessary in those cases. However, once we are able to configure an LLM that has a high agreement with human, the high assess–reassess consistency is encouraging sign that such an LLM can produce reliable ratings.

In this study, although human raters themselves showed only moderate inter-rater reliability, the larger model aligned more closely with human scores than the smaller models did. Additionally, human raters gave lower average scores than the LLMs, suggesting that humans may apply stricter evaluation criteria - a pattern consistent with prior findings (Morjaria et al., 2024). Therefore, incorporating at least one human rater, or ensuring that LLM raters are demonstrably aligned with human evaluations, remains essential when human judgment serves as the gold standard.

Future implementations should consider periodic human validation of LLM-generated scores to promote fairness and reliability. Importantly, assessments should avoid assigning LLM raters to some students and human raters to others, as this could introduce systematic biases. Further research with larger sample sizes should also explore the use of Generalizability Theory to analyze the impact of rater variability across combinations of human and LLM raters (Shavelson et al., 1992).

Another notable finding was related to scores generated by LLMs using persona prompts. Contrary to our expectation that persona prompting (asking LLM to pretend to be an expert) would enhance alignment with human expert ratings, these prompts instead led to higher scores compared to standard prompts and showed poorer alignment with human ratings. This suggests that persona-based prompting, at least in this case, may not effectively replicate human expert evaluation processes. Future research could explore alternative prompting strategies, such as few-shot prompting, or adding a few examples within prompts, as a potentially more effective method to improve alignment between LLM-generated ratings and human ratings, particularly for items exhibiting notable discrepancies, such as Item 4.

Ultimately, advancing the use of LLMs for assessment requires careful attention not only to consistency within models, but also to variation between models and, most critically, to their alignment with human judgments, as overlooking any of these dimensions risks undermining the validity of LLM-generated scores and even leading to systematic biases.

Limitations

A major limitation of this study was the relatively small sample size. Having only two human raters restricted our ability to establish a robust human expert benchmark. This could have influenced the reliability comparisons with LLM-generated scores. Additional human raters would reduce measurement error. Moreover, with scores from only 21 medical students, we were unable to comprehensively explore or decompose sources of measurement error contributing to discrepancies between variations of LLM-generated and human ratings. Future studies should include larger datasets, enabling the application of Generalizability Theory to better identify and quantify multiple facets of error, such as variability due to raters, tasks, or specific scoring criteria.

Acknowledgments

We thank Prof. Andrew Dean Ho, Kenji Kitamura, and Mikayla Do for their valuable guidance and advice.

References

Alison L Bailey, Alexander Johnson, Natarajan Balaji Shankar, Hariram Veeramani, Julie A Washington, and Abeer Alwan. 2025. Addressing bias in spoken language systems used in the development and implementation of automated child language-based assessment. *Journal of Educational Measurement*.

Peter Baldwin, Victoria Yaneva, Kai North, Le An Ha, Yiyun Zhou, Alex J Mechaber, and Brian E Clauser. 2025. The vulnerability of ai-based scoring systems to gaming strategies: A case study. *Journal of Educational Measurement*, 62(1):172–194.

Yavar Bathaee. 2017. The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.*, 31:889.

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, 24(1):1391.

S Coderre, HHPH Mandin, Peter H Harasym, and Gordon H Fick. 2003. Diagnostic reasoning strategies and diagnostic success. *Medical education*, 37(8):695–703.

Sophie Fürstenberg, Tillmann Helm, Sarah Prediger, Martina Kadmon, Pascal O Berberat, and Sigrid Harendza. 2020. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, 20:1–7.

Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Catharina M Haring, Bernadette M Cools, Petra JM van Gurp, Jos WM van der Meer, and Cornelis T Postma. 2017. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. *BMC Medical education*, 17:1–9.

Andrew D. Ho. 2025. Metaphors, mantras, and mnemonics: Communication competencies in educational measurement. In *Presidential Address, Annual Meeting of the National Council on Measurement in Education*, Denver, CO. Presidential Address.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Leo Morjaria, Levi Burns, Keyna Bracken, Anthony J Levinson, Quang N Ngo, Mark Lee, and Matthew Sibbald. 2024. Examining the efficacy of chatgpt in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, 3(1):32–43.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

OpenAI. 2023. *Audio: Create translation*. Accessed: 2025-06-05.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Richard J Shavelson, Noreen M Webb, and Glenn L Rowley. 1992. *Generalizability theory*. American Psychological Association.

Barbara Simmons. 2010. Clinical reasoning: concept analysis. *Journal of advanced nursing*, 66(5):1151–1158.

Christine A Tanner. 2006. Thinking like a nurse: A research-based model of clinical judgment in nursing. *Journal of nursing education*, 45(6):204–211.

University of Victoria Libraries. 2024. Prompt design: Temperature parameter. https://libguides.uvic.ca/promptdesign/temp. Accessed: 2025-06-05.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv e-prints*, pages arXiv–2401.

A Regular prompt

The regular prompt we used is shown below:

Regular Prompt

At this point, you will assess the quality of the third-semester medical student conducting a history-taking conversation.
Your assessment should include the following eight criteria
1. Assess whether the user has taken control of the conversation to obtain the necessary information.
2. Assess whether the user recognizes all relevant information.
3. Assess whether the user formulates targeted questions so that he can capture and specify the symptoms in detail.
4. Assess whether the questions of the user suggest that specific causes or circumstances lead to certain symptoms.
5. Assess whether the user asks questions in a logical sequence.
6. Assess whether the user reassures the patient that he has received the correct information from the patient.
7. Assess whether the user has summarized his collected information before ending the conversation.
8. Assess whether the user has collected sufficient information of high quality at an appropriate speed.
Assign each of the eight criteria a score according to the following scheme:
1 - Does not meet the criterion 2 - Rather does not meet the criterion 3 - Partially meets the criterion 4 - Rather meets the criterion 5 - Fully meets the criterion

Explain the evaluation with two sentences.

B Expert persona prompt

The expert persona prompt we used is shown below:

Expert Persona Prompt

Clinical decision-making (CDM) is a central process in healthcare where physicians gather, evaluate, and interpret relevant information about a patient's health status to make informed decisions regarding diagnosis and treatment. At this point, act as a rater with medical expertise who is evaluating medical students for their CDM mastery. You will assess the quality of the third-semester medical student conducting a history-taking conversation. Provide outputs that a rater with medical expertise would create.
Your assessment should include the following eight criteria 1. Assess whether the user has taken control of the conversation to obtain the necessary information.
2. Assess whether the user recognizes all relevant information.
3. Assess whether the user formulates targeted questions so that he can capture and specify the symptoms in detail.
4. Assess whether the questions of the user suggest that specific causes or circumstances lead to certain symptoms.
5. Assess whether the user asks questions in a logical sequence.
6. Assess whether the user reassures the patient that he has received the correct information from the patient.
7. Assess whether the user has summarized his collected information before ending the conversation.
8. Assess whether the user has collected sufficient information of high quality at an appropriate speed.
Assign each of the eight criteria a score according to the following scheme:
1 - Does not meet the criterion 2 - Rather does not meet the criterion 3 - Partially meets the criterion 4 - Rather meets the criterion 5 - Fully meets the criterion

Explain the evaluation with two sentences.

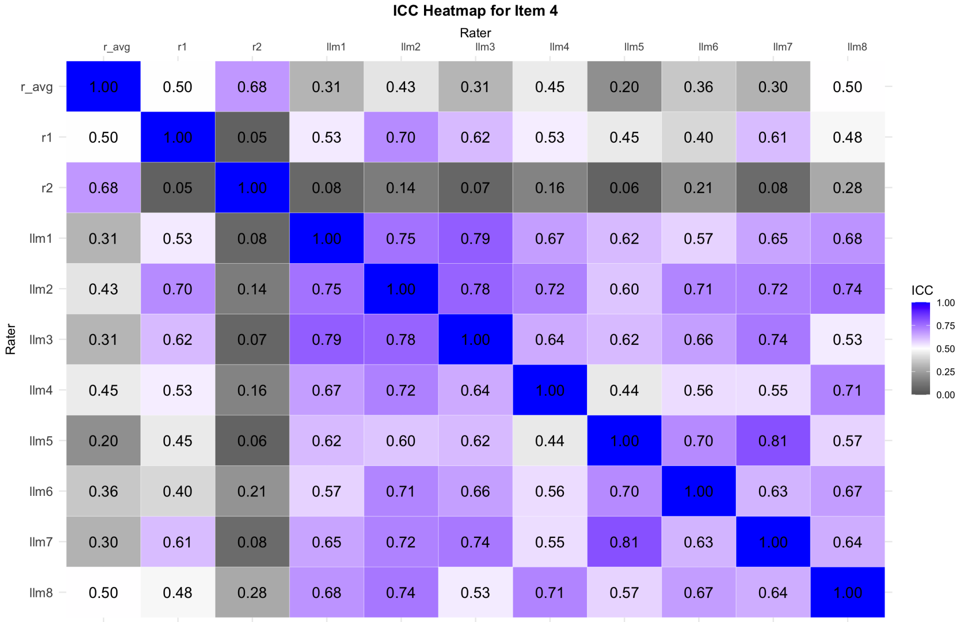


Figure 4: Pairwise Intraclass Correlation Coefficients (ICC) between human and LLM raters across rating conditions for Item 4

C Pairwise ICC for Item 4 (Figure 4)

D Main effects & interaction regression table (Table 2)

	Coefficient (SE)
(Intercept)	3.523*** (0.070)
Large Model	-0.228* (0.096)
High Temperature	0.263** (0.093)
Expert Prompt	0.274** (0.087)
Large Model:High Temperature	-0.257* (0.130)
Large Model:Expert Prompt	-0.142 (0.132)
High Temperature:Expert Prompt	-0.138 (0.121)
Large Model:High Temp.:Expert Prompt	0.149 (0.185)
Num. Obs.	189
R^2	0.293
Adj. R^2	0.265
AIC	137.3
BIC	166.5
RMSE	0.33

Table 2: Fixed effects regression on average clinical reasoning scores, with predictors for model size (Large Model), temperature (High Temperature), and prompt style (Expert Prompt), including interaction terms. Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Assessing AI skills: A washback point of view

Meirav Arieli-Attali[1*], **Beata Beigman Klebanov**[2*], **Tenaha O'Reilly**[2], **Diego Zapata-Rivera**[2], **Tami Sabag-Shushan**[3], and **Iman Awadie**[3]
[1]Fordham University, USA & One Assessment, Israel
[2]ETS Research Institute, USA
[3]National Authority for Testing and Evaluation, Israel

Abstract

The emerging dominance of AI in the community's perception of skills of the future makes assessing AI skills necessary to help guide learning. Creating an assessment of AI skills shares some challenges with other assessments but also poses new ones. We examine these from the point of view of washback and exemplify using two exploration studies conducted with 9th grade students.

1 Introduction

Washback is the impact of testing on curriculum, teaching, and learning. Positive washback occurs when practicing for the test results in wholesome learning. Negative washback occurs when preparing for the test results in a narrowing of the learning, such as the exclusion of important practices to focus on the skills targeted by the test or on test-taking strategies (Hughes, 1989).

In principle, the moment one makes choices regarding inclusion or exclusion of materials on a test, one sets things up for washback. Any assessment is going to focus on a particular construct – the skill that is the target of the assessment – and necessarily leave out other related skills. Thus, in the context of assessing writing proficiency, the choice of an essay task may have encouraged more attention to this genre; Burstein et al. (2014) found "an extraordinary prevalence of the essay" in USA K-12 context. Furthermore, once one designs a task – making decisions about timing, format, and scoring – there might be a further narrowing of the practice towards what the scoring rubric demands or implies. The debate about the artificial nature of a five-paragraph persuasive essay as a "test genre" is an example. The fact that the assessment task isn't identical to professional or vocational practice does not necessarily undermine validity; persuasive writing for a test, for example, shares important rhetorical elements with OpEd writing for the New York Times (Beigman Klebanov et al., 2019). The divergence between test and non-test contexts is nothing new, in itself. However, we believe assessment of AI skills comes with some new challenges.

2 Moving target

One challenge is specifying exactly what the target construct for the assessment is, within the general umbrella of AI skills. To articulate the construct, one would typically consult the relevant skills frameworks (such as OECD (2025) or UNESCO (2024)) and stakeholders (such as employers, educators, test takers) and model an assessment task on a common application of the target skill. Since AI tools are evolving, their typical and effective uses likewise continuously evolve. Moreover, the pace of the changes in AI tools and capabilities is such that it is likely to outpace the assessment development cycle of ideating, implementing, piloting, revising, and administering an assessment. Thus, the common tasks of today may change by tomorrow, so the assessment, when operationalized, would focus on the skills of yesterday and thus open up the possibility of negative washback, where preparing for the test would entail practicing an outdated use case of the AI technology. To address the challenge of assessing a 'moving target', one could (a) focus on fundamental elements that are less likely to change fast, such as understanding the implications of training on huge amounts of text data, and/or focus on enabling skills such as critical thinking; and (b) implement mechanisms to continuously review and revise the frameworks, constructs, and tasks, using the principles of evidence-centered design (Mislevy et al., 2003).

We note, in addition, that continuously updating the assessment to stay close to the current real-life

*Corresponding authors: mattali@fordham.edu and bbeigmanklebanov@ets.org

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 274–280
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

use comes with its own challenges. While student engagement and positive washback potential can thereby be improved, the very verisimilitude of the assessment may introduce into the performance variance that is authentic – it is part of real-life activities being mimicked – but might be construct-irrelevant from the point of view of the assessment. We will leave this issue for a more concrete discussion in the context of the exploration studies.

3 Broadening the construct

The second challenge is related to the consequences of selecting the focus of an assessment to be a particular set of AI skills. As explained, any selection would entail a de-selection of other skills that are not part of the target construct. To address this challenge, one could imagine a suite of assessments targeting various AI skills and rotation or sampling of the different assessments, as needed. However, picking specific AI skills does not only de-focus other AI skills. It could also de-focus skills that are not necessarily tied to AI. This is because much of the change brought about by AI is about using AI for tasks people did without AI before; with AI, these can be done faster and perhaps better; in any case, they are done differently.

Making the AI-way to perform a task the target of an assessment may de-focus the non-AI way of doing it. For example, using AI to help brainstorm an idea for a project is an increasingly common use of AI; yet people also think about project ideas themselves and brainstorm with other people. If a task on an assessment of AI skills asks students to brainstorm using AI, would that create washback potential whereby students will *have to* use AI for brainstorming to learn to do it more effectively and efficiently, at the expense of brainstorming with others or thinking creatively for themselves?

The issue of technology phasing out some human skills isn't a new consideration in the broader landscape of technological innovation. Weaving machines and calculators got integrated into the cultural fabric, largely displacing hand-weaving and calculating large sums on paper. What gives us pause is that the skills to be displaced might turn out to be those that are fundamental for the well-being of individuals and societies, such as creativity and thinking together with other people. Would it be wise, or, indeed, ethical, to impact the ongoing societal debate about the importance of skills like creativity by setting up an assessment that opens a possibility for washback against these skills, the element of choice in negative or positive washback notwithstanding?

Discussing assessment validity and washback in language testing, Messick (1996) argued that in order for washback, either positive or negative, to be tied to an assessment, one needs to show that it happened as a result of the assessment. Given the large variety of factors that go into curriculum design and educator choices of how to implement relevant learning and practice, Messick argued that it is more fruitful for assessment developers to consider issues in the assessment that could produce negative washback. Chief among these is construct under-representation. If, in order to succeed on an assessment, one needs to exercise only some of the skills that go into the real-world version of the target construct, the assessment under-represents the construct. Applying this reasoning to the discussion above, perhaps we should consider AI skills as part of broader constructs. Creativity can be exercised with AI or without AI; focusing on only one of these – either one – is likely to under-represent the construct in its current real-life use. Ergo, an assessment that includes brainstorming with AI should also include brainstorming without AI.

4 Exploration studies

In what follows, we exemplify the points raised above via two exploration studies we conducted with 9th grade students from public high schools in Israel; the students came from middle-high SES background in both studies. The first was conducted in June 2024, the second – in March 2025. The first study took place when students and teachers in Israel were only beginning to explore generative AI tools, most haven't tried them yet at all. By the time of the second study, all schools in Israel have introduced AI tools via an "AI month" – activities that included introducing to students several AI tools via specific tasks they needed to complete and submit. Teachers and students alike took part in these activities. The purpose was to show how AI can benefit education and to encourage teachers to implement AI in their teaching. Awareness of cyber security and AI's fake information and hallucinations were also part of this month's activities.

The first study included 10 students in one-on-one cognitive labs, where an experimenter observed a student working on the task. The goal was to explore how students react to and interact with the emerging generative AI. The second study included 72 students in 3 classrooms, working independently on a structured task on computers and smartphones. According to the regulations by the Israeli Department of Education, high school students (grades 9-12) are allowed to access the internet, including generative AI, directly, with parental consent and with a teacher's mediation. The task for the second study was developed based on the first study's insights and our evolving conceptual framework, inspired by the ECD (Mislevy et al., 2003). We implemented revisions to the framework in an iterative way following the "moving target" of the evolving capabilities of AI.

4.1 Study1 – June 2024

The task in this study consisted of three phases: (1) pre-task planning; (2) information gathering; and (3) preparing a 'product'. The task asked students to plan a 2-day class trip, guiding students through the necessary elements, e.g., choose a site, plan the arrival, choose or plan activities in the site's vicinity, find appropriate places to sleep nearby (hostel or camping) and eat (restaurants or takeouts). The 'product' students were asked to prepare was a trip brochure, one that can be published or sent out to the trip participants (their classmates).

The pre-task planning phase included planning verbally or in writing in front of the experimenter. For the information gathering phase, students were referred to ChatGPT (the experimenter created a free account for them) and asked to fact-check its responses using a search engine (e.g., Google). For the third phase of preparing the trip brochure, students were given Word or PowerPoint templates they could fill with pictures and the information they gathered. Students were told to use critical thinking and creative thinking, as well as to imagine that the trip they are planning could be a trip they take with their classmates. In other words, the task aimed to resemble an authentic use of AI tools and the internet to plan a trip, where it is necessary to verify the information given by the AI (e.g., correct names and details of sites or activities) and ascertain the feasibility of the plan (e.g., the distances between sites can be covered within the allocated time). Each student worked on the task while an experimenter sat beside them. The experimenter gave instructions at the start of the task and took the observer role with little interference unless needed, following the guidelines of cognitive labs in educational measurement (Arieli-Attali et al., 2023). The time allocation was 90 minutes.

4.2 Study 2 – March 2025

Based on our evolving Media & AI literacy framework and insights from Study 1, we designed a computerized scenario-based task, where students followed a storyline in which they were asked to help a tour guide plan a trip for a youth group. The task was structured such that the tour guide was the one who is planning the trip step-by-step, posing questions or needs in which he is requesting students' help in gathering the information for him. Thus, students were not asked to do the planning themselves nor did they have freedom in directing it; rather, they were requested to gather bits of information at each step to help the tour guide with his planning. Media & AI literacy items were incorporated as part of the information gathering process; critical and creative thinking items were incorporated as part of the storyline. Students had access to ChatGPT by opening a different tab on their device; this activity was not logged. The time allocated to the task was 90 minutes.

We now report on some insights from both studies to illustrate our main arguments above.

5 Discussion: Moving target

5.1 Changing AI capabilities

The target construct of AI literacy was composed primarily from the three previously well-researched constructs of (1) digital literacy; (2) media information literacy; and (3) critical thinking. One needs basic digital skills to operate digital tools on a computer or a smartphone in order to perform any AI literacy task. Media information literacy is needed not only in order to understand how and where to search for online information and identify its sources, but also to create, share or publish information to achieve one's goals. As information online is not always reliable, students need to apply their critical thinking skills in any online interaction, including when using generative AI.

As discussed above, one challenge is to define the skills of *today*, as AI capabilities change rapidly. Examining students' work in the two studies, less than one year apart, illustrated this point. In June 2024, the LLMs in Hebrew (e.g., ChatGPT, Gemini, Claude) were providing numerous fake details (or hallucinations) and make mistakes in phrasing in Hebrew. Thus, in the first study, we could focus on asking students to fact-check and edit the AI response, exhibiting their critical thinking skills. For example, some of the AI responses at that time included restaurants that do not exist, fake distances that would take more than eights hours' drive between the breakfast site and the lunch site (it takes less time than that to cross the country north-to-south or east-to-west), wrong details about the sites or the activities available at the sites. Less than a year later, the LLM responses were almost entirely accurate and very well phrased. Thus, while it is still the case that one would need to check the important details before embarking on the trip – in the manner of "measure twice, cut once" prudent planning – the editing trace of the final assessment product would be unlikely to contain substantial revisions. From the point of view of washback, it was no longer the case that students would truly grapple with fake information through this task, so using the assessment task to help set them on a course towards developing and practicing a critical attitude was no longer a viable option, at least not if one were using the publicly available AI tools as-is, without, for example, intentionally introducing incorrect information through engineering a prompt that would mediate between the student and the generative AI tools.

5.2 Challenges of approximating real-life use

Mirroring a real-life use case of AI can help set things up for positive washback through real-life applications of the practices encouraged by the assessment. Authenticity was thus a leading aspect of task development.

In the first study, the task was to plan a trip from scratch, with little guidance and only a few constraints. The task asked student to imagine that they are really going to invite their friends to this trip. As students were performing the task, our experimenters were watching them thinking out loud, and documenting their actions. Examining the trip brochures as the task "product" each student submitted and comparing those to the experimenters' protocols on student actions during task performance yielded the surprising insight that while it was evident that the task required and the students exhibited their critical and creative thinking skills in the process of planning the trip, the products were poor evidence of these processes. Some of the more critical and creative students ended up with a relatively poor brochure, due to poor digital or graphic skills or poor decision-making skills.

For example, one student's brochure was a paragraph describing the trip she planned, having no pictures, links, maps or any visuals or arguments that may persuade her friends to join the trip. In addition, some of the information was not fully accurate. The product would receive a low score. However, the experimenter protocol of that student revealed a thorough search, taking into account different conflicting considerations, and validating the information in many cases except one or two cases where she failed to do so; unfortunately, the latter found their way into the final brochure. Although the critical skills were not executed to the best, only the worst of them were evident in the outcome, masking all the other cases where they were executed correctly. It seemed that it was easier to find traces of mistakes in the final products rather than of correct conduct.

In addition, going through the protocols revealed that some students failed the task completely due to poor decision-making skills. They ended up hesitating and trying out different routes, and although they exhibited good media and AI skills and even good critical thinking – they finished the task without any product at all. Thus, while this aspect could suggest strong student engagement with the task and its authenticity in that a one-hour planning activity might not yield any plan that satisfies the traveller's standards, it introduced decision-making as one of the constructs assessed, which we judged to be outside of our desired assessment focus.

Finally, we found that the difficulty of the task depended very much on how students decided to approach it. Some students chose an "easy" trip, part of which was already stated in webpages of the chosen sites, while others chose to take a more challenging route of creating everything from scratch, trying to come up with their own creative combinations, some of which turned out to not be feasible at all. It was the case that those who chose the easier task performed better – in terms of the quality of the final product – than those who chose the harder task. Thus, the task itself was not comparable between students, creating an additional challenge from the point of view of scoring. Even putting this issue aside and examining the products themselves convinced us that creating a common scoring rubric would be extremely challenging due to variation across the submitted brochures.

The issue of variation in both process and product that comes with a closer approximation to real-life is not a unique challenge for assessing AI skills. However, the sheer extent of possibilities for variation may be a hallmark of real life in the era of fast-paced advances in AI technology. That is, the fact that one could quickly come up with, check, and discard a lot of different ideas is related to the strengths of generative AI in idea generation and in instant provision of a wealth of relevant information on almost any conceivable topic. Similarly, the possibility of a large variation in the quality of the visual designs of the brochures created *in mere minutes* may have come about due to the generative AI-induced amplification of differences in the student's independent design skills: It isn't only the visual artists among the students who could come up with visually compelling brochures in a matter of minutes, but also those students who could articulate the imagined designs in a textual prompt.

Based on the results of study 1, we developed a much more structured version of the task for the second study, so that students had less freedom in deciding on the type of trip and more opportunities to show clear evidence of their media and AI skills, as well as more direct focus on their critical and creative skills alongside their media and AI skills.

Scenario-based assessment is a promising paradigm for structured tasks where multiple aspects of a skill can be targeted through different elements of the scenario, thus potentially supporting both standardization and authenticity (Sabatini et al., 2020). In scenario-based assessment, the different discrete items, each targeting an aspect of the skill, are integrated into a thematically coherent whole, where the storyline resembles enough the real-world situation that it allows for representing more aspects of the target real-world skill.

In the second study, we designed a storyline where a tour guide needed to plan a trip for his youth group and is asking for assistance in the planning. The guide needs information which he asks the students to search for and verify. Although this task lacked some of the agency and authenticity of the original where students did the full planning themselves, this task included discrete items within the storyline aimed to elicit student critical thinking skills. For example, as part of the tour preparation, the guide is looking to create a post about the site the youth are going to visit, finding in social media a slogan stating a specific (incorrect) fact about the site. The tour guide then asks students whether he should share that slogan. This discrete item within the scenario elicits student critical thinking of fact-checking information before sharing.

In principle, an alternative to structuring the task could be to use not only the final product but also process data as a basis for assessment. For example, students could be asked to submit preliminary ideas, the LLM prompts and search queries they used, and/or reflections as they move through the task; the richer data could potentially support giving students credit for exhibiting the target thinking patterns even if the outcomes are not clearly reflected in the product. To allow for drawing evidence of skills from process data, one would need to clearly articulate the target construct and think through the extent to which fruitful thinking patterns in the context of the task can be reliably identified irrespective of the quality of the final product.

6 Discussion: Broadening the construct

As part of media-information literacy and critical thinking skill, an essential skill that was particularly identified as needed for AI literacy is "prompt engineering", that is, the ability to write to the LLM an appropriate request that will yield the desired response. Generally speaking, as students are more accurate and detailed in their request to the LLM, they may get a better response. Specifically, in the trip planning task, if the assessment focuses on the prompt engineering aspect of the task and lets the gen-AI do the planning, we might neglect the human – non-AI – skill of planning. We sought to learn about pre- and post–AI–use planning by designing a task where students first plan without AI and later plan with the aid of AI. We examined this issue in both studies.

In the first study, students were first asked to plan a blueprint of a class trip verbally while talking to the experimenter, and the experimenter documented what students said. This first stage was primarily aiming to elicit students' planning skill, while it also served as an engagement means to ease the transition to the AI task. The task instructions included several restrictions to the desired class trip so as to give some structure to the open-ended task, yet left it open enough to allow for students' planning. The instructions were: "You should plan a two-day class-trip for your class to a historic site in the vicinity of... ; you should find a place to spend the night (hostel or camping) and activities only for the first day. The activities should be appropriate for a group of students your age. You should plan for one morning activity and one afternoon activity, and lunch in-between. You should ignore for now budget or security considerations." After students finished telling the experimenter their plan and said that they are satisfied with it, they were then asked to open ChatGPT and ask it to detail or improve their blueprint. At this point they were facing the computer, and the experimenter was sitting behind them documenting their actions.

The main observation reported by the experimenters was that there was almost no connection between the initial blueprint plans students described at the first stage and the second part of the task. This was manifested in two main ways: (1) Students lacked good prompt engineering skills, that is, they made general requests from the LLM, ignoring what they already came up with; (2) The LLM response took a different route and students continued with the LLM route, forgetting their own. Thus, although students did invest creativity and effort in generating initial ideas, they either did not feel committed to these plans enough to pursue them with the help of LLM, or did not know how to do that and opted for the more generic suggestions by the LLM in response to a general prompt. We inferred from this experience that unless the students' initial ideas were elicited effectively and recorded to serve as part of the assessment data, the post-AI version is unlikely to reflect these ideas. This disjoint nature of the two brainstorming experiences is a challenge in designing a coherent scenario-based task that would cover both effectively.

In the second experiment, although the task was much more structured and required less overall planning, we did preserve the aspect of planning on a smaller scale. At some point it the task, the tour guide asks students to plan a one-hour activity around the theme of the site they're visiting. The students were given instructions (or constraints) about the activity – the theme of the site, the timeframe of the activity (one hour), the materials they have (ancient coins), and that the activity needs to be a group activity to suit a group of students of grade 9. After students submitted their planned activity, they were asked to open ChatGPT and now ask the LLM to detail or improve their plan, and the write down the AI response. There followed a question asking students to compare the AI activity plan to their own initial activity plan. The results largely replicated those of the first study – only 21% of the students actually made a thoughtful comparison, explaining in detail what AI added and why it was a better plan (14%) or why they decided to reject AI's elaboration and stay with their plan (7%). The rest of the students did not engage in the activity as intended: 30% said either AI's or their own plan was better, without explanation, while 45% said they did not know or did not answer the question at all. The remaining four students said that they used ChatGPT to help them come up with the original idea to begin with, therefore no comparison was necessary. While disengaged responses or disconnected own and AI plans dominated, we did obtain, from the 21% of the students, the intended behavior, where the two plans were connected and a meaningful evaluation and comparison were conducted. We are considering ways to encourage this behavior, both during test and in the form of washback – a learning activity where students can practice having their own ideas meaningfully elaborated by AI.

Limitations. In the current discussion, we exemplified assessment of AI skills in a stand-alone, non-disciplinary way. Additionally, the focus was on practical skills rather than on understanding how AI works or on AI ethics. We leave a discussion of contextualization and of ethics to future work.

7 Conclusion

Considering the emerging need to assess AI skills, we present some challenges related to fixing an AI-skills-focused construct to target in an assessment. One challenge is the rapid evolution of AI capabilities, which may lead to assessing today AI skills of yesterday; the other is the hazard of under-representing a broad, human-centered construct by focusing on the AI-reliant way to exercise the relevant skill. We illustrated via two exploration studies the need to revise and refine both the conceptual framework and the tasks themselves, in order to capture the changes in AI capabilities and AI practices and experiences.

We consider washback – the impact of testing on teaching and learning – to be an important motivation, keeping in mind that an assessment task can be used as a model by teachers to prepare students. Grounding both assessment and instruction of AI skills in common AI Literacy frameworks provides the first part of the bridge between assessment and instruction. Beyond this common ground, we have an opportunity to support positive washback by creating tasks that do not only provide good assessment data but have sufficient richness to capture a broad construct and enough authenticity to engage students in relevant practice – with the caveat that "relevant practice" in the age of AI might require frequent construct revision and updating.

References

Meirav Arieli-Attali, Irvin R Katz, and Gabrielle Cayton-Hodges. 2023. The many faces of cognitive labs in educational measurement. *ASK: Research and Methods*, 32(1):91–120.

Beata Beigman Klebanov, Chaitanya Ramineni, David Kaufer, Paul Yeoh, and Suguru Ishizaki. 2019. Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing*, 36(1):125–144.

Jill Burstein, Steven Holtzman, Jennifer Lentini, Hillary Molloy, Jane Shore, Jonathan Steinberg, Meg Vezzu, and N Elliiot. 2014. Genre research and automated writing evaluation: Using the lens of genre to understand exposure and readiness in teaching and assessing school and workplace writing. In *Annual Meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.*

Arthur Hughes. 1989. *Testing for language teachers*. Cambridge university press.

Samuel Messick. 1996. Validity and washback in language testing. *Language testing*, 13(3):241–256.

Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.

OECD. 2025. Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (review draft). Technical report, OECD, Paris.

John Sabatini, Tenaha O'Reilly, Jonathan Weeks, and Zuowei Wang. 2020. Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, 20(1):1–23.

UNESCO. 2024. AI competency framework for students. Technical report, UNESCO, France.

Using Generative AI to Develop a Common Metric in Item Response Theory

Peter Baldwin
Office of Research Strategy, National Board of Medical Examiners, Philadelphia, USA
pbaldwin@nbme.org

Abstract

Item response theory (IRT) models are subject to scale indeterminacy, causing parameters to be arbitrarily scaled. Consequently, parameters from independently calibrated test forms are not directly comparable without first estimating the linear transformation that aligns their respective scales. This paper introduces a novel procedure that uses large language models (LLMs) to estimate the transformation's slope and intercept. The method is evaluated using empirical data from a medical licensure exam. Results indicate that the LLM-based approach consistently recovers the slope across conditions, while intercept recovery is moderately sensitive to differences in average item difficulty between forms and improves as that difference narrows.

1 Introduction

When examinees take different test forms designed to measure the same trait, their scores must be adjusted for comparability. For number-correct scores (or transformations thereof), this process is called *equating*. In IRT, parameters are invariant up to a linear transformation; thus, while equating per se is unnecessary, a similar scaling adjustment is still required to ensure comparability across independently calibrated forms. After this adjustment, model parameters are expressed on a common scale—sometimes called developing a *common metric* (Stocking and Lord, 1983).

This scaling is necessary because the origin and unit of the latent scale must be arbitrarily fixed—directly or indirectly—to identify the model. Independently calibrated forms will therefore generally differ in scale, requiring a linear transformation before parameter estimates can be compared. This paper addresses this problem and proposes a procedure that uses GPT-based LLMs to estimate the slope and intercept of the required transformation. The method is illustrated using empirical data from the medical licensure domain.

2 Background

2.1 Problem Definition

Although scale indeterminacy affects all IRT models, we illustrate the issue using the two-parameter logistic model (2PL):

$$P(\theta) = \frac{1}{1 + e^{-Da(\theta - b)}}, \quad (1)$$

where $\theta \in \mathbb{R}$ denotes proficiency, $a \in \mathbb{R}_{>0}$ is item discrimination (equal to 4 times the item response function's (IRF) maximum slope), and $b \in \mathbb{R}$ is the item difficulty, the point on the difficulty/proficiency scale where the IRF inflects). $P(\theta)$ gives the probability of a correct response for an examinee with proficiency θ.[1]

A key feature of IRT is parameter invariance: item parameters are independent of examinee sample, and examinee proficiencies are independent of item set (Hambleton et al. 1991). However, this invariance holds only up to a linear

[1] Note: D is a scaling factor equal to $D = 1.702$ that allows the more mathematically tractable logistic formulation to closely resemble the normal ogive function, which preceded the logistic function in the historical development of IRT (Birnbaum, 1968).

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 281–289
October 27-29, 2025

transformation: for any slope j and intercept k, the transformation $a' = a/j$, $b' = bj + k$, and $\theta' = \theta j + k$ leaves $P(\theta)$ unchanged.

This scale indeterminacy is expected—these model parameters are not directly observable, requiring arbitrary scaling—but it complicates comparisons across independently calibrated test forms. Conventions exist for identifying IRT models (e.g., setting θ's mean and SD to 0 and 1, respectively) but they do not guarantee a shared scale across forms, since these constraints are applied separately to each. A linear transformation is still required. We denote its slope and intercept γ and η, respectively.

To estimate γ and η, something common across forms is needed (Baldwin and Clauser 2022), typically in the form of shared parameters (e.g., anchor items). These common parameters, being invariant up to a linear transformation, can be used to estimate the linear relationship between scales. Many well-known linking methods take this approach (Hambleton and Swaminathan 1985; Kolen and Brennan 2014). Absent common items, ancillary covariates that correlate with model parameters can sometimes be used (e.g., Mislevy et al. 1993; Wiberg and Bränberg 2015).

A single-group design, in which both forms are administered to the same examinees, allows estimation of the transformation constants via shared proficiencies. However, this approach is often infeasible due to examinee burden. More common is the *non-equivalent groups with anchor test* design. Although less demanding for examinees, it relies on item parameter invariance—an assumption that may be violated due to item exposure, evolving curricula, or changes in exam preparation, leading to *item parameter drift*.

To address these limitations, we propose using generative AI to create shared parameters across forms. Specifically, GPT-based LLMs are tasked with estimating item-level success probabilities for typical examinees from defined groups. These probabilities are used to derive a common set of synthetic proficiency parameters across forms—analogous to a single-group design—enabling estimation of the transformation constants without requiring common items, examinees, or external covariates.

The proposed approach is illustrated using empirical data from the medical licensure domain. It performed well, particularly for slope estimation, with high consistency across all conditions. Intercept estimates were more sensitive to differences in average item difficulty between forms, with smaller gaps yielding more accurate results.

2.2 Related Work

A review of the literature did not identify any studies that use LLMs directly to develop a common metric. However, several studies address related challenges, particularly item difficulty prediction—a long-standing topic in educational and psychological measurement (e.g., Beinborn et al., 2015; Huang et al., 2017; Ha and Yaneva, 2018). Current LLM-based approaches to difficulty prediction fall into two categories: (a) item-parameter prediction and (b) item-specific examinee-group performance prediction. The latter, while not identical to the task described here, is more closely aligned. Each approach is discussed below.

Item-parameter prediction estimates classical or IRT-based indices from item text. For example, Razavi and Powers (2025) used GPT-based models to predict difficulty for K–5 math and reading items. Their feature-based ensemble models outperformed direct rating methods, reaching correlations up to $r = 0.87$ with empirical difficulties. However, accuracy may decline in domains requiring specialized knowledge or complex reasoning. For instance, in a shared task using medical multiple-choice questions (MCQs), Yaneva et al. (2024) reported that difficulty estimation remains challenging in this domain.

The second approach—item-specific performance prediction—models how systems or subgroups perform on individual items. Studies have linked item difficulty for question-answering systems to human performance (e.g., Yaneva et al., 2019; Uto et al., 2024; Liu et al., 2025; Maeda, 2025), though not always with high precision. More relevant here are studies modeling interactions between examinee subgroups and items. Feng et al. (2025) used chain-of-thought prompting and synthetic response generation to predict MCQ difficulty for defined cohorts. Park et al. (2024) used AI models as proxies for students at different skill levels. While promising, such methods raise concerns about bias in synthetic responses and highlight the need for further validation.

3 Methodology

3.1 Proposed Procedure

Let $P_{g,m,i}$ denote the predicted probability that a typical examinee from group g will answer item i correctly, according to LLM m. Likewise, for test form f, let $\mathbf{P}_{g,m,f}$ be the vector of predicted probabilities for that set of items. Now, suppose an IRT model is fit to an empirical dataset for form f, yielding item parameter estimates. These estimates can then be used to estimate a proficiency value $\hat{\theta}_{g,m,f}$ that, in some way, best describes $\mathbf{P}_{g,m,f}$.

Because this process can be replicated multiple times, let $\hat{\theta}_{g,m,f,r}$ denote the estimate of $\theta_{g,m,f}$ associated with the rth such replication. To improve stability, we can then take the average across R replications:

$$\bar{\hat{\theta}}_{g,m,f} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{g,m,f,r}. \tag{2}$$

This average reduces the impact of sampling variability from LLM output and estimation noise.

Because IRT proficiencies are form-invariant, were $\theta_{g,m,f}$ to be estimated using two different test forms, the resulting two $\bar{\hat{\theta}}_{g,m,f}$ will be the same (excepting random error) *up to a linear transformation*:

$$\bar{\hat{\theta}}_{g,m,f_B} \approx \bar{\hat{\theta}}_{g,m,f_N}\gamma + \eta, \tag{3}$$

where the subscripts f_B and f_N, indicate base form or new form, respectively, and γ and η represent the slope and intercept of the transformation needed to place f_N's estimates on the scale of f_B.

Given G examinee groups and M LLMs, this procedure yields $G \times M$ mean estimated proficiencies, $\bar{\hat{\theta}}_{g,m,f}$, for each form. The transformation constants γ and η can then be estimated using the *mean and sigma* method (Marco 1977; Kolen and Brennan 2014), which matches the means and standard deviations of the two sets:

$$\hat{\gamma} = \frac{SD\left(\bar{\hat{\theta}}_{g,m,f_B}\right)}{SD\left(\bar{\hat{\theta}}_{g,m,f_N}\right)} \tag{4}$$

$$\hat{\eta} = \bar{\bar{\theta}}_{f_B} - \gamma\bar{\bar{\theta}}_{f_N}, \tag{5}$$

where

$$\bar{\bar{\theta}}_{f} = \frac{1}{GM}\sum_{g=1}^{G}\sum_{m=1}^{M}\bar{\hat{\theta}}_{g,m,f} \tag{6}$$

$$SD\left(\bar{\hat{\theta}}_{g,m,f}\right) = \sqrt{\frac{1}{GM}\sum_{g=1}^{G}\sum_{m=1}^{M}\left(\bar{\hat{\theta}}_{g,m,f} - \bar{\bar{\theta}}_{f}\right)^2}. \tag{7}$$

Notably, this application differs from traditional difficulty prediction in that there is no requirement that $\bar{\hat{\theta}}_{g,m,f}$ reflect the actual proficiency of a typical examinee in group g. That is, accuracy of $\bar{\hat{\theta}}_{g,m,f}$ is less important than form-invariance.

4 Experiments

4.1 Examinee Response Data

We evaluated the procedure using empirical data from the Step 2 exam of the United States Medical Licensing Examination (USMLE®) sequence. Step 2 is typically taken by medical students after their third year of medical school, following their core rotations, and consists of multiple simultaneously administered test forms, each with ~318 MCQs. This study used ~220 items (text-based and table-based) from a single form, along with responses from ~1,500 examinees.

Responses were modeled using the 2PL (Equation 1). Because all items came from a single form, item parameters were estimated on a common scale. Parameters were scaled such that proficiencies had a mean of 0 and SD of 1, simplifying interpretation.

4.2 LLM Data

For each MCQ, a prompt was generated instructing the LLM to act as "an expert medical education analyst" with "thorough knowledge of how medical students and residents perform on USMLE®-style multiple-choice questions." The LLM was then told: "You are tasked with predicting the performance of the typical examinee from each of five different examinee groups on the following USMLE® multiple-choice question." The prompt included the MCQ, its correct answer, exam label (Step 2), item type (e.g., "diagnosis"), and topic area (e.g., "cardio: infectious disorders"). The five examinee groups—first- through fourth-year medical students and first-year medical residents (PGY-1)—were listed, followed by the judgment task: "Think carefully (internally) about each group's level of training, typical preparedness, and likelihood of arriving at the correct answer… Factor in both knowledge and potential guessing... Provide one probability… for each of the five groups… [that] represents the probability that a typical examinee within that group will answer the question correctly."

For each of fifty replications, the prompt was submitted separately to three large language models (LLMs)—GPT-o1, GPT-o3, and GPT-4.1—via the OpenAI API (OpenAI, 2024). To ensure item security, we used private deployments of these models through Azure OpenAI. While this was the implementation used here, the procedure itself is model-agnostic.

In this way, $G=5$, $M=3$, and $R=50$, yields a set $5\times3\times50=750$ $P_{g,m,i}$ for each of the approximately 220 MCQs totaling approximately $5\times3\times50\times220=165{,}000$ predicted probabilities across the ~220 items.

4.3 Experiments

The full set of items was randomly divided into two equal-length artificial test forms: a *base form* and a *new form*. Form difficulty was manipulated as two study factors: (a) across-form difference in mean item difficulty and (b) across-form difference in the standard deviation of item difficulties. Each factor had 11 symmetrically spaced levels (from -0.25 to 0.25 in 0.05 increments), varied independently, resulting in 21 conditions (10 for mean differences, 10 for SD differences, and 1 baseline condition). An exploratory analysis using a fully crossed design found that slope estimates were largely insensitive to changes in mean item difficulty, and intercept estimates were similarly unaffected by changes in item difficulty spread. For this reason, rather than employing a fully crossed design, we explored only conditions in which exactly one parameter was varied at a time, while holding the other parameter fixed at zero.

While no specific proficiency estimation procedure is required, this study used a two-step empirical Bayes approach with Newton–Raphson optimization designed to ensure monotonic deviance reduction through step-size constraints and backtracking. Initial estimates were computed using diffuse (flat) priors. The empirical mean and standard deviation of these estimates then served as Gaussian priors in a refined second phase. At each iteration, values were optimized by minimizing deviance—the sum of the negative log-likelihood of LLM-predicted response probabilities and a Gaussian prior penalty. Newton updates were derived analytically from the 2PL model (Equation 1), using closed-form gradients and Hessians based on item parameters from the empirical dataset. Step sizes were clamped, and backtracking ensured monotonic deviance reduction. The estimation procedure terminated once convergence criteria (step sizes < 1×10^{-6}) were met.

Once $\hat{\theta}_{g,m,f,r}$ values were computed for all replications, they were averaged as described in Equation 2, yielding 15 values for each artificial test form. These proficiency estimates were then used to estimate the slope and intercept of the transformation line using Equations 4 and 5.

Using the same set of 165,000 predicted probabilities, the procedure—item assignment, proficiency estimation, and transformation recovery—was repeated multiple times. Because item assignments were randomized (within specified difficulty constraints), the resulting $\hat{\gamma}$ and $\hat{\eta}$ varied across repetitions. This variation reflects sensitivity to item selection, although it is smaller than would be expected had each repetition drawn from a new item pool. For this reason, the mean $\hat{\gamma}$ and $\hat{\eta}$ across repetitions were calculated for each of the 21 difference-in-form-difficulty conditions, and a new repetition—considered *typical*—was generated that produced $\hat{\gamma}$ and $\hat{\eta}$ within 0.005 of these means. These "typical" artificial test forms served as stable reference points for evaluating sampling variability more accurately via bootstrapping. One thousand bootstrap draws were created by independently sampling, with replacement, both LLM replications and items within each form. For each bootstrap draw, a $\hat{\gamma}_d$ and $\hat{\eta}_d$ were estimated. These bootstrap-estimated transformation constants were then used to approximate the sampling distributions of $\hat{\gamma}$ and $\hat{\eta}$.

4.4 Evaluation Criteria

Because all items were jointly calibrated, the "true" transformation function was the identity function: $\gamma = 1$ and $\eta = 0$. Accuracy was evaluated by the proximity of estimated values to these targets.

5 Results

5.1 Recovery of Transformation Function Slope

Figure 1 shows the estimated slopes as a function of the difference in item-difficulty standard deviations between forms. Also shown are the boundaries for the middle 95% of the distribution

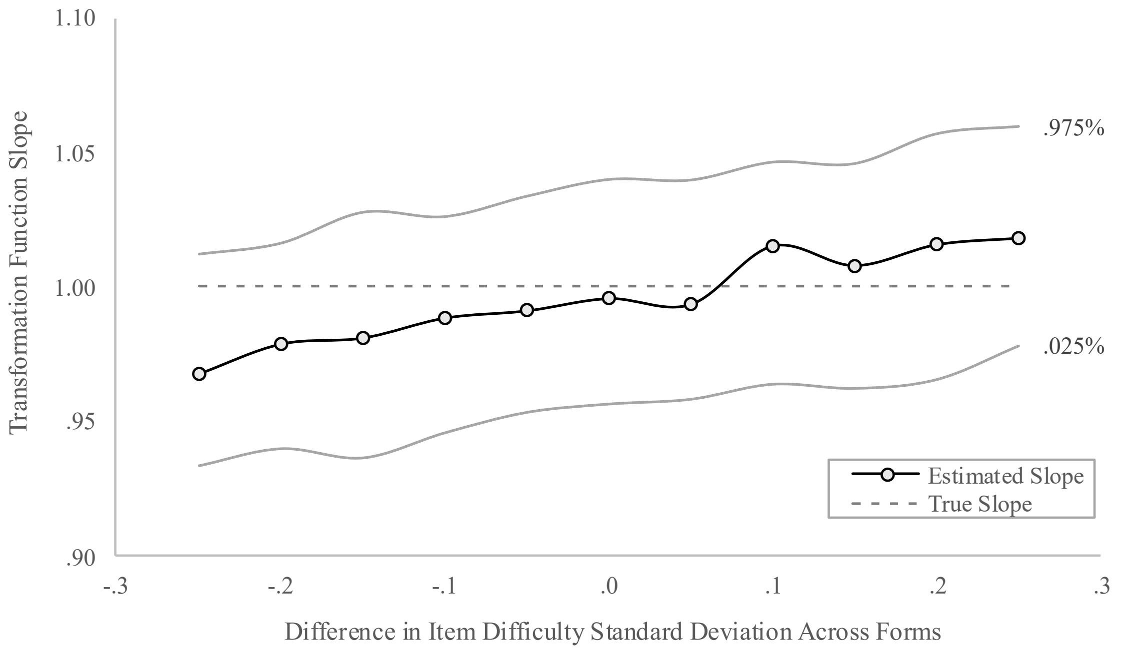


Figure 1: Estimated transformation function slope (black line) as a function of difference in item difficulty standard deviations. The boundaries for the middle 95% of bootstrap slopes are also given (light grey lines); the broken grey line shows the true slope.

Figure 2: Estimated transformation function intercept (black line) as a function of difference in mean item difficulty. The boundaries for the middle 95% of bootstrap intercepts are also given (light grey lines); the broken grey line shows the true slope. Note: the vertical axis scale spans .70 whereas for Figure 1, this axis spans only .20.

of bootstrap slopes ($\hat{\gamma}_d$; light grey) and the true slope ($\gamma = 1$; broken grey line). Across all conditions, the estimated slope deviates by no more than 0.03 from the true value. The estimates are most accurate when the difference in item-difficulty standard deviations between forms is close to zero. Likewise, the span of the 95% bootstrap sampling distribution is less than 0.09 for all conditions.

5.2 Recovery of Transformation Function Intercept

Figure 2 shows the estimated intercepts as a function of the difference in mean item difficulty between forms, following the same structure as Figure 1. The boundaries for the middle 95% of bootstrap intercepts ($\hat{\eta}_d$; light grey) and the true intercept ($\eta = 0$; broken grey line) are also shown. Intercept estimates improve as the difference in mean item difficulty across forms approaches zero. However, the vertical axis in Figure 2 spans 3.5 times the range of Figure 1, indicating greater variability. In the most extreme condition, the absolute difference between the true and estimated intercepts ($|\hat{\eta} - \eta|$) reaches 0.11. This difference does not fall below 0.05 until the across-form difference in mean item difficulty is $\leq$ 0.10. Similarly, the span associated with the middle 95% of the bootstrap intercepts is greater than that observed for the slopes—with spans up to 0.45.

6 Conclusion

6.1 Discussion

This paper describes a procedure for estimating the transformation constants required to place independently calibrated test forms on a common scale. It follows a single-group (common-person) design (Kolen and Brennan, 2014), but instead of using common examinees, proficiency estimates for a typical test taker from each of five predefined groups are used. These estimates are based on judgment tasks given to three LLMs: GPT-o1, GPT-o3, and GPT-4.1. The method was demonstrated using real examinee-response data from the USMLE® Step 2 exam.

The procedure recovered transformation-function slopes with high precision: across all conditions, slope estimates deviated from true values by no more than 0.03. Intercept estimates were more sensitive to model–data misfit and exhibited greater variability, particularly when mean form difficulties differed substantially. This likely reflects residual dependencies between the LLM-generated proficiency estimates and the item pool, undermining the assumption of conditional independence.

In IRT, proficiency parameters are item-set invariant. While proficiency *estimates* are never fully independent of the items used to derive them, the LLM-generated estimates in this study appeared especially sensitive to item characteristics. This suggests a degree of conditional dependence that may stem from misalignment between LLM-predicted probabilities and the modeled item response function. Because the success of the proposed procedure relies on form-invariant proficiency estimates, such dependencies likely contributed to the observed difficulties in recovering intercepts.

Developing a common metric across test forms administered at different points in time presents a challenge: *common items must exhibit invariance over time*. For testing programs where item parameter drift is a concern, this is a vexing problem. The procedure proposed here does not require items to have this property. Instead, it relies on LLMs to produce form-invariant proficiency estimates, and it is these estimates—rather than common item parameters—that are used to estimate the transformation constants needed to create a common scale across forms. If successful, the proposed procedure represents a considerably more secure method for maintaining a common scale over time. This will be especially attractive to testing programs that administer high-stakes tests following an episodic testing design.

6.2 Limitations

Although the procedure is not specific to any testing program, content domain, IRT model, or LLM, it was demonstrated using medical-domain items from the USMLE®, the 2PL IRT model, and three OpenAI LLMs. These design choices limit the generalizability of the findings.

The USMLE® assesses highly specialized technical content and is taken by a relatively homogeneous examinee population. It is unclear whether the findings extend to more general domains. However, previous studies have reported stronger performance for LLM-based predictions in broader content areas (e.g., Uto et al., 2024; Liu et al., 2025; Maeda, 2025; Razavi and Powers

2025), suggesting that the current results may underestimate the method's effectiveness in less technical contexts.

While the 2PL model is widely used, some testing programs—particularly in K–12 settings—prefer models like the 3PL, which incorporate additional complexity and assumptions about guessing behavior. Although the proposed procedure is not restricted to any single IRT model, it remains unclear whether LLM-based predictions align equally well under models other than the 2PL.

The OpenAI models used in this study are widely known, but they are neither the only nor necessarily the most effective LLMs for this task. Alternative models—used individually or in ensembles—may offer improved accuracy and consistency. Moreover, prediction quality is likely influenced by prompt phrasing and model settings. This study employed a fixed prompt and the default temperature, but future work should examine how variations in prompt structure and sampling parameters affect prediction accuracy and downstream performance. Finally, data security remains a critical concern. This study used private LLM deployments with no data logging or model training from inputs. However, not all models offer this level of protection—an important consideration for testing programs concerned with safeguarding test content.

Finally, the number of examinee groups and the number of items per form were chosen to suit the illustrative nature of this study. These design aspects may influence the quality of estimated transformation constants and the method's scalability. Other programs are likely to involve different group structures or item counts, and the procedure's performance under such conditions remains untested.

6.3 Future Work

Although the procedure performed well under most conditions, it may still fall short of the precision required for high-stakes applications, and several avenues for improvement remain.

First, the procedure is not limited to a fixed number of LLMs. Although this study used three widely known models, incorporating additional models—or employing ensemble strategies—may further improve the quality and stability of the proficiency estimates used to derive the transformation constants.

Similarly, although five examinee groups were used here, additional or alternative groupings could further enhance performance. Exploring optimal group configurations and model-specific strengths across subpopulations may yield more robust results. Increasing the number of examinee groups would also increase the number of independent estimates contributing to the transformation calculation, potentially improving precision.

The greatest limitation of the procedure lies in intercept estimation, specifically the bias in $\hat{\eta}$ when test forms differ in average difficulty. This issue may be mitigated through improved form design. For example, assembling forms to have closely matched mean difficulties can reduce the conditions under which intercept estimation becomes unstable. Importantly, precise individual difficulty estimates are not needed for this purpose; only mean difficulty must be controlled. This may be more tractable using existing difficulty-prediction methods.

Finally, alternative test assembly and delivery strategies could further improve the method's performance. For example, consider a scenario in which a large number of forms—comprising both unique and anchor items—are administered concurrently. These forms could be placed on a common metric using traditional IRT linking techniques (e.g., nonequivalent groups with anchor tests). Now suppose that multiple such administrations occur over time, as in an episodic testing design. In this case, large pools of administration-specific items, already placed on a common scale within each administration, could serve as input to the proposed procedure, yielding substantially larger item sets for estimating the required across-administration transformation. Future research should investigate such strategies, which focus on optimizing test design conditions rather than altering the procedure itself.

Acknowledgments

The author thanks the National Board of Medical Examiners for supporting this work.

References

Peter Baldwin and Brian E. Clauser. 2022. Historical perspectives on score comparability issues raised by innovations in testing. *Journal of Educational Measurement* 59(2):140–160. https://doi.org/10.1111/jedm.12318

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics* 2:517–530. https://doi.org/10.1162/tacl_a_00200

Allan Birnbaum. 1968. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479, Reading, MA. Addison-Wesley. https://ia601405.us.archive.org/32/items/in.ernet.dli.2015.139135/2015.139135.Statistical-Theories-Of-Mental-Test-Scores.pdf

Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew Lan. 2025. Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. *arXiv preprint* arXiv:2503.08551. https://arxiv.org/abs/2503.08551

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 389–398, New Orleans, Louisiana. Association for Computational Linguistics. https://aclanthology.org/W18-0548/

Ronald K. Hambleton and Hariharan Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Boston, MA. Kluwer-Nijhoff. https://link.springer.com/book/10.1007/978-94-017-1988-9

Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA. Sage Publications.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1):1352–1359. https://doi.org/10.1609/aaai.v31i1.10740

Radhika Kapoor, Sang T. Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W. Domingue. 2025. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint* arXiv:2502.20663. https://arxiv.org/abs/2502.20663

Michael J. Kolen and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices* (3rd ed.). New York, NY. Springer. https://link.springer.com/book/10.1007/978-1-4939-0317-7

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* 56(3):1028–1052. https://doi.org/10.1111/bjet.13570

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ. Lawrence Erlbaum Associates. https://www.routledge.com/Applications-of-Item-Response-Theory-To-Practical-Testing-Problems/Lord/p/book/9780898590067

Hotaka Maeda. 2025. Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement* 85(2):221–244. https://doi.org/10.1177/00131644241281053

Gary L. Marco. 1977. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement* 14(2):139–160. https://www.jstor.org/stable/1434012

Robert J. Mislevy, Kathleen M. Sheehan, and Marilyn S. Wingersky. 1993. How to equate tests with little or no data. *Journal of Educational Measurement* 30(1):55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

OpenAI. 2024. OpenAI API models documentation. https://platform.openai.com/docs/models

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.477/

Pooya Razavi and Sonya J. Powers. 2025. Estimating item difficulty using large language models and tree-based machine learning algorithms. *arXiv preprint* arXiv:2504.08804. https://arxiv.org/abs/2504.08804

Martha L. Stocking and Frederic M. Lord. 1983. Developing a common metric in item response theory. *Applied Psychological Measurement* 7(2):201–210. https://doi.org/10.1177/014662168300700208

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2024. Question difficulty prediction based on virtual test-takers and item response theory. In *CEUR Workshop Proceedings*, Vol. 3772. https://ceur-ws.org/Vol-3772/paper1.pdf

Marie Wiberg and Kenny Bränberg. 2015. Kernel equating under the nonequivalent groups with covariates design. *Applied Psychological Measurement* 39(5):349–361. https://doi.org/10.1177/0146621614567939

Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the difficulty of multiple-choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019)*, pages 11–20, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/W19-4402/

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.39/

Augmented Measurement Framework for Dynamic Validity and Reciprocal Human-AI Collaboration in Assessment

Taiwo Feyijimi[1], **Daniel Oyeniran**[2], **Olukayode Apata**[3], **Henry Makinde**[4], **Hope Adegoke**[4], **John Ajamobe**[3], **Justice Dadzie**[2],
[1] College of Engineering, University of Georgia,
[2]College of Education, University of Alabama,
[3]College of Education and Human Development, Texas A&M University,
[4]College of Education, University of North Carolina, Greensboro.
Correspondence: taiwo.feyijimi@uga.edu

Abstract

The swift penetration of Generative AI presents unprecedented opportunities and profound challenges for educational measurement, straining traditional validity and authenticity. This study introduces the Augmented Measurement Framework (AMF), a novel proactive conceptual model designed to transcend reactive responses by fostering synergistic human-AI collaboration. Grounded in four core principles, the AMF champions: Reciprocal Co-creation, for continuous human-AI learning and development within the assessment lifecycle; a Continuous Validity Ecosystem, ensuring dynamic measurement quality as AI evolves and contexts shift; Explainable Augmentation, enhancing human judgment through transparent AI insights and ethical deployment; and Pedagogical Resonance, aligning AI tools with sound learning outcomes and authentic assessment practices. This framework offers a paradigm shift for creating valid, fair, and pedagogically sound assessments that empower educators and learners alike. The paper explores its practical applications, policy implications, and charts a critical research agenda for advancing this essential framework.

1 Introduction

The educational landscape is currently navigating a period of transformation, largely catalyzed by the rapid advancements and increasing ubiquity of Artificial Intelligence (AI), particularly Generative Artificial Intelligence (GenAI). Tools such as ChatGPT and other sophisticated AI systems have evolved from simple text-generation utilities into powerful engines capable of producing complex, human-like content across a vast spectrum of domains (Khlaif et al., 2025). This technological surge offers immense potential for innovation in teaching, learning, and, critically, educational assessment. However, it simultaneously presents formidable challenges that shake the foundations of established pedagogical and measurement practices.

The core of the current predicament lies in the capacity of GenAI tools to enable students to produce outputs that may not genuinely reflect their own competencies, understanding, knowledge, and skills (Corbin et al., 2025). This capability directly confronts traditional assessment methods designed to ascertain individual student learning. AI's integration into academic practices is so profound that it blurs the line between acceptable assistance and inappropriate delegation at every stage of the assessment process, from initial ideation through to final editing (Corbin et al., 2025). The current situation has fostered a significant degree of uncertainty among educators and students alike, leading to what some describe as a sense of unease, even trauma, arising from the rapid erosion of established pedagogical norms and the challenge to professional identity, as they attempt to reconcile their educational values with the new realities imposed by AI technology (Corbin et al., 2025).

In response to these disruptions, numerous frameworks have emerged to help educators adapt, aiming to articulate boundaries between acceptable and unacceptable uses of AI in assessment (Corbin et al., 2025). Despite variety in the reasons for assessment redesign, most of the reasons are responsive focusing on developing "AI-resistant" assessments or implementing policies of simple prohibition or permission (Khlaif et al., 2025). However, these often reactive and incremental approaches frequently fall short in fostering a deeply integrated future, either limiting the transformative potential of AI or failing to address the fundamental redefinition of learning and assessment required in an AI-pervaded educational environment. This paper posits that the current juncture demands more than incremental adjustments or reactive measures. It calls for a proactive and theoretically grounded rethinking of the role of AI in educational mea-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 290–296
October 27-29, 2025

surement. The central purpose advanced here is the need for a new conceptual framework, the Augmented Measurement Framework (AMF). The AMF seeks to move beyond the "assessment crisis" by proposing a model where human intelligence and artificial intelligence work synergistically (Feyijimi et al., 2025) to create assessment processes that are more valid, fair, transparent, and ultimately, more conducive to authentic learning.

2 Conceptual Foundation

As educational institutions grapple with the integration of GenAI, various conceptual models have been proposed to guide practice and policy. One such model, the "Against, Avoid, Adopt, and Explore" framework, offers a pragmatic lens for faculty members to consider their responses to AI in assessment, reflecting differing levels of engagement and concern (Khlaif et al., 2025). Another significant contribution is the Human-Centric AI-First (HCAIF) framework, developed through extensive research and experimentation (Verhoeven & Hor, 2025). The HCAIF model is built upon five pillars: Preparation, Personalized Learning, Classroom Engagement, Summative Assessment, and Personalized Monitoring, and is underpinned by the crucial factors of Attribution (students showing how AI was used) and Reflection (students analyzing their AI use).

The challenges accompanying these adaptations are substantial. A primary concern is the difficulty in distinguishing between authentically student-produced work and AI-assisted or AI-generated submissions, a problem that strikes at the heart of academic integrity (Khlaif et al., 2025). Furthermore, the time and resource demands for redesigning assessments to be "AI-resistant" or to meaningfully incorporate AI are significant, often straining already burdened educators (Khlaif et al., 2025). Equity and accessibility concerns also loom large; disparities in student access to AI tools and digital infrastructure can create unfair assessment landscapes (Khlaif et al., 2025).

Compounding these issues are resistance to change from both faculty and students accustomed to traditional methods, and a pervasive lack of clear institutional guidelines and adequate professional development for educators (Khlaif et al., 2025). To understand how educational communities navigate these changes, particularly in defining acceptable AI use, concepts from sociology, such as boundary work and social boundary theory, offer valuable analytical tools (Corbin et al., 2025). These theories help explain how groups construct, maintain, and negotiate symbolic boundaries between legitimate and illegitimate practices, especially during periods of technological disruption (Corbin et al., 2025). To bridge this gap and cultivate the sophisticated human-AI partnership required, the concept of 'Augmented Intelligence' offers a powerful paradigm for envisioning the future of AI in education.

In an augmented intelligence model, the role of the teacher undergoes a significant transformation. Teachers shift from being the primary purveyors of knowledge to becoming facilitators, mentors, and guides who help students navigate personalized learning paths supported by AI (Chiu & Rospigliosi, 2025; Feyijimi et al., 2025). This shift allows educators to focus on higher-order skills such as critical thinking, ethical reasoning, creativity, and emotional intelligence (Feyijimi et al., 2025).

The integration of AI into educational measurement carries a profound ethical responsibility. As AI systems become more powerful and pervasive, ensuring their ethical development and deployment is not merely an adjunct consideration but a foundational requirement. There are numerous ethical concerns in AI driven assessment such as algorithmic bias (Feyijimi et al., 2025; Walden University, 2025), extensive data collection required (Bhadwal, 2024), academic integrity (Feyijimi et al., 2025; Khlaif et al., 2025) among others. A proactive approach to mitigating the ethical concerns goes beyond reducing harm but designing AI systems that embody and promote pedagogical values and equitable learning opportunities through the very act of measurement (Wynants et al., 2025). This proactive stance, which forms an intrinsic element of the proposed Augmented Measurement Framework, ensures that ethical considerations are not merely reactive safeguards but are foundational to the design and deployment of AI in assessment, fostering human flourishing and equity from inception. For AI to be ethically deployed and effectively integrated into educational measurement, its decision-making processes cannot remain opaque. Explainable AI (XAI) addresses this by referring to AI systems that can reveal how their decisions and recommendations are made, moving away from the "black box" paradigm where AI reasoning is hidden (European Commission, 2024; Feyijimi et al., 2025). The Defense Advanced Research Projects Agency (DARPA) defines XAI systems as those capable of explaining their reasoning, highlighting their strengths and limitations, and forecasting their future behavior (Gunasekara & Saarela, 2025). XAI is thus not merely a technical feature but a fundamental enabler of ethical AI, fostering trust, enabling informed human oversight, and supporting the pedagogical goals of fairness and understanding in AI-augmented educational measurement.

3 Dynamic Validity in AI-Augmented Assessment

Traditional conceptions of validity in educational measurement, while foundational, face significant challenges when applied to the rapidly evolving landscape of AI-augmented assessment. Classical validity theory, culminating in Messick's unified framework, encompasses various forms of evidence (e.g., content-related, construct-related, criterion-related) to support the interpretation and use of assessment scores for specific purposes (Aghazadeh et al., 2015; Messick, 1995). AI-driven assessments, particularly those employing machine learning, introduce several complexities that traditional validation approaches struggle to accommodate such as adaptivity, continuous learning, opacity and complexity of data.

To address these limitations, this paper proposes the concept of Dynamic Validity. Dynamic Validity is conceptualized as an ongoing, context-aware, and adaptive evaluation of an AI assessment system's inferences, decisions, and actions. It is not a one-time certification but a continuous process that monitors and seeks to ensure the system's ability to maintain measurement quality – including accuracy, fairness, relevance, and utility – as the AI model itself evolves, student populations change, instructional contexts shift, and our understanding of the assessed constructs develops.

Dynamic Validity is a multifaceted construct, encompassing several interconnected dimensions which include adaptive validity, continuous validity and ongoing assessment, dynamic fairness and algorithmic equity, explainability's role in dynamic validation and evolving consequential validity. These dimensions collectively underscore the necessity of adaptive mechanisms to re-evaluate validity as AI models evolve, the continuous monitoring of measurement quality for sustained reliability, the proactive pursuit of algorithmic equity to mitigate bias, the integration of explainability for transparent validation, and the iterative assessment of the broader educational and societal impacts of AI systems.

The concept of Dynamic Validity resonates with approaches in other fields that deal with rapidly evolving evidence bases such as Living Systematic Reviews (LSRs) in healthcare which ensures continuously updated review of new research findings and provide current and relevant evidence to support decision-making (Lansky & Wethington, 2020). Analogously, AI-augmented assessment systems require "Living Validity Arguments."

This implies establishing processes for periodic re-validation, transparent reporting of performance metrics, and mechanisms for incorporating new research findings about the AI's effectiveness and impact. Such an approach acknowledges that validity is not a fixed state to be achieved, but a continuous commitment to ensuring the responsible and effective use of AI in educational measurement. Indeed, Dynamic Validity is not merely a theoretical concept but a foundational pillar, intrinsically woven into the fabric of the Augmented Measurement Framework, ensuring that AI's evolution in assessment remains aligned with core educational values and robust measurement standards through constant vigilance.

4 Augmented Measurement Framework (AMF)

The AMF is proposed as a conceptual model (Figure 1) designed to guide the development, implementation, and ongoing evaluation of AI in educational assessment. It aims to foster a synergistic relationship between humans and AI, moving beyond simplistic views of AI as either a threat to be mitigated or a panacea for all assessment challenges (Feyijimi et al., 2025). The AMF is structured around four core, interconnected principles that collectively promote an assessment ecosystem that is valid, fair, transparent, pedagogically sound, and empowering for both educators and learners. This particular combination and emphasis on their synergistic interaction distinguishes the AMF, moving beyond singular focus areas to address the holistic challenges and opportunities of AI in educational measurement by envisioning a truly collaborative future. These principles are not discrete entities but rather deeply interdependent, forming a dynamic ecosystem where Reciprocal

Co-creation informs the ethical development of assessment tools, whose measurement quality is then continuously assured through the Continuous Validity Ecosystem, while Explainable Augmentation fosters trust and provides transparent insights, all serving to optimize Pedagogical Resonance with sound educational practices and desired learning outcomes.

The reciprocal co-creation principle of the AMF posits that educators and AI systems should engage in a continuous, bi-directional process of learning and development within the assessment lifecycle. It moves beyond the traditional paradigm where educators are passive end-users of pre-packaged AI tools or where AI functions merely as a static instrument. Mechanisms for enacting Reciprocal Co-creation include collaborative design from inception, AI-assisted, educator-refined content generation, nuanced feedback loops for AI improvement and interactive machine learning.

The continuous validity ecosystem principle asserts that validity in AI-augmented assessment is not a static property achieved at a single point in time, but an ongoing process embedded within an ecosystem of human oversight and AI adaptation. Mechanisms for establishing a Continuous Validity Ecosystem include systematic human review and oversight, AI-assisted prioritization of human review, ongoing performance monitoring and adaptation, living validity argument documentation and avoiding AI-only training cycles.

The explainable augmentation principle underscores that the primary role of AI in educational measurement is to augment and enhance – not replace – the professional judgment of educators. The principle can be achieved with the following process integration of XAI techniques, actionable insights for educators, supporting pedagogical diagnosis, and building trust through transparency. The final principle of pedagogical resonance proposed that AI assessment tools and processes must be subservient to, and supportive of, sound pedagogical principles and desired learning outcomes. Mechanisms for ensuring Pedagogical Resonance include alignment with learning theories, formative and actionable feedback, educator customization and control, emphasis on authentic assessment and holistic evaluation impact.

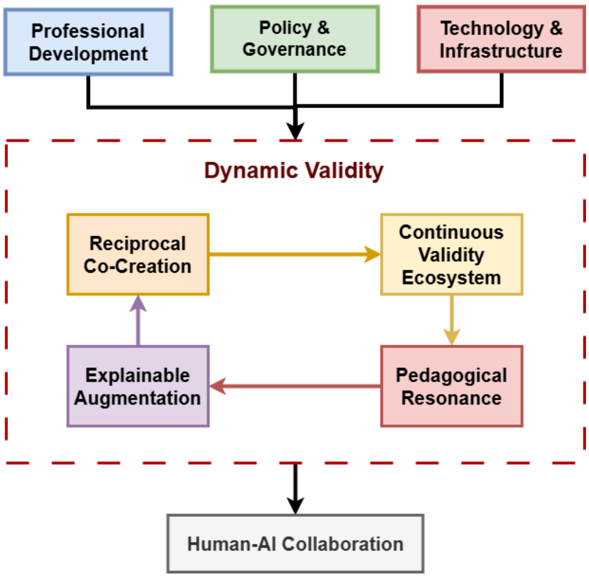


Figure 1: Augmented Measurement Framework. This figure visualizes the interconnectedness of the pillars and components within the AMF. Three enabling pillars: Professional Development, Policy & Governance, and Technology & Infrastructure feed into a Dynamic Validity envelope that continuously surrounds a four-step co-creative cycle consisting of Reciprocal Co-creation, Continuous Validity Ecosystem, Explainable Augmentation, and Pedagogical Resonance. This validity lens ensures that AI-supported assessments are fair, transparent, and aligned with learning goals. The cycle culminates in true human-AI collaboration, where educators and algorithms work together to produce valid, interpretable, and learner-centered results.

5 Operationalizing the Augmented Measurement Framework

The principles of the Augmented Measurement Framework (AMF) – Reciprocal Co-creation, Continuous Validity Ecosystem, Explainable Augmentation, and Pedagogical Resonance – provide a conceptual blueprint. Translating this blueprint into practice requires considering illustrative scenarios, implications for teacher professional development, necessary policy adjustments, and supportive technological enablers. Operationalizing the AMF is not about implementing a fixed set of tools, but fostering an evolving ecosystem where human expertise and AI capabilities synergize to improve educational measurement.

To concretize how the AMF might function, an illustrative scenario will be discussed in this proposal.

A consortium of STEM educators, representing diverse cultural backgrounds, collaborates with AI developers to create more culturally responsive as-

sessment items. The process begins with educators identifying potential biases (e.g., culturally specific contexts, language) in existing STEM assessment banks. The AI, equipped with NLP capabilities and access to diverse textual and cultural knowledge bases, assists in generating alternative assessment items or scenarios that are potentially more inclusive or relevant to different student populations (Martin et al., 2025). Educators then meticulously vet these AI-generated alternatives for cultural appropriateness, linguistic clarity, pedagogical soundness, and alignment with STEM learning objectives (Reciprocal Co-creation, Pedagogical Resonance). They might, for instance, adapt an AI-suggested physics problem to use examples or contexts familiar to students in their specific communities. The AI learns from these educator modifications and preferences, improving its ability to generate culturally sensitive content over time. Furthermore, the AI's Explainable AI (XAI) capabilities provide transparency into why certain items were flagged for bias or how alternative items were generated, offering educators actionable insights into the underlying patterns and reasoning, thereby enhancing trust and facilitating more informed collaborative decisions (Explainable Augmentation). The Continuous Validity Ecosystem involves ongoing monitoring of item performance across different demographic groups to ensure fairness and identify any emergent biases, with educators playing a key role in interpreting these data and guiding further AI refinement (Continuous Validity for fairness).

The successful implementation of the AMF heavily relies on equipping educators with new knowledge, skills, and dispositions. Professional development (PD) must evolve beyond basic operational training for specific AI tools. Key areas for PD include comprehensive AI literacy, data literacy for assessment, ethical AI use and bias mitigation, skills for human-AI collaboration and co-creation and pedagogical adaptability. Effective PD for the AMF is not just about transferring technical skills; it's about fostering a new "assessment mindset." Educators should see themselves as augmented decision-makers, critical evaluators of AI-generated insights, and active agents in shaping the AI assessment ecosystem, rather than passive recipients of AI directives or mere implementers of externally developed tools.

Supportive institutional and systemic policies are crucial for the AMF to take root. Achieving this requires ethical guidelines and data governance, investment in research and AMF aligned development, procurement and development standards, equity and access, and educator agency and time allocation. Policy frameworks supporting the AMF must strike a balance between fostering innovation and establishing robust safeguards.

The vision of the AMF relies on the availability and continuous improvement of certain technological capabilities such as user friendly XAI interfaces, platforms for collaborative designs and validation, robust and secure data infrastructure, interoperability standard and advanced AI models. Operationalizing the AMF is an ambitious but necessary endeavor. It requires a concerted effort from researchers, developers, educators, and policymakers to build the tools, practices, and supportive environments that can realize the promise of a truly augmented and human-centered approach to educational measurement.

6 Navigating Challenges and Charting Future Directions

While the Augmented Measurement Framework (AMF) offers a promising vision for the future of AI in educational assessment, its realization is not without significant challenges. Successfully navigating these hurdles and advancing the field requires a clear understanding of the practical, technical, ethical, and societal complexities involved, alongside a focused research agenda. However, the very design of the AMF, with its emphasis on Reciprocal Co-creation, a Continuous Validity Ecosystem, Explainable Augmentation, and Pedagogical Resonance, inherently provides mechanisms to proactively address many of these challenges, transforming potential obstacles into opportunities for innovation within a human-centered design.

The implementation of AMF-aligned AI assessment systems faces several practical and technical obstacles such as scalability of human oversight, data requirements and quality, computational costs and infrastructure, developing intuitive XAI interfaces and interoperability and integration.

Beyond technical issues, the ethical and societal implications of AI-augmented measurement demand careful and continuous attention such as mitigating algorithmic bias and ensuring fairness, protecting student data privacy and security, academic integrity in the age of GenAI, the digital divide and equitable access and over-reliance on AI and De-skiling (Feyijimi et al., 2025). Successfully navigating these multifaceted challenges hinges on fostering deep, interdisciplinary collaboration.

The AMF is a conceptual framework that requires empirical validation and refinement. A robust research agenda is needed to explore its principles and operationalization. There is need to develop and validate dynamic validity metrics, structure a longitudinal impact study, and investigating pedagogies for human-AI co-creation in assessment are some of the most crucial research agenda for achieving the goals of AMF. Among these, the immediate imperative lies in developing and empirically validating dynamic validity metrics, as these are fundamental to establishing the trustworthiness and utility of AMF-aligned systems in diverse educational contexts.

7 Conclusion

The integration of Artificial Intelligence into educational measurement stands at a critical juncture. The transformative power of AI, particularly GenAI, offers unprecedented opportunities to personalize learning, provide richer feedback, and gain deeper insights into student understanding. However, this potential is accompanied by significant challenges to traditional assessment paradigms, raising fundamental questions about validity, fairness, authenticity, and the very nature of human learning in an AI-pervaded world. This paper has argued that navigating this complex terrain requires a paradigm shift, a move away from viewing AI as a mere instrument of automation or efficiency, towards conceptualizing it as a collaborative partner in a dynamic and ethically grounded assessment ecosystem.

The Augmented Measurement Framework (AMF) has been proposed as a conceptual guide for this shift. Unlike prior reactive or incremental approaches, the AMF provides a proactive, holistic blueprint for integrating AI, championing a synergistic relationship where the unique strengths of human intelligence are weaved into the fabrics of AI-augmented assessment. Built upon four interconnected principles: Reciprocal Co-creation, Continuous Validity Ecosystem, Explainable Augmentation, and Pedagogical Resonance – the AMF offers a pathway to harness AI's capabilities responsibly. It champions a synergistic relationship where the unique strengths of human intelligence (pedagogical expertise, ethical judgment, contextual understanding) and artificial intelligence (data processing, pattern recognition, adaptive capabilities) are combined to create assessment processes that are more than the sum of their parts.

Operationalizing the AMF is undoubtedly a complex undertaking, fraught with technical, practical, and ethical challenges. It demands significant investment in research, the development of new technologies and professional competencies, and the establishment of supportive policy environments. However, the pursuit of such a framework is not merely a technical or academic exercise; it is an ethical imperative. The ultimate success of AI in educational measurement will not be judged solely by its technical sophistication or its efficiency gains. Instead, it will be measured by its capacity to foster more humane, equitable, and meaningful learning experiences for all students (Khlaif et al., 2025). AI must be a tool that helps to close achievement gaps, not widen them; that promotes critical thinking and creativity, not rote compliance; and that empowers educators, not diminishes their professional role.

Adopting frameworks like the AMF represents more than a technical upgrade for educational institutions; it signifies a cultural shift. It requires a collective commitment to ongoing learning, critical reflection, interdisciplinary collaboration, and adaptive governance of AI technologies. Realizing this vision demands a concerted, sustained effort from researchers, developers, educators, and policymakers alike, ensuring that the necessary tools, practices, and supportive environments are co-created to serve this profoundly human endeavor. The journey towards this future is ongoing, but by embracing principles of synergy, dynamic validity, transparency, and pedagogical integrity, we can strive to ensure that AI serves to make educational assessment not only more powerful but also more purposeful and profoundly human.

References

Aghazadeh, M. A., Jayaratna, I. S., Hung, A. J., Pan, M. M., Desai, M. M., Gill, I. S., & Goh, A. C. (2015). External validation of global evaluative assessment of robotic skills (GEARS). Surgical Endoscopy, 29, 3261–3266.

Bhadwal, A. (2024, September). Benefits and challenges of implementing AI in student assessment. Mera Tutor. https://www.meratutor.ai/blog/ai-in-student-assessment/, accessed on June 12, 2025.

Chiu, T. K., & Rospigliosi, P. A. (2025). Encouraging human-AI collaboration in interactive learning environments. Interactive Learning Environments, 33(2), 921–924.

Corbin, T., Dawson, P., Nicola–Richmond, K., & Partridge, H. (2025). ‘Where's the line? It's an absurd line’: towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education*, 1–13.

Feyijimi, T. R., Aliu, J. O., Oke, A. E., & Aghimien, D. O. (2025). ChatGPT's Expanding Horizons and Transformative Impact Across Domains: A Critical Review of Capabilities, Challenges, and Future Directions. Computers, 14(9), 366. https://doi.org/10.3390/computers14090366

Khlaif, Z. N., Alkouk, W. A., Salama, N., & Abu Eideh, B. (2025). Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. Education Sciences, 15(2), 174.

Lansky, A., & Wethington, H. R. (2020). Living systematic reviews and other approaches for updating evidence. American Journal of Public Health, 110(11), 1687.

Martin, A. F., Tubaltseva, S., Harrison, A., & Rubin, G. J. (2025). Participatory co-design and evaluation of a novel approach to generative AI-integrated coursework assessment in higher education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American psychologist, 50(9), 741.

Troussas, C., Krouska, A., & Sgouropoulou, C. (2025). The role of augmented intelligence and pedagogical theories in digital learning. In Human-computer interaction and augmented intelligence: The paradigm of interactive machine learning in educational software (pp. 39–93). Cham: Springer Nature Switzerland.

Verhoeven, B., & Hor, T. (2025, February). Higher education is undergoing a paradigm shift toward more human-centric, adaptive learning models that will equip students for an AI-powered future. Association to Advance Collegiate Schools of Business (AACSB). https://www.aacsb.edu/insights/articles/2025/02/a-framework-for-human-centric-ai-first-teaching, accessed on June 12, 2025.

Walden University. (2025). 5 pros and cons of AI in the education sector. What teachers need to know about artificial intelligence in the classroom. Available at https://www.waldenu.edu/programs/education/resource/five-pros-and-cons-of-ai-in-the-education-sector, accessed on June 12, 2025.

Wynants, S., Childers, G., De La Torre Roman, Y., Budar-Turner, D., & Vasquez, P. (2025). ETHICAL principles AI framework for higher education.

Patterns of Inquiry, Scaffolding, and Interaction Profiles in Learner-AI Collaborative Math Problem-Solving

Zilong Pan
Lehigh University

Shen Ba
The Education University of Hong Kong

Zilu Jiang
Johns Hopkins University

Chenglu Li
University of Utah

Abstract

This study investigates inquiry and scaffolding patterns between students and MathPal, a math AI agent, during problem-solving tasks. Using qualitative coding, lag sequential analysis, and Epistemic Network Analysis, the study identifies distinct interaction profiles, revealing how personalized AI feedback shapes student learning behaviors and inquiry dynamics in mathematics problem-solving activities.

1 Introduction & Background

As generative artificial intelligence (GAI) becomes increasingly integrated into K–12 settings, educators and researchers are starting to explore how GAI tools can meaningfully support student learning through interactive, adaptive assistant (Lang et al., 2025). Recent studies have shown that conversational AI tools can act as learning companions, offering personalized scaffolding and guiding students through complex tasks in real time, particularly in K–12 contexts (Kim & Kwon, 2025; Li et al., 2024). In mathematics education, where students often struggle with abstract reasoning and procedural complexity in the subject such as algebra, AI agents have the potential to provide personalized, real-time feedback that bridges the gap between instruction and independent inquiry.

The idea of instructional scaffolding has long emphasized the importance of timely support that is tailored to the learner's evolving needs (Belland, 2017). In AI-integrated learning environments, scaffolding can take the form of prompts, hints, motivational encouragement, or step-by-step guidance, each playing a crucial role in sustaining engagement and deepening understanding especially as students are solving math problems independently (Rittle-Johnson & Koedinger, 2005; Tay & Toh, 2023). Similarly, inquiry-based learning frameworks have highlighted how students' questioning behaviors, exploration strategies, and reflection processes contribute to meaningful learning (Artigue & Blomhøj, 2013). Yet, less is known about how these two dynamics—inquiry and scaffolding—manifest when students interact with AI agents in real-time problem-solving contexts.

This gap calls for a closer examination of how AI-mediated interactions shape learning processes in mathematics, particularly how students and ai agents collaboratively navigate complex problem-solving tasks. While previous studies have demonstrated the effectiveness of intelligent tutoring systems in enhancing performance and engagement (Niño-Rojas et al., 2024; Zhang & Jia, 2017), relatively few have examined the dialogic and adaptive qualities of these interactions—particularly from the dual perspectives of student-initiated inquiry and AI-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 297–305
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

generated scaffolding. Moreover, although some work has explored AI's responsiveness to student input (Atherton et al., 2024; Kim et al., 2025), there remains a lack of fine-grained analysis on how specific types of student inquiries elicit particular forms of scaffolding, and how these patterns vary across learners.

To address this gap, this study examines how high school students interact with MathPal (MathPal, 2023), an AI conversational agent designed to provide personalized support during math problem-solving tasks. Grounded in growth mindset and dialogic learning principles, MathPal engages students through structured yet adaptive conversation, offering hints, strategies, and encouragement as they work through algebraic problem sets. The integration of MathPal into real classroom settings offers a rich opportunity to explore the nature of AI-mediated learning interactions and to better understand how students navigate mathematical problem-solving tasks colligatively with an AI agent.

This research is guided by two key aims. First, researchers seek to analyze the patterns of inquiry and scaffolding interactions that emerge during student–MathPal problem-solving conversations. By classifying the types of questions students pose and the forms of support provided by AI, researchers aim to identify recurring patterns and interactional flows that characterize the learning dialogue. Second, this study intends to investigate how students differ in their interaction trajectories with MathPal, and whether distinct profiles of interaction can be identified across the classroom. Understanding these profiles may help educators and designers develop more adaptive AI systems that tailor scaffolding strategies to students' unique learning needs and inquiry styles, which lead to the following two research questions:

RQ1: What are the patterns of inquiry and scaffolding interactions between students and MathPal during math problem-solving activities?

RQ2: What distinct student interaction profiles with MathPal emerge during math problem-solving, and how do these profiles reflect differences in learning behaviors?

2 Methods

2.1 Participants

Participants were 48 ninth-grade students enrolled in three Algebra I classes at an urban high school located in the northeastern United States. The school serves a diverse student population with approximately 50% of students identified as economically disadvantaged. All participants were taught by the same high school mathematics teacher who has 12 years of teaching experience. Teacher consent, parental permission, and student assent were obtained in accordance with Institutional Review Board (IRB) approval prior to the commencement of the study and data collection.

2.2 Math AI Agent--MathPal

MathPal is an AI-powered conversational agent designed to serve as an interactive math learning partner for high school students. It supports students in understanding mathematical concepts and offers strategies and hints to guide them in solving practice problems. Informed by growth mindset concept (Dweck, 2006), which is the belief that abilities can be developed through effort, effective strategies, and constructive feedback, MathPal encourages students to persist through challenges and view mistakes as opportunities for learning.

MathPal can be integrated seamlessly with digital math learning platforms, providing students with real-time, responsive guidance in a conversational and supportive tone. Its design promotes resilience and confidence in problem-solving by offering personalized support aligned with students' learning needs. As shown in Figure 1, MathPal can read the screen and assist students in solving problems by providing strategies or hints without revealing the answer, using growth mindset–aligned language to facilitate problem-solving process.

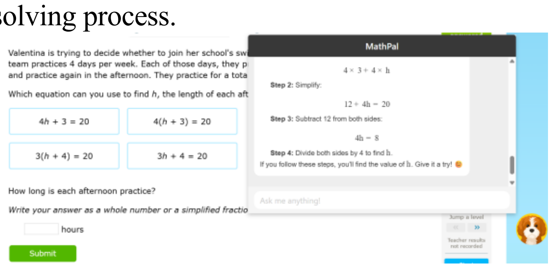


Figure 1: MathPal Conversation Interface

Additionally, MathPal promotes focus and motivation by redirecting off-topic discussions toward relevant mathematical content, linking students' personal interests to core math concepts. When students interact with MathPal, whether to ask about algebraic concepts or problem-solving steps and strategies, their inputs are interpreted by the system, routed through reliable back-end computational engines, and reformulated into accessible and student-friendly language.

2.3 Research Contexts

MathPal was integrated into the 9th-grade Algebra I curriculum and embedded within a virtual mathematics learning program. During daily math lessons, following the teacher's instruction, students transitioned to the virtual program for individual practice, typically engaging in math problem sets for approximately 15–20 minutes per class session. As students worked through various problem-solving tasks, they interacted with MathPal to receive guidance and tailored feedback. The types of tasks and instructional content supported by MathPal were aligned with daily instructional concepts such as solving linear function problems.

At the beginning of the semester, teachers provided an on-boarding session to students on installing and using MathPal, including demonstrations of its features and functions. Students then utilized MathPal throughout the semester (14 weeks) as a regular component of their mathematics learning experience (see Figure 2). During independent practice, teachers circulated around the classroom to address students' questions, whether related to the use of MathPal or the mathematical content itself. This support ensured that students remained engaged and could navigate the problem-solving tasks effectively.

Figure 2. Daily Math Lesson with MathPal

2.4 Data Analysis

Qualitative analysis was used to examine the conversational exchanges between students and MathPal during problem-solving activities. A total of 1,214 conversational threads were generated by the 48 participating students.

To better understand how students interact with MathPal during problem-solving activities, an inductive approach to qualitative analysis was used. Because engaging with an AI agent to solve problems collaboratively is a first-time experience for many students, the goal was to allow inquiry patterns to emerge directly from the data. Accordingly, open coding was first applied to identify recurring concepts and types of student input, followed by axial coding to organize these codes into broader themes and subthemes representing students' engagement with the tool (Strauss & Corbin, 1998). The researchers independently coded the first 300 threads of the data and developed separate codebooks. Through a process of comparison, contrast, and triangulation, they reconciled differences and collaboratively refined the codes into a single, unified codebook. This final codebook was then applied to complete the coding of the full dataset. Table 1 presents the codebook used to categorize students' inquiry types during problem-solving interactions. These were classified into five categories: solution-focused, conceptual, computational, formula/procedural, and clarification-seeking inquiries.

Theme	Definition	Example
Clarification Seeking	Students ask for repetition, elaboration, or clearer explanation of problem-solving process.	Can you go through the steps again?
Computational Inquiry	Students inquire about performing a specific mathematical computation.	I need to calculate cos(π/4) without a calculator.
Conceptual Inquiry	Students seek to understand the underlying meaning behind a mathematical concept.	What exactly is a derivative.
Formula/ Procedural Inquiry	Students ask for a formula, rule, or step-by-step procedure for solving a problem.	Can you provide the steps for computing the determinant.
Solution-Focused Inquiry	Students directly request help in solving a specific problem.	Can you help me solve this equation

Table 1. Codebook for Student Inquiry Types

To better understand how MathPal supported students in progressing through the problem-solving process, a deductive coding analysis (Fereday & Muir-Cochrane, 2006) was conducted on the AI-generated responses to student inquiries. The coding scheme was adapted from the STEM scaffolding types proposed by Belland (2017), which enables researchers to capture diverse forms of cognitive, metacognitive, and strategic support essential to inquiry-based learning, thus aligns well with the problem-solving context of this study. Researchers followed a similar coding and triangulation process as used in the student data analysis, collaboratively refining and consolidating the codebook through consensus (see Table 2).

Theme	Definition	Example
Conceptual Scaffolding	Provides explanations or definitions to help students understand underlying mathematical concepts.	A linear function is a function that…
Manage-ment Scaffolding	Help students stay on task or maintain focus during the problem-solving process.	Let's focus on solving the math problem!
Metacog-nitive Scaffolding	Encourages students to reflect on their thinking, or self-monitor their problem-solving process.	Let's go through this again…Ve-rify these to ensure correctness. Let me know how that works out.
Motivation-al Scaffolding	Offers encourage-ment or support to sustain students' confidence or engagement in the learning task.	I'll provide you with hints and guidance. Let's work together on some math problems!
Strategic Scaffolding	Guides students through a step-by-step process or strategy to solve a problem effectively.	To solve the problem of….…, follow these steps……

Table 2. Codebook for MathPal's Scaffolding Types

After completing the coding of both students' and MathPal's data based on the established codebook, data visualization and lag sequential analysis were conducted to address the first research question, which focused on the temporal patterns of inquiry types during problem-solving interactions.

To address the second research question, Epistemic Network Analysis (ENA; Shaffer, 2017) was employed to examine the co-occurrence and structural patterns of student–AI interactions. ENA utilizes a sliding window (also referred to as a "stanza") to analyze interaction data, enabling the identification of codes that co-occur frequently within a defined context. In this study, the unit of analysis for ENA was the individual student, with each student's interaction history with the AI defined as "conversation". For this analysis, the stanza size was set to seven, a standard parameter commonly used for modeling conversational data.

Following the initial ENA modeling, K-means clustering was applied to uncover distinct patterns of student-AI interaction. Using students' network projections on the ENA plane, the K-means algorithm was employed to determine the optimal number of clusters, ensuring that within-cluster similarity was maximized while between-cluster similarity was minimized. Subsequently, the clustering results were integrated into the ENA modeling to visualize and compare student-AI interactions across the identified clusters.

3 Results

3.1 RQ1: Inquiry and Scaffolding Patterns During Problem-Solving

To gain a comprehensive understanding of the general interaction patterns between students and MathPal, pie charts were created to illustrate the distribution of student inquiry types and MathPal's scaffolding strategies. Figure 3 reveals that Solution-Focused Inquiry accounted for the largest proportion of student utterances (42.29%), followed by Computational Inquiry (20.48%). This indicates that students primarily engaged with MathPal when tackling multi-step tasks or performing calculations. For example, a student asked, "how to do this problem step by step," which reflects a focus on procedural progression rather than conceptual exploration.

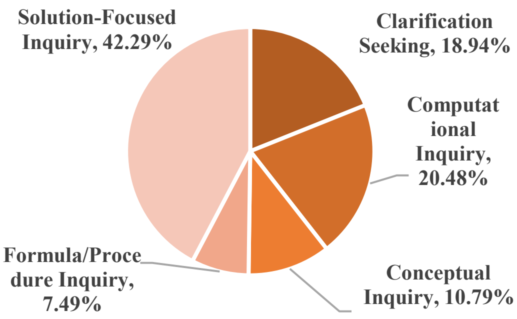


Figure 3: Student Inquiry Type Distribution

On MathPal's side (Figure 4), Strategic Scaffolding was the most frequently applied support type (52.88%), emphasizing the system's tendency to guide students through structured problem-solving steps. Together, the pie charts suggest that student-AI interactions were heavily oriented toward procedural assistance, with MathPal primarily functioning as a step-by-step guide rather than a conceptual tutor. This raises important pedagogical considerations for designing AI that not only supports task completion but also fosters deeper mathematical understanding.

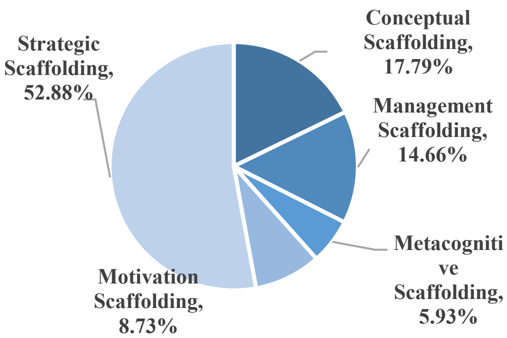


Figure 4: MathPal Scaffolding Type Distribution.

To further illuminate how specific types of student inquiries elicited corresponding scaffolding from MathPal, a Sankey diagram was constructed to visualize the directional flow of interactions between learner input and MathPal-generated scaffolding. This diagram captures the distribution of coded interactions across 454 episodes, illustrating how the nature of student queries on the left influenced the form of scaffolding provided on the right.

As indicated in figure 5, the dominant flow in the diagram stems from Solution-Focused Inquiry (n = 192), which overwhelmingly leads into Strategic Scaffolding (n = 300). This directional flow demonstrates that when students present complex problem-solving questions, such as those requiring multi-step reasoning, MathPal responded primarily with strategic procedural guidance. A typical case illustrates this flow: a student asked, "how to do this problem step by step", which was answered with: "To solve the equation -8d + 11d = 9d, follow these steps...". Such flows underscore the alignment between problem complexity and strategic decomposition in AI support, suggesting that MathPal's scaffolding logic is tightly calibrated to respond to problem-solving requests.

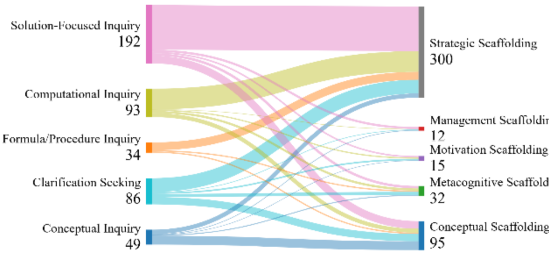


Figure 5: Student Inquiry Types and Corresponding MathPal Scaffolding Support

Computational Inquiries (n = 93) also predominantly flowed toward Strategic Scaffolding. The directional thickness of this stream in the Sankey diagram reflects frequent AI responses aimed at operational clarity. For example, queries like "put 10, 20,15, 30... in a number line" prompted responses emphasizing ordered thinking, reinforcing the trend that MathPal prioritizes task structure and planning in response to numerical and operational queries.

Another substantial pathway leads from Clarification Seeking (n = 86) and Conceptual Inquiry (n = 49) into Conceptual Scaffolding (n = 95). These flows signify situations where learners express epistemic uncertainty, often about definitions, categories, or relational understanding, and the AI responded with conceptual elaboration. This pattern affirms that conceptual depth on the part of the student correlates with knowledge-level scaffolding from MathPal, further evidencing a pedagogically aligned flow system.

In contrast, weaker flows were observed toward Management (n = 12), Motivational (n = 15), and Metacognitive Scaffolding (n = 32). These less prominent flows appeared thin across all inquiry types, indicating either low incidence of such student needs or limited detection and activation by MathPal. Notably, even from Solution-Focused Inquiries, which is the most dominant source, only a small number of instances flowed into these support categories.

The visual marginality of these flows in the Sankey chart, particularly those leading to motivational, management, and metacognitive scaffolding, points to instructional dynamics where motivational and self-regulatory supports are engaged less frequently in AI-student interactions.

In summary, the Sankey diagram functionally maps the dialogic flow from student-generated inquiries to AI-generated scaffolds, offering a visual representation of how MathPal interprets and responds to varying learner needs. The flow patterns reveal a predominant reliance on strategic and conceptual scaffolding, primarily triggered by Solution-Focused and Clarification-Seeking inquiries, respectively. The directionality and volume of these flows underscore the AI's responsiveness to cognitive and procedural challenges. From an instructional perspective, these patterns invite further consideration of how scaffolding models might be expanded to proactively integrate motivational and metacognitive supports, particularly in contexts where student engagement, persistence, or reflection could enhance learning outcomes.

While the Sankey diagram visualizes aggregate inquiry–scaffolding flows, it does not test whether specific patterns are statistically significant. To investigate these patterns systematically, we conducted a Lag Sequential Analysis to examine whether specific types of student inquiries consistently elicited particular AI scaffolds in a statistically meaningful way. This analysis examined the extent to which specific student inquiry types were predictably followed by particular MathPal scaffolding beyond what would occur by chance. By statistically modeling these turn-by-turn sequences, we were able to identify significant dialogic contingencies that provide deeper insight into the dynamics of student–AI interaction during problem-solving.

The results from lag-sequential analysis revealed statistically significant patterns of interaction between students and MathPal during high school mathematics problem-solving activities. As summarized in Table 3, each student inquiry type was followed by distinct forms of AI scaffolding at rates significantly higher than expected by chance ($z > 1.96$, $p < .05$). For example, Problem-Solving Inquiries were most frequently followed by Strategic Scaffolding ($z = 10.54$) and Management Scaffolding ($z = 10.48$), suggesting that MathPal responds to action-oriented problem-solving attempts by providing procedural guidance and task organization support—both of which are essential for success in secondary-level mathematics.

Inquiries focused on formulas and procedures led to similarly strong responses, particularly Strategic Scaffolding ($z = 9.38$) and Metacognitive Scaffolding ($z = 6.71$), reflecting MathPal's tendency to blend step-by-step guidance with prompts for reflection. Conceptual Inquiries elicited Conceptual Scaffolding ($z = 5.93$), indicating that MathPal adjusts to cognitively deeper questions with explanation-focused support. Additionally, Clarification-Seeking moves triggered a broad range of responses, including Conceptual ($z = 4.40$), Metacognitive ($z = 4.05$), Strategic ($z = 4.00$), and Motivational Scaffolding ($z = 3.77$), showing that the AI detects confusion as an opportunity for multi-dimensional support.

Overall, the results highlight MathPal's capacity to align its scaffolding strategies with the nature of student input, supporting the notion of contingent scaffolding in AI-driven learning environments, particularly within high school mathematics problem-solving contexts.

Student Inquiry	**MathPal Scaffolding**	**Z-Score**
Solution-Focused Inquiry	Strategic Scaffolding	10.54
Solution-Focused Inquiry	Management Scaffolding	10.48
Formula/Procedure Inquiry	Strategic Scaffolding	9.38
Formula/Procedure Inquiry	Metacognitive Scaffolding	6.71
Conceptual Inquiry	Conceptual Scaffolding	5.93
Computational Inquiry	Strategic Scaffolding	4.99
Computational Inquiry	Metacognitive Scaffolding	4.62
Computational Inquiry	Motivation Scaffolding	4.44
Clarification Seeking	Conceptual Scaffolding	4.40
Clarification Seeking	Metacognitive Scaffolding	4.05
Clarification Seeking	Strategic Scaffolding	4.00
Clarification Seeking	Motivation Scaffolding	3.77

Table 3: Significant Transitions Between Student Inquiry Types and MathPal Scaffolding

3.2 RQ2: Student-AI Collaborative Problem-Solving Profiles

While lag sequential analysis uncovered statistically significant turn-level patterns between inquiry types and scaffolding responses, it does not reveal how these interaction patterns accumulate across learners. To address this, we employed Epistemic Network Analysis (ENA) to model individual students' inquiry–scaffold co-occurrence structures and identify latent profiles of interaction. ENA diagrams were developed to better understand students-MathPal collaboration patterns during problem-solving activities. As shown in Figure 6, the dots represent the projections of individual student networks , while the square markers indicate the average projections of students within each cluster. The spatial distances within the ENA diagram reflect differences in network structures, where closer dots indicate more similar interaction patterns, and greater distances signify more distinct structures. ENA analysis revealed four distinct interaction clusters, corresponding to different patterns of student-MathPal interactions.

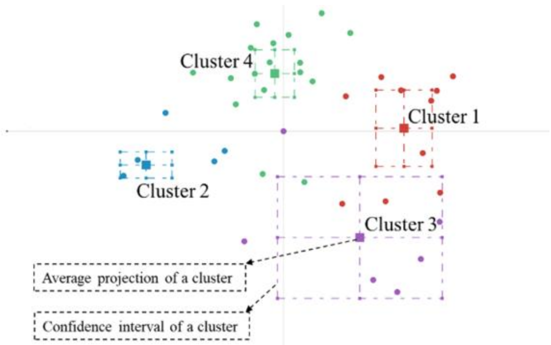


Figure 6: Distributions of four clusters of student-AI interactions

The specific connection patterns within each cluster, visualized in Figure 7, further illustrate these distinctions. In Cluster 1, the strongest connection was observed between computational inquiry and strategic scaffolding. Meanwhile, there were several moderate connections such as clarification seeking-strategic scaffolding and solution-focused inquiry-strategic scaffolding. Comparatively, the most prominent connection in Cluster 2 was between solution-focused inquiry and strategic scaffolding, suggesting a goal-driven interaction pattern. In contrast, Cluster 3 exhibited more balanced connections overall, with a slightly stronger connection between strategic scaffolding and motivation scaffolding. Lastly, Cluster 4 featured prominent connections among solution-focused inquiry, computational inquiry, and strategic scaffolding, as well as a moderate connection between solution-focused inquiry and conceptual scaffolding, suggesting a blend of conceptual understanding and problem-solving strategies.

Each cluster reflects an integrated pattern of how students regulate their interactions with MathPal and their learning. By identifying these clusters (and more importantly their specific configurations), we can provide students with personalized support and also identify students who are not engaging with MathPal effectively in Problem-solving activities.

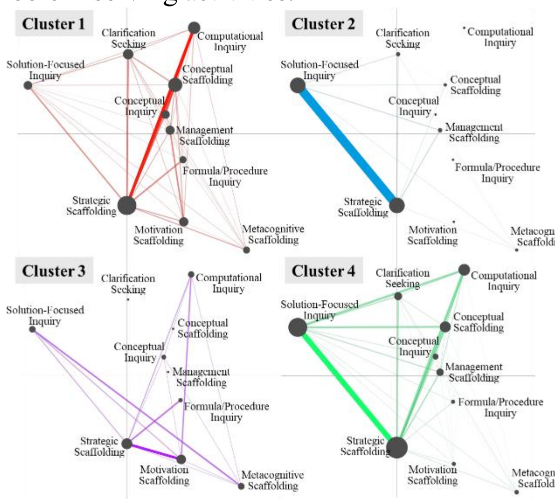


Figure 7: Four types of student-AI interaction patterns

4 Discussion & Implications

This study examined how students interacted with MathPal, an AI-based scaffolding tool, during math problem-solving. Two core findings emerged: (1) students followed consistent inquiry–scaffolding patterns centered on procedural support, and (2) distinct student–AI collaborative problem-solving profiles reflected varying learning behaviors. These findings extend current research on AI tutoring systems (Holstein et al., 2019) and offer new insight into how generative AI can be adapted for more personalized, effective learning support in math problem-solving contexts.

4.1 Personalizing AI Scaffolding Based on Student Inquiry

Most student inquiries were solution-focused or computational, prompting primarily strategic scaffolding from MathPal. While this guided students through procedural tasks, it offered less conceptual or motivational support. This imbalance suggests AI agents should better detect moments of confusion or disengagement to deliver metacognitive or motivational scaffolds.

The findings align with the idea of contingent scaffolding, which emphasizes the need for support to be responsive to students' evolving cognitive and affective states (Van de Pol et al., 2010). Enhancing AI systems with adaptive feedback mechanisms that dynamically shift between strategic, conceptual, and motivational scaffolding may promote deeper learning by moving students beyond surface-level task completion toward the development of enduring mathematical understanding and problem-solving skills (Aleven et al., 2016).

4.2 Personalizing Support Through Interaction Profiles

The distinct interaction profiles offer actionable insights for designing more personalized and effective AI learning support. By recognizing students' habitual inquiry patterns, AI systems can adapt their scaffolding strategies to align with individual cognitive and motivational needs. For instance, students who primarily engage in procedural or solution-focused exchanges may require targeted prompts that promote self-explanation, conceptual elaboration, or metacognitive reflection (Belland, 2017). Conversely, students exhibiting more exploratory or conceptual dialogue might benefit from strategic scaffolds that help organize their thinking and support knowledge gaining. Such adaptive tailoring not only supports differentiated learning trajectories but also enables educators to detect patterns of disengagement or superficial inquiry and intervene to promote deeper engagement and sustained learning growth (Roll & Winne, 2015).

4.3 Theoretical and Methodological Perspectives

From a theoretical lens, the findings align with socio-cognitive perspectives that view learning as a co-constructed process. In the context of this study, students and the AI agent collaboratively constructed knowledge to complete problem-solving tasks through dialogic interaction, learner-driven inquiry, and scaffolded support (Mercer & Howe, 2012). The observed variation in student–AI interaction patterns highlight the importance of context-sensitive scaffolding, reinforcing prior work that emphasizes the dynamic and situated nature of collaborative learning with intelligent tools (Sawyer, 2014). These profiles not only reflect variation in students' inquiry behaviors but may also signal different AI scaffolding patterns in math problem-solving contexts.

Methodologically, this study demonstrates the value of combining qualitative coding with lag sequential analysis and ENA to capture both the structure and flow of human–AI interaction. ENA, in particular, proved effective in visualizing co-occurring patterns that distinguish learner engagement types. The use of clustering to define emergent profiles further enhances ENA's utility in learning analytics, offering a powerful lens for exploring personalized learning trajectories in real-world settings.

5 Conclusion

This study explored how high school students interact with a math AI agent, MathPal, during problem-solving activities. Findings revealed consistent patterns of inquiry and scaffolding, with students frequently seeking procedural help and the AI responding primarily with strategic guidance. Through ENA, four distinct interaction profiles emerged, reflecting differences in how students engage with AI support. These results highlight the need for adaptive AI scaffolding that responds not only to task demands but also to learners' conceptual and motivational needs. By leveraging interaction patterns, educators and designers can create more personalized, responsive AI systems that better support students' math learning in high school classrooms.

References

Vincent Aleven, Elizabeth A. McLaughlin, R. Amos Glenn, and Kenneth R. Koedinger. 2016. Instruction based on adaptive learning technologies. In *Handbook of Research on Learning and Instruction*, 2nd edition, pages 522–560.

Michèle Artigue and Mogens Blomhøj. 2013. Conceptualizing inquiry-based education in mathematics. *ZDM – The International Journal on Mathematics Education*, 45:797–810.

Pete Atherton, Wasiq Khan, and Luke Topham. 2024. AI and student feedback. In *Proceedings of EDULEARN24*.

Brian R. Belland. 2017. *Instructional Scaffolding in STEM Education: Strategies and Efficacy Evidence*. Springer Nature.

Carol S. Dweck. 2006. *Mindset: The New Psychology of Success*. Random House, New York.

Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92. https://doi.org/10.1177/160940690600500107

Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, 6(2):27–52. https://doi.org/10.18608/jla.2019.62.3

Keunjae Kim and Kyungbin Kwon. 2025. A systematic review of the evaluation in K-12 artificial intelligence education from 2013 to 2022. *Interactive Learning Environments*, 33(1):103–131. DOI: 10.1080/10494820.2024.2335499

Minseong Kim, Jihye Kim, Tami L. Knotts, and Nancy D. Albers. 2025. AI for academic success: Investigating the role of usability, enjoyment, and responsiveness in ChatGPT adoption. *Education and Information Technologies*, pages 1–22.

Qi Lang, Minjuan Wang, Minghao Yin, Shuang Liang, and Wenzhuo Song. 2025. Transforming education with generative AI (GAI): Key insights and future prospects. *IEEE Transactions on Learning Technologie*s, 18:230–242. https://doi.org/10.1109/TLT.2025.3537618

Li Li, Yu Fengchao, and Enting Zhang. 2024. A systematic review of learning task design for K-12 AI education: Trends, challenges, and opportunities. *Computers and Education: Artificial Intelligence*, 6:100217. DOI: 10.1016/j.caeai.2024.100217

MathPal. 2023. [Chrome extension]. *Google Chrome Web Store*. https://chromewebstore.google.com/detail/mathpal/gdjlloipbkjmdkdaegicmnccahlfiabd

Neil Mercer and Christine Howe. 2012. Explaining the dialogic process of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction*, 1:12–21. https://doi.org/10.1016/j.lcsi.2012.03.001

Francisco Niño-Rojas, Diana Lancheros-Cuesta, Martha Tatiana Pamela Jiménez-Valderrama, Gelys Mestre, and Sergio Gómez. 2024. Systematic review: Trends in intelligent tutoring systems in mathematics teaching and learning. *International Journal of Education in Mathematics, Science and Technology*, 12(1):203–229. DOI: https://doi.org/10.46328/ijemst.3189

Bethany Rittle-Johnson and Kenneth R. Koedinger. 2005. Designing knowledge scaffolds to support mathematical problem solving. *Cognition and Instruction*, 23(3):313–349. DOI: 10.1207/s1532690xci2303_1

Ido Roll and Philip H. Winne. 2015. Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, 2(1):7–12. DOI: https://doi.org/10.18608/jla.2015.21.2

Sawyer, R. Keith. 2014. *The Cambridge Handbook of the Learning Sciences*, 2nd edition. Cambridge University Press.

David W. Shaffer. 2017. *Quantitative Ethnography. Madison*, WI: Cathcart Press.

Anselm Strauss and Juliet Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 2nd edition. Thousand Oaks, CA: Sage.

Yong Khin Tay and Tin Lam Toh. 2023. A model for scaffolding mathematical problem-solving: From theory to practice. *Contemporary Mathematics and Science Education*, 4(2):ep23019. https://doi.org/10.30935/conmaths/13308

Janneke van de Pol, Monique Volman & Jos Beishuizen. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3):271–296. https://doi.org/10.1007/s10648-010-9127-6

Bilan Zhang and Jia Jiyou. 2017. Evaluating an intelligent tutoring system for personalized math teaching. In *Proceedings of the 2017 International Symposium on Educational Technology* (ISET), pages 126–130. IEEE.

Pre-trained Transformer Models for Standard-to-Standard Alignment Studies

Hye-Jeong Choi[a], Reese Butterfuss[b], Meng Fan[a], and Emily Dickinson[a]

[a] *HumRRO*
[b] *Certiverse*

Abstract

The current study evaluated the accuracy of five pre-trained large language models (LLMs) in matching human judgment for standard-to-standard alignment study. Results demonstrated comparable performance across LLMs despite differences in scale and computational demands. Additionally, incorporating domain labels as auxiliary information did not enhance LLMs performance. These findings provide initial evidence for the viability of open-source LLMs to facilitate alignment study and offer insights into the utility of auxiliary information.

1 Introduction

Large language models (LLMs) are increasingly used in educational and psychological measurement activities. Their evolving sophistication and ability to represent deep, contextual semantics make them viable tools to support subject matter experts (SMEs) in reviewing large volumes of text-based context, such as educational standards (e.g. Butterfuss & Doran, 2025; Kim et al., 2023; Kusumawardani & Alfarozi, 2023; Zhou & Ostrow, 2022). However, little guidance exists on the effective use of LLMs in such contexts. Our goal was to compare popular, pretrained LLMs in a common measurement context (i.e., standard-to-standard alignment) to provide initial evidence on which LLMs may be particularly useful for measurement tasks that require extensive review of large bodies of text.

Alignment is a critical aspect of validity evidence for any assessment (AERA, APA, NCME, 2014). Standards-to-standards alignment is a process to examine how well two distinct sets of content standards target the same content (Neidorf et al., 2016). In general, it requires SMEs to review two sets of standards and determine alignment such that each standard in one set is evaluated against the standards in the second set until any or all standards that capture the same meaning are identified. It is a time-consuming process because it requires evaluation of potentially thousands of possible pairs of content standards. Recently, the potential for NLP and LLMs as a supporting tool in this process has been presented (e.g., Butterfuss & Doran, 2024; Zhou & Ostrow, 2022), but there is a lack of work that provides guidance on which LLMs to choose for such tasks.

This study aimed to address two research questions: (1) how do five popular pre-trained transformer models compare in standards-to-standards alignment studies? and (2) does auxiliary information (e.g., domain label) impact LLMs performance? Educational standards, typically presented as brief, abstract statements, often include examples to provide clarity and context. It is also not uncommon to have the exact same standard appear under different domains. Understanding how such auxiliary information influences LLMs performance is crucial for developing more effective automated alignment tools.

2 Methods

2.1 Data

The alignment study dataset used for the current study consisted of individual standards from 33 states and aligned each state standard to the corresponding the National Assessment of Educational Progress (NAEP) standard for grades 4, 8, and 12 for science. Each standard was classified into one of three domains: life science (LS), physical science (PS), and earth & space science (ES). The number of potential pairs ranged

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 306–311
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

from approximately 18,000 to 60,000. The number of standards represented within each state varied. In the original work, SMEs judged each possible pair of standards as aligned, partially aligned, or not aligned. Thus, we used the SME decision as "ground truth" for evaluating the LLMs. More details about the dataset and original alignment study can be found in the published report (Dickinson et al., 2021).

2.2 Pre-trained transformer models

We accessed the LLMs via the Hugging Face Transformers library, a popular open-source library that provides a simple and consistent way to use pre-trained models for various NLP tasks. As of 2025, the Hub hosts over 50,000 models, many of which are based on Transformer architectures.

The LLMs transform each content standard into an embedding, or numeric representation of the meaning of the text. Once every standard is transformed into an embedding, then the relations among the embeddings can be evaluated using cosine similarity. Cosine similarity is a metric used to measure how similar two vectors are irrespective of their magnitude. It calculates the cosine of the angle between two vectors, determining whether they point in roughly the same direction. Commonly used in text analysis, recommendation systems, and information retrieval. While it behaves similarly to a correlation in some contexts, cosine similarity specifically only measures directional similarity, not linear correlation or magnitude.

In this study we used the cosine similarity between every possible pair of standards that can be made from the two sets. Doing so allows us to gauge which standard pairs are more similar than others. Critically, standard pairs that share high semantic overlap (i.e., large cosine similarity values) are more likely to be aligned than standard pairs that share little semantic overlap (Butterfuss & Doran, 2025).

To calculate cosine similarity, we utilized five LLMs which are widely used, including all-distilroberta-v1, all-MiniLM-L6-v2, multi-qa-MiniLM-L6-cos-v1, all-mpnet-base-v2, and gtr-t5-large. All of these are sentence embedding models that can be used to calculate cosine similarity between texts. The mathematical formula for calculating cosine similarity remains the same across all these models. However, LLMs vary in the specific linguistic features their embeddings emphasize, and thus LLMs differ in which aspects of meaning contribute to cosine similarity values. Due to this variability, we extracted embeddings for each standard using five different popular LLMs:

- **all_distilroberta_v1 (DistilRoBERTa-v1).** It is a distilled version of the RoBERTa (Liu et al., 2019) model to cover a wide range of topics and styles. It is a smaller, more efficient model that's designed to be faster and more computationally efficient.

- **all_MiniLM_L6_v2 (MiniLM-L6-v2).** MiniLM is designed for efficiency and smaller size. It's useful for text classification, sentiment analysis, or question answering. They are particularly useful for deployment in resource-constrained environments, such as mobile devices or edge computing platforms (Wang et al., 2020).

- **multi_qa_MiniLM_L6_cos_v1 (MultiQA-MiniLM-L6)**. It is a variant of the MiniLM model that is designed for multi-question answering tasks, such as answering multiple questions about a given text passage, identifying relevant passages or sentences that answer multiple questions, and generating answers to multiple questions based on a given text passage.

- **all_mpnet_base_v2 (MPNet-Base-v2)**. A model known for its efficiency and performance on a wide range of NLP tasks, including text classification, sentiment analysis, question answering, and more. It's particularly useful when you need a model that can handle long-range dependencies and contextual relationships in text data.

- **gtr_t5_large (GTR-T5-Large)**. It is known as a powerful language model. It can be used text generation and summarization, question answering and reading comprehension, sentiment analysis and opinion mining, and language translation and machine translation.

2.3 Three approaches to set a threshold

We employed three threshold setting approaches to pair state and NAEP standards: cosine similarity value, percentile, and rank order. First, we used predetermined cosine similarity values: if the cosine similarity of two-paired standards was greater than the predetermined cosine similarity value, we classified the state-to-NAEP standard pair as aligned. We used three different values as the predetermined value (i.e., 0.4, 0.5, 0.6).

Second, we used a percentile to set the cut score cosine similarity value. As mentioned in the previous section, cosine similarity is a measure of direction but not magnitude. Using percentile can resolve potential scaling issues across LLMs. We used three percentiles (i.e., 70, 80, 90) to obtain the threshold of cosine similarity to pair standards.

Finally, we utilized a rank-order approach to classify aligned standard pairs: if the cosine similarity of standard pairs was within the predetermined top n highest cosine similarity, we classified those standards as aligned. We used top 3, 5 and 10. After we classified each pair as either aligned or not aligned based on those criteria, we evaluated those results with SMEs judgment.

LLMs performance was evaluated using overall accuracy, recall and F1 metrics. Overall accuracy (either hit rate or precision) refers to the probability of capturing the true matches (according to human judgment) within condition. Recall measures the proportion of actual positive instances that were correctly identified by the model. It is a metric used to evaluate the completeness of a classification model's positive predictions. The F1-score is a metric that combines precision and recall into a single value, providing a balance between these two sometimes competing metrics. Precision measures the proportion of correctly identified positive instances among all instances that the model predicted as positive. It's particularly useful when you need a single measurement to evaluate a classification model's performance.

3 Results

3.1 Comparison of five pre-trained transformer models

Table 1 presents descriptive statistics for the cosine-similarity values generated by each LLM for each grade. Overall, the correlations between LLMs were high (higher than .76). In both conditions, the results indicated that the models produced cosine-similarity values that were scaled slightly differently. In particular, means of GTR-T5-Large were higher and standard deviations were smaller than other LLMs.

Table 1 Descriptive statistics of cosine similarity by LLMs for each grade

	MPNet	Distil RoBERT	Mini LM	GTR-T5	Multi QA
Grade 4 (N=18,744)					
Mean	.22	.18	.19	.55	.17
STD	.14	.14	.14	.07	.14
Grade 8 (N=55,857)					
Mean	.22	.18	.20	.56	.18
STD	.13	.13	.14	.06	.14
Grade 12 (N=59,829)					
Mean	.20	.16	.18	.55	.16
STD	.13	.13	.14	.06	.13

Note. MPNet = MPNet-Base-v2; DistilRoBERT = DistilRoBERTa-v1; MiniLM = MiniLM-L6-v2; GTR-T5 = GTR-T5-Large; MultiQA = MultiQA-MiniLM-L6

Table 2 summarizes comparison of LLMs to SMEs with respect to the overall accuracy, recall, and F1. Note that we used three different methods to classify pairs of standards: cosine similarity value, percentile, and rank order. First, notably, the free, open-source models fared nearly as well as the costlier, more computationally intensive model (GTR-T5-Large). Overall, the correlations between LLMs were high (higher than .76). Second, all LLMs performed similarly in capturing the true pairs with respect to F1 and accuracy. However, recall indicates percentile and rank performed much better to identify the true pairs for all five models. When cut score was used, GTR-T5-Large performed differently from other four models. That was because cosine similarity from GTR-T5-Large tended different and larger than four other models.

Table 2 Overall comparison LLMs results with SMEs rating under different conditions

Model	Stats	Cut Value		Percentile		Rank	
		M	STD	M	STD	M	STD
MPNet-Base-v2	F1	.24	.21	.29	.13	.35	.13
	Recall	.22	.26	.91	.11	.94	.05
	Accuracy	.98	.01	.86	.10	.89	.07
Distil RoBERTa -v1	F1	.20	.21	.29	.13	.35	.13
	Recall	.16	.21	.90	.11	.94	.06
	Accuracy	.98	.01	.86	.10	.90	.07
MiniLM -L6-v2	F1	.24	.21	.29	.13	.35	.13
	Recall	.20	.24	.90	.12	.94	.06
	Accuracy	.98	.01	.86	.10	.90	.07
GTR-T5 -Large	F1	.19	.19	.29	.14	.35	.14
	Recall	.53	.44	.91	.10	.95	.04
	Accuracy	.79	.30	.86	.10	.89	.07
MultiQA -MiniLM -L6	F1	.21	.19	.27	.12	.34	.13
	Recall	.16	.20	.87	.13	.92	.06
	Accuracy	.98	.00	.85	.10	.90	.07

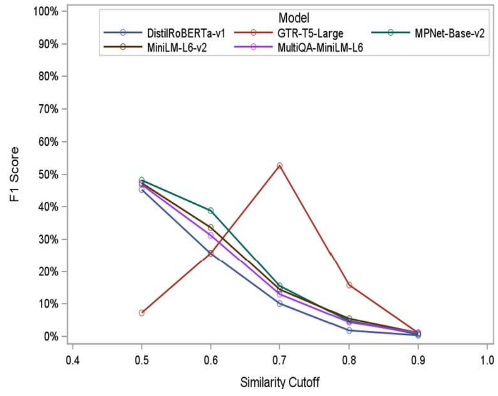


Figure 1 F1 score trend of LLMs across cosine similarity points (Grade 4)

Table 4 Comparison of LLMs at different ranks (Grade 4, N=18,744)

Model	Rank	Accuracy	Recall	F1
MPNet-Base-v2	3	.95	.87	.50
MPNet-Base-v2	5	.90	.95	.34
MPNet-Base-v2	10	.76	.99	.19
DistilRoBERTa-v1	3	.95	.83	.49
DistilRoBERTa-v1	5	.90	.93	.34
DistilRoBERTa-v1	10	.76	.99	.19
MiniLM-L6-v2	3	.95	.85	.49
MiniLM-L6-v2	5	.90	.94	.34
MiniLM-L6-v2	10	.76	1.00	.19
GTR-T5-Large	3	.95	.91	.51
GTR-T5-Large	5	.90	.97	.35
GTR-T5-Large	10	.76	1.00	.19
MultiQA-MiniLM-L6	3	.95	.83	.49
MultiQA-MiniLM-L6	5	.90	.91	.34
MultiQA-MiniLM-L6	10	.76	.99	.19

Table 3 Comparison of LLMs at different percentiles (Grade 4, N=18,744)

Model	%ile	Accuracy	Recall	F1
MPNet-Base-v2	70	.73	1.00	.17
MPNet-Base-v2	80	.83	.97	.24
MPNet-Base-v2	90	.92	.83	.37
DistilRoBERTa-v1	70	.73	.99	.17
DistilRoBERTa-v1	80	.83	.94	.24
DistilRoBERTa-v1	90	.92	.83	.37
MiniLM-L6-v2	70	.73	.99	.17
MiniLM-L6-v2	80	.83	.95	.24
MiniLM-L6-v2	90	.92	.84	.37
GTR-T5-Large	70	.73	.99	.17
GTR-T5-Large	80	.83	.97	.24
GTR-T5-Large	90	.92	.89	.40
MultiQA-MiniLM-L6	70	.73	.98	.17
MultiQA-MiniLM-L6	80	.83	.95	.24
MultiQA-MiniLM-L6	90	.92	.79	.35

Next, we will present the results focusing on Grade 4 as the results with other grades were similar. Figure 1 depicts F1 across several cosine similarity points from .50 to .90 for all five language models for Grade 4. GTR-T5-Large performed best when the cosine similarity was set at .70 whereas four languages models performed best when the cosine similarity was set at .50. Table 4 presents the comparison of LLMs with different ranks. As expected, LLMs captured the true pair with lower ranks (rank=3). However, recall indicates LLMs captured the true pair with rank=10. In other words, the NAEP standards, with which the state standard is aligned, appear among the top ten pairs rank-ordered by cosine similarity. Table 3 shows the comparison of LLMs with different percentile. Again, LLMs performed similarly: the higher percentile, the better in capturing the true pairs with respect to accuracy whereas the lower percentile, the better in terms of recall. Both accuracy and recall indicate LLMs with either 70 percentile or rank order 10 well captured the true pairs. In other words, the NAEP standards, with which the state standard is aligned, appear among the top ten or even top five pairs rank-ordered or 70 or 80 percentiles by cosine similarity.

3.2 Effects of domain information effect on cosine similarity

Table 7 presents cosine similarity distributions when domain labels were added for each grade. Note that the N counts for all grades were slightly larger than the N counts in Table 1. This was because the same standard was assigned into different domains. The descriptives were similar with ones in Table 1; however, those values were slightly lower. Also, the correlations between cosine similarity measures for standard pairs with domain were similar with ones without domain, ranging from .75 to .92.

Next, we compare how LLMs performed to capture the true pairs compared with cosine similarity without domain. Again, we present the results for Grade 4 as the results with other grades were similar. Figure 2 depicts F1 scores across several cosine similarity points from .50 to .90 for all five language models for Grade 4. The results show a similar pattern with Figure 1; with respect

Table 7 Descriptive statistics of cosine similarity with domain by LLMs for each grade

	MPNet	Distil RoBERT	Mini LM	GTR -T5	Multi QA
Grade 4 (N=18,846)					
Mean	.21	.17	.17	.55	.15
STD	.13	.13	.14	.06	.13
Grade 8 (N=56,932)					
Mean	.22	.17	.19	.56	.16
STD	.12	.13	.13	.06	.13
Grade 12 (N=59,927)					
Mean	.20	.15	.17	.55	.14
STD	.13	.13	.13	.06	.13

Note. MPNet = MPNet-Base-v2; DistilRoBERT = DistilRoBERTa-v1; MiniLM = MiniLM-L6-v2; GTR-T5 = GTR-T5-Large; MultiQA = MultiQA-MiniLM-L6

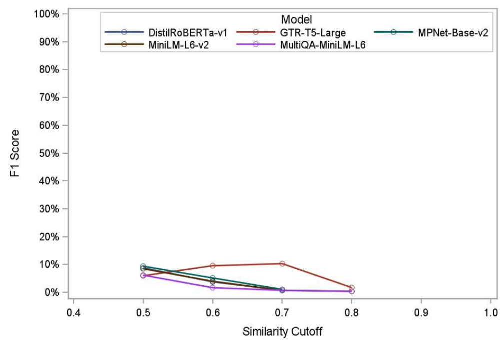


Figure 2 F1 Score by cosine similarity cut for each language model with domain (Grade 4)

to F1, GTR-T5-Large outperformed and performed the best when the cosine similarity was set at .70. Overall, however, all the five LLMs performed slightly worse with domain labels.

Table 6 and 7 show that domain information improved accuracy but reduced recall and F1. Adding domain labels did not enhance overall model performance.

4 Summary and discussion

The results of the current study indicated that scaling differences among LLMs in raw cosine similarity values meant that using a raw cosine value threshold may not be feasible, particularly when comparing multiple LLMs. Overall, when percentile or rank order was used, the results suggest that the five LLMs performed comparably with respect to accuracy for standards-to-standards alignment of science content standards. Specifically, the models were generally 90% accurate at capturing the "true matches" according to human judgment above the 90 percentile or within the top five highest-cosine pairs. Put another way, for a given state standard, the SME-aligned NAEP standard had appeared either the 90 percentile or among the top five cosine similarity pairs.

Table 6 Comparison of LLMs with domain at different ranks (Grade 4, N=18,876)

Model	Rank	Accuracy	Recall	F1
MPNet-Base-v2	3	.97	.23	.28
MPNet-Base-v2	5	.95	.31	.25
MPNet-Base-v2	10	.88	.44	.17
DistilRoBERTa-v1	3	.97	.21	.26
DistilRoBERTa-v1	5	.95	.28	.24
DistilRoBERTa-v1	10	.88	.43	.17
MiniLM-L6-v2	3	.97	.23	.27
MiniLM-L6-v2	5	.95	.30	.25
MiniLM-L6-v2	10	.88	.44	.17
GTR-T5-Large	3	.97	.26	.30
GTR-T5-Large	5	.95	.33	.25
GTR-T5-Large	10	.87	.47	.17
MultiQA-MiniLM-L6	3	.97	.20	.25
MultiQA-MiniLM-L6	5	.95	.29	.24
MultiQA-MiniLM-L6	10	.88	.45	.17

Table 5 Comparison of LLMs with domain at different percentiles (Grade 4, N=18,876)

Model	%ile	Accuracy	Recall	F1
MPNet-Base-v2	70	.70	.44	.08
MPNet-Base-v2	80	.79	.35	.09
MPNet-Base-v2	90	.88	.23	.10
DistilRoBERTa-v1	70	.70	.44	.08
DistilRoBERTa-v1	80	.79	.33	.08
DistilRoBERTa-v1	90	.88	.23	.10
MiniLM-L6-v2	70	.70	.45	.08
MiniLM-L6-v2	80	.79	.35	.09
MiniLM-L6-v2	90	.88	.23	.10
GTR-T5-Large	70	.70	.46	.08
GTR-T5-Large	80	.79	.38	.09
GTR-T5-Large	90	.89	.25	.11
MultiQA-MiniLM-L6	70	.70	.44	.08
MultiQA-MiniLM-L6	80	.79	.33	.08
MultiQA-MiniLM-L6	90	.88	.21	.09

Using Grade 4 from our real-world alignment study as an example, the current method would reduce the number of pairs that SMEs must compare from nearly 18,000 pairs to around 2,840 pairs. Moreover, the current findings suggest that any of the popular, open-source LLMs we compared may yield such benefits. Thus, for contexts similar to those in the current work, researchers and practitioners may be well-suited to choose any of the models we evaluated given their comparable performance. Also, the current findings highlight a potentially enormous efficiency increase by dramatically reducing the number of pairs SMEs must consider via economical LLMs and a relatively simple percentile or rank-order approach using cosine-similarity.

Unfortunately, adding “domain” to the standard did not improve LLMs performance to capture the true matches. Subsequent work in this area is needed to examine the added benefits of including more contextual information for the LLMs when extracting embeddings for each standard (i.e., content domain descriptions), as well as the conditions under which it is useful to include or omit accessory information that some content standards include, such as exemplary information or explanatory information. The results did not show LLMs performance were substantially impacted by grade.

Critically, the current LLM approach does not *replace* humans in making alignment decisions. Instead, the method provides a simple, economical way to support SMEs in making alignment decisions more efficiently by leveraging an organizational structure based on semantic similarity and constraining the number of viable pairs that must be considered. Overall, the current study represents a judicious, human-centered use of AI in a laborious routine measurement task.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Butterfuss, R., & Doran, E. (2025). Advancing measurement with large language models: Implications for content review. *Journal of Educational Measurement, 62*(1), 45–63.

Dickinson, E. R., Gribben, M., Schultz, S. R., Spratto, E., & Woods, A. (2021). *Comparative analysis of the NAEP science framework and state science standards* (Tech. Rep.). National Assessment Governing Board. Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/NAEP-Science-Standards-Review-Final-Report-508.pdf

Kim, D. E., Hong, C., & Kim, W. H. (2023, July). Efficient Transformer-based Knowledge Tracing for a Personalized Language Education Application. In *Proceedings of the Tenth ACM Conference on Learning@ Scale* (pp. 336-340).

Kusumawardani, A., & Alfarozi, M. (2023). Exploring AI-driven tools for curriculum mapping. *International Journal of Educational Technology*, 18(2), 110–125.

Neidorf, T. S., Binkley, M., Galia, J., & Stephens, M. (2016). *Assessing alignment among curriculum, instruction, and assessments: Methodologies and findings*. OECD Publishing.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

Zhou, Z., & Ostrow, K. S. (2022, June). Transformer-Based Automated Content-Standards Alignment: A Pilot Study. In *International Conference on Human-Computer Interaction* (pp. 525-542). Cham: Springer Nature Switzerland.

From Entropy to Generalizability: Strengthening Automated Essay Scoring Reliability and Sustainability

Yi Gui
University of Iowa
yi-gui@uiowa.edu

Abstract

Generalizability Theory with entropy-derived stratification optimized automated essay scoring reliability. A G-study decomposed variance across 14 encoders and 3 seeds; D-studies identified minimal ensembles achieving G $\geq$ 0.85. A hybrid of one medium and one small encoder with two seeds maximized dependability per compute cost. Stratification ensured uniform precision across complexity.

1 Introduction

Automated Essay Scoring (AES) systems based on transformer architectures have transformed large-scale writing assessment by offering both scalability and consistency that rival human raters (Shermis & Burstein, 2013). However, before deployment in high-stakes contexts—such as college admissions or professional licensure—AES models must meet stringent psychometric standards for reliability. In Classical Test Theory, indices like Cronbach's alpha confound multiple error sources into a single coefficient, obscuring the distinct contributions of prompt heterogeneity, model stochasticity, and text complexity (Nunnally & Bernstein, 1994). **Generalizability Theory (G-Theory)** overcomes these limitations by decomposing observed-score variance into multiple facets and interactions, yielding the generalizability coefficient (G-coefficient) as an overall index of dependability (Brennan, 2001; Cronbach et al., 1972).

In any G-Theory study, the object of measurement—the entity whose universe score we seek to estimate—is critical. For AES, as in large-scale writing assessments like the GMAT AWA, these objects are the test-takers themselves (Gao et al., 2015). Here, each essay's observed model score is viewed as an estimate of that test-taker's "universe score" across all combinations of prompts, model seeds, and cross-validation folds.

A persistent obstacle to AES reliability is uneven precision across essays of varying complexity: richly worded, syntactically complex essays tend to yield larger residual errors, thereby reducing the G-coefficient for those subsets (Shermis & Burstein, 2013). **Shannon entropy**—computed from token-frequency distributions—offers a principled, near–zero-cost measure of textual complexity (Shannon, 1948), but raw entropy correlates strongly ($r > .80$) with essay length. To decouple length from unpredictability, we standardize entropy (z-score) and stratify essays into equal-size buckets that contribute uniformly to variance-component estimates.

The present study applies a Generalizability-Theory approach to automated essay scoring by incorporating standardized entropy buckets into both G-Study and D-Study phases. After estimating variance components for prompts, entropy strata, encoder architectures, seed initializations, and folds, we analytically predict G-coefficients for alternative measurement designs—varying the number and combination of transformer models and random seeds—without refitting mixed models. Our aim is to identify minimal ensembles that meet a target reliability (G $\geq$ 0.85), thereby guiding more efficient and cost-effective AES deployments.

This study is guided by three primary questions.

1) In a fully crossed G-study of test-takers (essays) × encoders × seeds × folds, what proportions of total score variance are attributable to encoder choice, seed

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 312–328

initialization, and fold assignment, and how practically meaningful are these facets for overall reliability?

2) Can essays stratified into equal-size buckets based on standardized Shannon entropy be seamlessly integrated into a G-Theory design, thereby supporting more precise D-study planning?

3) In the D-study phase, which minimal combinations of encoder architectures and seed replications suffice to reach the predefined reliability threshold (G≥0.85), and how does this entropy-informed approach reduce unnecessary computation in AES system development?

2 Related Work

Evolution of Scoring Reliability Methods in Writing Assessment.

Early research on essay scoring focused on Classical Test Theory (CTT) concepts of reliability, using measures like inter-rater correlation or Cohen's kappa to judge consistency between human graders. A kappa value of 0.7 is often cited as an acceptable minimum for essay scoring reliability, as it accounts for roughly half of score variance. Traditional writing assessments thus aimed for high inter-rater agreement to ensure reliable scores. However, CTT-based reliability is limited in that it offers a single coefficient (like Cronbach's alpha or inter-rater correlation) for a given test under fixed conditions, without disentangling multiple error sources. This is problematic for essay tasks, where score variance may arise from multiple facets – differences in raters, prompts, or occasions. Researchers recognized that a more nuanced framework was needed beyond what CTT provides.

Generalizability Theory (G-Theory) emerged as that framework, extending CTT by incorporating analysis of variance to parse out various sources of measurement error (persons, raters, tasks, etc.) and to estimate the dependability of scores under varying conditions. Cronbach et al. (1972) introduced the G-Theory model, defining the generalizability coefficient (G-coefficient) as an index analogous to reliability that reflects how well observed scores generalize to the universe of all possible scoring conditions. Unlike a single CTT reliability, the G-coefficient can account for, say, multiple essay prompts or rater inconsistencies simultaneously. Subsequent works (Brennan, 2001; Shavelson & Webb, 1991) further formalized G-Theory and its use in performance assessments. In writing assessment research, G-Theory has been used to examine how many raters or prompts are needed to achieve dependable scores and to diagnose where inconsistencies arise. For instance, Huang (2008) applied G-theory to ESL writing tests and found that using multiple tasks and raters significantly improved score accuracy. These studies demonstrated that classical inter-rater reliability indices could be inadequate, and that G-Theory provides deeper insight into the facets affecting essay score reliability.

Generalizability Theory Applied to AES.

As Automated Essay Scoring (AES) systems have matured—especially with the advent of transformer-based models—researchers have turned to Generalizability Theory (G-Theory) to rigorously evaluate their reliability. Williamson et al. (2012) first argued that an AES engine must demonstrate stable performance not just on a single prompt but across diverse essay tasks and forms. In the high-stakes context of standardized tests—where the objects of measurement are the test-takers—Gao et al. (2015) extended that framework in their GMAC AWA study by modeling facets such as prompts, essay types (fixed), rating engines versus human raters, and occasions. Because a fully crossed design was impractical, they employed overlapping G-Studies and D-Studies to approximate universe-score variance, reporting operational G-coefficients around 0.83.

Subsequent empirical work has confirmed the value of this approach. Han and Sari (2024) applied G-Theory to compare human raters with ETS's e-rater on a set of EFL essays, finding that human raters introduced more score variance than the automated engine. When automated and human scores were combined, overall dependability improved—underscoring how AES can mitigate human-rater inconsistency. Bridgeman et al. (2012) similarly showed that an AES system maintained comparable reliability across gender and ethnic subgroups, suggesting that machine scoring does not exacerbate demographic biases in measurement error.

Together, these studies illustrate two key points: first, that G-Theory provides a nuanced, facet-level understanding of where scoring variability arises; and second, that modern AES engines can achieve reliability on par with—or even exceeding—that of human raters, provided they are evaluated across a representative range of prompts and test-taker populations.

Entropy-Based Approaches and Score Uncertainty

Information-theoretic metrics—most notably Shannon entropy—have gained traction as tools for characterizing essay complexity and anticipating scoring uncertainty in automated systems. In classical educational measurement, conditional standard errors of measurement (CSEM) acknowledge that precision can vary across score levels or item types; by analogy, essays whose linguistic patterns are highly unpredictable may elicit greater variability in both human and machine-generated scores. Shannon entropy, computed from an essay's token-frequency distribution, quantifies this unpredictability: higher entropy reflects richer vocabulary and structural diversity, which can challenge consistency in scoring. Several studies have extended this concept by measuring relative entropy (Kullback–Leibler divergence) between an essay's word distribution and a reference language model, demonstrating that essays with greater divergence tend to produce more dispersed human ratings (Atkinson & Palma, 2025).

Other work has leveraged entropy of next-token probabilities from pretrained transformers to flag low-confidence segments, finding that these high-entropy regions correspond to larger prediction errors. Such entropy-derived features thus serve as data-driven proxies for CSEM, identifying essays on which automated scorers are likely to be less reliable (Atkinson & Palma, 2025).

Although entropy-informed methods are not yet ubiquitous in production AES pipelines, they offer a promising complement to aggregate reliability indices and G-Theory analyses. Integrating entropy as either a stratification facet or a diagnostic feature enables more nuanced dependability assessments, pinpointing when an individual essay score may warrant caution. As AES systems continue to evolve, embedding information-theoretic insights can enhance both psychometric rigor and operational transparency, ensuring that automated evaluations maintain equitable precision across the full spectrum of student writing complexity.

3 Methods

Data

The PERSUADE 2.0 corpus originally comprised over 25,000 persuasive essays from U.S. students in grades 6–12. We restricted our analysis to ninth through twelfth graders, yielding 13,815 unique essays each annotated with a holistic score and writer demographics (prompt, task type, grade, gender, socioeconomic and ELL status). Because each examinee submitted exactly one essay, essay IDs are fully confounded with test-taker identity, so all essay-level variance components mirror examinee-level differences. To ensure balanced coverage across topics, tasks, and grade levels, we performed a stratified split by prompt × task × grade, sequestering 10% (1,381 essays) as an unlabeled "Holdout Evaluation Set" and retaining 12,428 essays for G-study and D-study modeling (Figures 1–2).

Each essay in the holdout set was then scored according to a fully crossed design of 14 transformer encoders × 3 random seeds × 5 cross-validation folds, for a total of 14 × 3 × 5 = 210 predictions per essay. All encoders shared a common ordinal logistic regression (OLR) head during fine-tuning and inference. Inference proceeded by tokenizing each essay into sliding-window segments matched to model-specific maximum sequence lengths, pooling segment representations into a single fixed-length embedding, and passing that embedding through the OLR head. The resulting 210 predicted scores per essay were collated into one dataset. The wall-clock times for the models are all presented in Table 10.[1]

To examine complexity effects, we computed raw Shannon entropy from token-frequency distributions (Figure 3, $r = 0.741$ with word count), then standardized these values (z-scores) and applied our equal-variance bucketing algorithm. Standardization inverted the length correlation ($r =$

1 Wall-clock time refers to the real-world elapsed time measured from the start to the end of each processing step, encompassing all computation, data loading, and any idle or I/O wait times—analogous to timing an event with a stopwatch.

−0.641), equalizing variability across essay lengths. Plotting sample variance against standardized entropy (Figure 4) reveals a flat cloud of points, indicating no systematic rise in inconsistency for more complex texts. Likewise, mean predicted score remains constant across entropy levels (Figure 5), confirming that our stratification yields uniform precision (variance) and fairness (mean score) regardless of textual complexity.

Entropy Standardization and Bucketing Method

To control for the confounding effect of essay length on token-based complexity measures, we first computed each essay's Shannon entropy H_i from its token-frequency distribution:

$$H_i = -\sum_t p_{i,t} log\,(p_{i,t}) \qquad (1)$$

where $p_{i,t}$ is the relative frequency of token t in essay i. These raw entropy values were then standardized to z-scores

$$z_i = \frac{H_i - \bar{H}}{SD(H)} \qquad (2)$$

centering and scaling the distribution so that z_i has mean zero and unit variance across the corpus.

Because our goal was to ensure that each complexity stratum contributed equally to the D-study's variance estimates, we partitioned the essays into buckets by **equalizing the sum of their observed score variances** rather than by simple quantile splits of z_i. Denoting by v_i the sample variance of the 210 predictions for essay i, we sought a set of cut-points $\{c_1,\ldots,c_{K-1}\}$ that induced buckets $B_k = \{i: c_{k-1} < z_i \leq c_k\}$ satisfying

$$\sum_{i \in B_k} v_i \approx \frac{1}{K}\sum_{i=1}^{N} v_i \; for\; k = 1,\ldots,K \qquad (3)$$

Starting from a maximum K, we iteratively reduced K until every prompt × bucket cell contained at least the pre-specified minimum number of essays. In practice, we sorted essays by z_i, formed cumulative sums of v_i, and placed cut-points at entropy values corresponding to equal increments of total variance. Essays were then assigned to buckets by thresholding their z_i against these cut-points. This "equal-variance" binning guarantees that each entropy stratum contributes the same total score dispersion—and, by enforcing a minimum cell size per prompt, preserves adequate data in every prompt × bucket combination for reliable variance-component estimation.

G-Study Design

To decompose score variance across measurement facets, we fit a series of linear mixed-effects models in R using the lme4 package. After reshaping the predictions into long format—one record per essay × encoder × seed × fold—and subsetting by entropy bucket, we specified the model

$$Predicted\_score_{ijkl} = \mu + u_{p[i]} + v_{b[i]} + w_{p[i],b[i]} + x_{e[i]} + \varepsilon_{ijkl} \qquad (4)$$

Where $u_{p[i]} \sim \mathrm{N}(0, \sigma_{\mathrm{p}}^2)$ captures prompt-level variance, $v_{b[i]} \sim \mathrm{N}(0, \sigma_{\mathrm{b}}^2)$ the bucket (entropy) effect, $w_{p[i],b[i]} \sim \mathrm{N}(0, \sigma_{\mathrm{pb}}^2)$ their interaction, $x_{e[i]} \sim \mathrm{N}(0, \sigma_{\mathrm{e}}^2)$ and the essay (test-taker) facet, and $\varepsilon_{ijkl} \sim \mathrm{N}(0, \sigma_{\mathrm{r}}^2)$ residual error. From each model we extracted the variance components $\sigma_{\mathrm{p}}^2, \sigma_{\mathrm{b}}^2, \sigma_{\mathrm{pb}}^2, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{r}}^2$ via VarCorr(), supplying the inputs for subsequent analytic D-study computations.

D-Study Design

Using those variance components, we predicted the generalizability coefficient G without refitting:

$$G = \frac{\sigma^2_{test-taker}}{\sigma^2_{test-taker} + \frac{\sigma^2_{prompt:bucket}}{n_{buckets}} + \frac{\sigma^2_{residual}}{n_{buckets} \times n_e \times n_s}} \qquad (5)$$

Here $n_{buckets}$, n_e , and n_s are the numbers of entropy buckets, encoders, and seeds. Two sweeps were performed:

Small-only ensembles: All $\binom{6}{\mathrm{k}} \times \binom{6}{\mathrm{s}}$ combinations of k=2–6 small encoders and s=1–3 seeds (399 runs) identified the minimal small-model sets achieving G≥0.85.

Mixed small + medium ensembles: For each subset of 1–4 medium encoders, we incrementally added 1–6 small encoders under 1–3 seeds (5,401 runs), halting further expansion once a medium–seed pair first attained G≥0.95.

Large-model exploration: Finally, we considered ensembles incorporating at least one large encoder (BERT-Large, RoBERTa-Large, GPT-2 Large, DeBERTa-V3 Large), applying a second early-stop rule: after two distinct large-inclusive sets surpassed G≥0.95, no further large-model sweeps were conducted.

These analytic D-study procedures rigorously chart reliability gains against computational expense, guiding selection of AES ensembles that meet dependability targets with minimal overhead.

4 Results

1. Overall Reliability Ceiling

The fully crossed G-study design—incorporating all 14 encoders, three seed initializations, and five-fold cross-validation—yielded near-perfect generalizability coefficients across entropy strata. Substituting $\mathrm{n_e} = 14$, $\mathrm{n_s} = 3$, and $\mathrm{n_{buckets}} = 3$ into the analytic D-study formula produced

Low-entropy bucket: G=0.990

Mid-entropy bucket: G=0.989

High-entropy bucket: G=0.989

An overall average of G≈0.99 confirms that exhaustive model diversity and replication effectively eliminate measurement error, establishing an upper bound against which all reduced-complexity ensembles are benchmarked.

After experimenting with different K values under a minimum cell-size constraint, we determined that only by requiring at least seven essays per prompt–bucket cell could the equal-variance algorithm produce three strata. Higher K values violated this constraint, and K = 2 proved too coarse for reliable variance estimation. With the minimum cell size set to seven, the algorithm converged on K = 3 buckets containing 451, 479, and 457 essays (Table 2), which we label as low, medium, and high complexity based on their mean standardized-entropy (z) scores. Table 3 details each prompt's essay counts within these buckets, confirming an even distribution of texts across topics and complexity levels.

Figure 6 illustrates that the holistic score distributions are virtually identical across low, medium, and high entropy strata. Each bucket peaks in the mid-score range (2–4) and tapers symmetrically toward the extremes (1 and 6), with no stratum showing a disproportionate concentration at any particular score band. This consistency confirms that our equal-variance bucketing preserved the overall score profile: textual complexity, as indexed by standardized entropy, does not systematically bias the distribution of human-assigned scores.

2. G-Study Facet Contributions

Despite this high ceiling, the initial G-study decomposition (Table 4) reveals that genuine test-taker differences remain the predominant source of score variability. Across all three entropy strata, the essay/test-taker component σ^2_{T} accounted for approximately 60–65 % of total variance. Residual error σ^2_{R} contributed another 11–14 %. In contrast, encoder choice σ^2_{E} explained only 12–18 %, seed initialization σ^2_{S} 2–4 %, and fold assignment σ^2_{F} less than 1 %. Interaction terms—encoder × seed, encoder × fold, seed × fold—were effectively zero (Table 4).

Thus, although true-score variance dominates AES reliability, the nontrivial contributions of model architecture and random seed justify their explicit modeling in both G- and D-study phases.

3. Small-Only D-Study: Trading Seeds for Model Diversity

A full sweep of the six "small" transformers (ELECTRA Small (Discriminator), DistilBERT-base (uncased), DeBERTa-v3-small, GPT-2 (small), MiniLM-L6-uncased, MobileBERT-uncased) crossed with 1–3 seeds (399 designs with 354 of them obtained G values ≥ 0.85) revealed clear trade-offs between architectural diversity and seed replication (Table 5; Figure 7). Under a single-seed design, two- and three-encoder ensembles fail to reach G=0.85, but four small encoders just meet it (G≈0.851), and adding one or two more models raises reliability to approximately 0.868 and 0.882, respectively. Introducing a second seed yields larger gains: three encoders under two seeds already surpass G=0.85, and six encoders exceed 0.90. With three seed replications, even the minimal two-model pairing (GPT-2-small + MobileBERT) reaches G≈0.921; adding a third model pushes G to roughly 0.942, and four-model ensembles climb to about 0.952. Beyond four encoders, marginal improvements taper off—five- and six-model ensembles with three seeds achieve G≈0.958 and 0.962, respectively (Figure 8).

Figure 9 renders these results as smooth surfaces in the (number of encoders, number of seeds, G-coefficient) space. All three seed planes rise steeply from two to three encoders before plateauing, indicating diminishing returns on adding more models. Conversely, each additional seed shifts the entire surface upward by an almost constant amount, confirming that seed replication is a more efficient lever for boosting reliability once a modest ensemble is in place. In resource-constrained settings (e.g., limited GPU memory), two or three encoders with three seeds offer the fastest route to $G \geq 0.85$; where parallelism is abundant, larger ensembles can edge closer to the reliability ceiling but with diminishing payoff.

4. Mixed D-Study: Leveraging Medium-Sized Models

To explore hybrid configurations, we swept 5,401 designs combining 1–4 medium transformers, 1–6 small transformers, and 1–3 seed replications. As in the small-only study, an early-stop rule was enforced: for each medium–seed pairing, additional small models were added only until $G \geq 0.85$ was first reached (see Table 8). We then ranked each ensemble by a simple "resource" metric (number of models + number of seeds) to identify the most cost-effective solutions (Table 9).

Remarkably, a two-model hybrid—one medium encoder (BERT-Base), one small encoder (DistilBERT-base), and two seeds (resource = 4)—achieved $G \approx 0.895$, exceeding every small-only design at the same resource level. This single-medium + single-small configuration appears as the top entry in Table 9. Although adding more small models or seeds continued to raise G, the incremental benefit per added compute unit diminished rapidly.

Encouraged by these mid-range gains, we briefly evaluated large transformers under a second early-stop: once two large-inclusive ensembles surpassed $G \geq 0.95$, we terminated further exploration on cost-benefit grounds. The reliability uplift from large models (≈1–2 percentage points) did not justify their 5–10× higher FLOPs, memory footprint, and energy cost.

Together, Tables 8 and 9 illustrate that compact mixed ensembles—anchored by a single medium transformer, a single small transformer, and minimal seed replication—deliver the highest generalizability per unit of compute.

5 Discussion

The D-Study sweeps reveal a clear trade-off between reliability and computational cost. Small-only configurations must marshal at least five combined encoder-and-seed resources to surpass a generalizability coefficient of 0.85, whereas a compact hybrid ensemble of one medium transformer, one small transformer, and two seed replications achieves approximately 0.90 with only four resources. This efficiency frontier underscores that thoughtfully chosen model diversity and minimal replication can meet rigorous reliability thresholds while markedly curbing FLOPs, GPU memory, and energy consumption.

Our G-Study also confirms that large transformers—despite reducing encoder-facet variance by just one to two percentage points—incur disproportionately high compute and environmental costs, making medium-sized architectures the practical backbone for high-stakes scoring. By treating each essay (and thus each examinee) as the object of measurement, the mixed-effects framework captures both content and ability variance in a single term, aligning our dependability estimates with established psychometric practice. Moreover, entropy-stratified bucketing ensured that low-, mid-, and high-complexity texts contributed equally to variance-component estimation, guarding against bias from essay length or richness and validating the fairness of our reliability analyses.

Looking ahead, enriching our simple resource metric with actual FLOPs, GPU-hours, and energy use would enable truly multi-objective D-studies. Exploring adaptive or continuous stratification and designs with multiple essays per examinee could further disentangle content from ability variance and broaden applicability beyond holistic scoring.

6 Conclusion

This study integrates standardized Shannon-entropy stratification within a Generalizability-Theory framework to guide efficient AES ensemble design. A fully crossed G-Study (14 encoders × 3 seeds × 5 folds) quantified error sources across prompts, entropy strata, encoder models, and seed initializations. Analytic D-Study formulas then predicted generalizability coefficients for over 5,800 hypothetical ensembles without refitting models, revealing that compact hybrids of medium and small transformers with limited seed replication achieve target reliability at minimal cost. By balancing psychometric rigor with computational pragmatism, our approach offers a principled roadmap for deploying reliable, fair, and sustainable AES systems across diverse writing complexities.

References

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, *15*(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Brennan, R. L. (2001). *Generalizability theory* [doi:10.1007/978-1-4757-3456-0]. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

Bridgeman, B., Catherine, T., & and Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley and Sons, Inc. https://doi.org/10.3102/00028312011001054

Gao, X., Brennan, R. L., & Guo, F. (2015). *Modeling Measurement Facets and Assessing Generalizability in a Large-Scale Writing Assessment* (GMAC ® Research Reports, Issue.

Han, T., & Sari, E. (2024). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, *37*(4), 961-985.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218. https://doi.org/https://doi.org/10.1016/j.asw.2008.10.002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shavelson, R., & Webb, N. (1991). *Generalizability Theory: A Primer*. https://doi.org/10.1002/9781118445112.stat00068

Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (M. D. Shermis & J. Burstein, Eds.). Routledge.

Williamson, D., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

A Appendices

Category	Model	Hidden Size	# Layers	# Heads	Approx. Parameters	Max Seq. Length
Small	ELECTRA Small (Discriminator)	256	12	4	14 M	512
	DistilBERT-base (uncased)	768	6	12	66 M	512
	DeBERTa-v3-small	768	6	12	86 M	512
	GPT-2 (small)	768	12	12	124 M	1024
	MiniLM-L6-uncased	384	6	12	22 M	512
	MobileBERT-uncased	512	24	4	25 M	512
Medium	BERT-base-uncased	768	12	12	110 M	512
	RoBERTa-base	768	12	12	125 M	512
	Longformer-base-4096	768	12	12	149 M	4 096
	GPT-2 (medium)	1 024	24	16	355 M	1024
Large	BERT-large-uncased	1 024	24	16	340 M	512
	RoBERTa-large	1 024	24	16	355 M	512
	GPT-2 (large)	1 280	36	20	774 M	1024
	DeBERTa-v3-large	1 024	24	16	304 M	512

Table 1 General Information Comparison of All (14) Encoder Models used

Bucket	N	Mean	Range	SD
Low	451	1.131	0.909~1.237	0.048
2	479	1.176	1.100~1.276	0.040
3	457	1.216	1.125~1.331	0.038

Table 2. Entropy Bucketing Result using the Equal Variance Method

Prompt	Entropy Bucket		
	L	M	H
Car-free cities	12	32	152
Distance learning	95	80	43
Does the electoral college work?	90	78	37
Driverless cars	50	81	58
Exploring Venus	38	78	70
Facial action coding system	62	71	84
Summer projects	104	59	13
Total	451	479	457

Table 3 Essay Distribution in Entropy Bucket by Prompt

Bucket	σ^2_{E}	σ^2_{S}	σ^2_{F}	σ^2_{T}	σ^2_{ExS}	σ^2_{ExF}	σ^2_{SXF}	$\sigma^2_{Residual}$
Low	0.00479	0	1.34×10^{-7}	1.231	0	0	0	0.163
Mid	0.00019	1.7×10^{-10}	0	1.274	0	0	0	0.157
High	0.00051	5.4×10^{-6}	0	0.954	0	0	3.1×10^{-9}	0.164

Table 4 Variance Components and Ceiling-G Coefficients by Entropy Bucket

# Encoders	# Seeds	Mean G
2	3	0.872
3	2	0.860
4	1	0.851

Table 5. Minimal small-only configurations achieving G ≥ 0.85

Seeds	Encoders	Encoder Set	G-Coefficient
1	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.866
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.893
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.909
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.920
2	2	DistilBERT-base , MobileBERT-uncased	0.893
	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.922
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.937
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.946
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.951
3	2	GPT-2 (small) , MobileBERT-uncased	0.921
	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.942
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.952
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.958
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.962

Table 6. Best G ⩾ 0.85 Designs By Seed And Encoder Count in Small-encoder Ensemble

N of Encoders	Small Encoders	Seeds	G
1	DistilBERT-base	3	0.872
2	GPT-2 small; MobileBERT-uncased	3	0.921
3	DistilBERT-base; GPT-2 small; MobileBERT-uncased	3	0.942
4	DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MobileBERT-uncased	3	0.952
5	ELECTRA Small (Discriminator) ; DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MobileBERT-uncased	3	0.958
6	ELECTRA Small (Discriminator) ; DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MiniLM-L6-uncased; MobileBERT-uncased	3	0.962

Table 7 Best Small-encoder Ensemble Designs by Number of Models.

Medium Encoder	# of M Encoders	Small Encoder	# of S Encoders	# of seeds	G
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	GPT-2 (small)	1	2	0.894
BERT-base-uncased	1	GPT-2 (small)	1	2	0.894

Table 8. Top 5 G ≥ 0.85 Mixed Ensemble Designs with Fewest Resources

Medium Encoder	# of M Encoders	Small Encoder	# of S Encoders	# of seeds	G
BERT-base-uncased , RoBERTa-base , Longformer-base-4096 , GPT-2 (medium)	4	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , RoBERTa-base , Longformer-base-4096	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , RoBERTa-base , GPT-2 (medium)	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , Longformer-base-4096 , GPT-2 (medium)	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.967
BERT-base-uncased , RoBERTa-base	2	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.967

Table 9. Mixed Designs with Top 5 G-coefficients

Category	Model	Train Embedding	Test Embedding	Full 3×5 CV & Merge
Small	ELECTRA Small (Discriminator)	0 h 9 m 29 s	0 h 1 m 3 s	0 h 12 m 45 s
	DistilBERT-base	0 h 10 m 18 s	0 h 1 m 10 s	0 h 17 m 42 s
	DeBERTa-v3-small	0 h 31 m 53 s	0 h 3 m 30 s	0 h 41 m 14 s
	GPT-2 (small)	0 h 41 m 54 s	0 h 4 m 31 s	0 h 52 m 46 s
	MiniLM-L6-uncased	0 h 4 m 50 s	0 h 0 m 32 s	0 h 8 m 49 s
	MobileBERT-uncased	0 h 21 m 12 s	0 h 2 m 15 s	0 h 28 m 10 s
Medium	BERT-base-uncased	0 h 21 m 6 s	0 h 2 m 27 s	0 h 29 m 48 s
	RoBERTa-base	0 h 21 m 18 s	0 h 2 m 25 s	0 h 30 m 8 s
	Longformer-base-4096	5 h 40 m 48 s	0 h 38 m 3 s	6 h 24 m 32 s
	GPT-2 (medium)	2 h 2 m 50 s	0 h 13 m 14 s	2 h 24 m 24 s
Large	BERT-large-uncased	1 h 4 m 41 s	0 h 7 m 17 s	1 h 20 m 10 s
	RoBERTa-large	1 h 10 m 46 s	0 h 7 m 58 s	1 h 27 m 23 s
	GPT-2 (large)	4 h 29 m 16 s	0 h 31 m 11 s	5 h 11 m 33 s
	DeBERTa-v3-large	2 h 58 m 33 s	0 h 20 m 48 s	3 h 28 m 12 s

Table 10. Wall-Clock Times for Embedding and 3×5 Cross-Validation Scoring of 14 Transformer Encoders

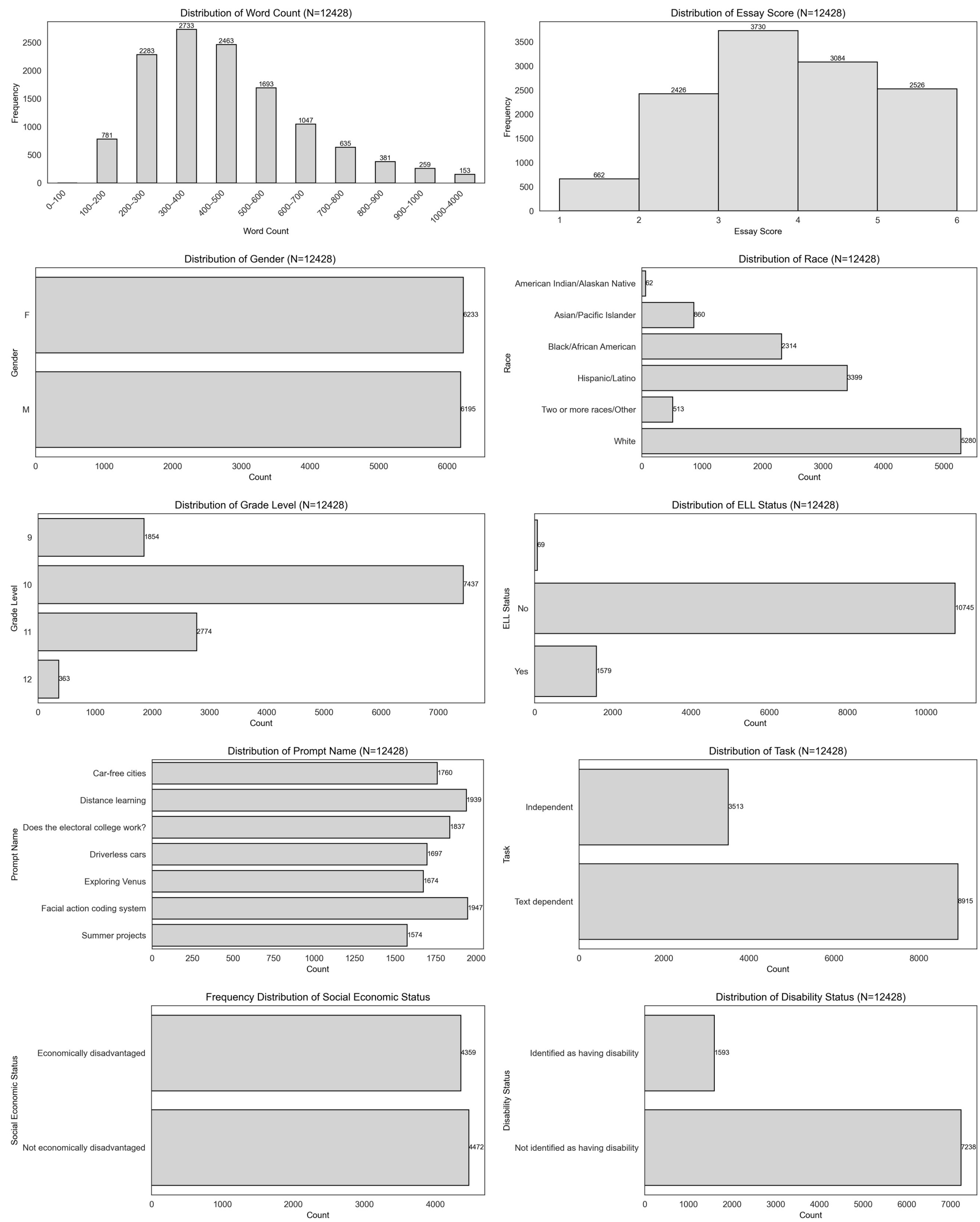


Figure 1. Descriptive Statistics of Essays in Training Set and the Test-takers' Demographic Information

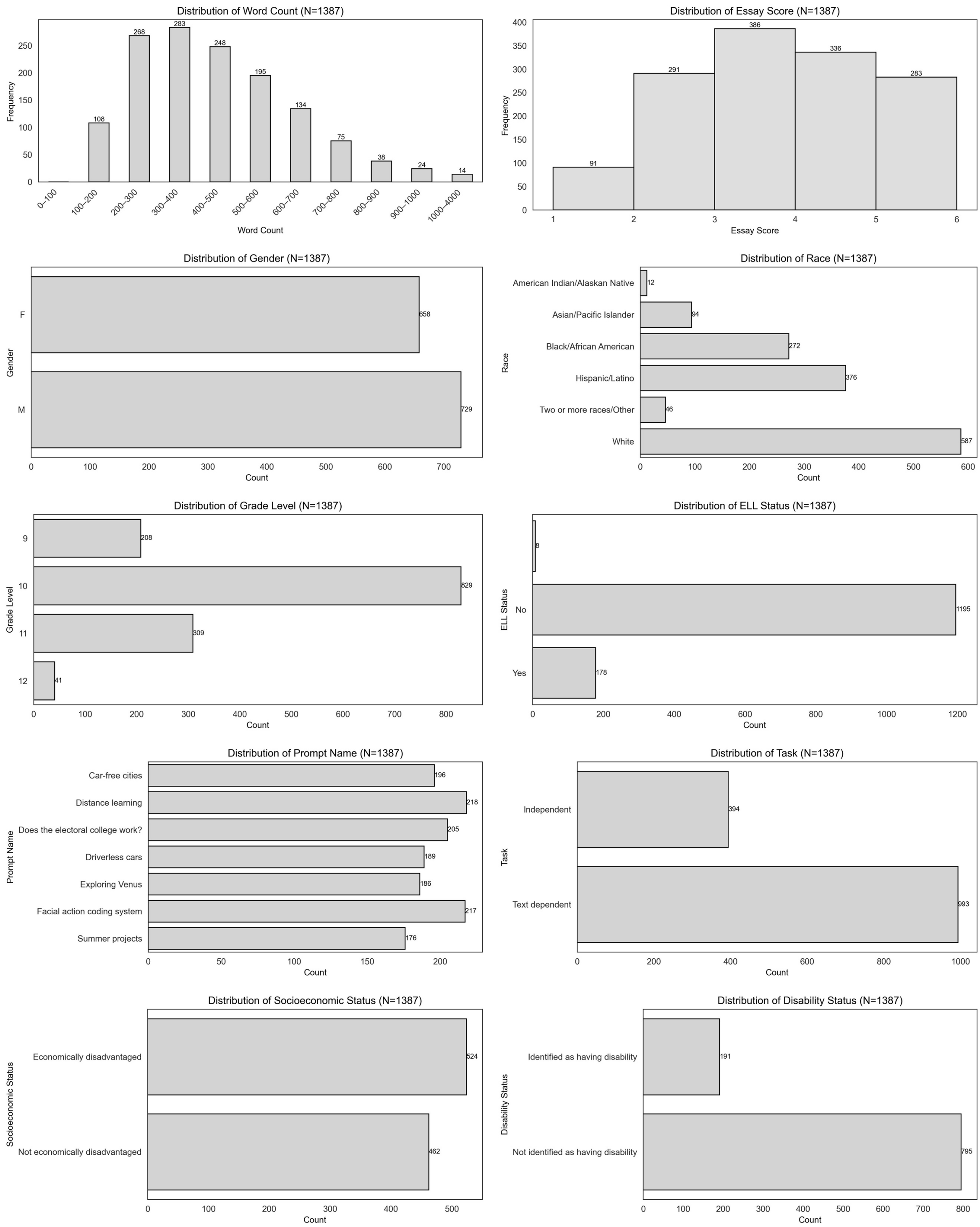


Figure 2. Descriptive Statistics of Essays in “Holdout_evaluation_set” and the Test-takers’ Demographic Information

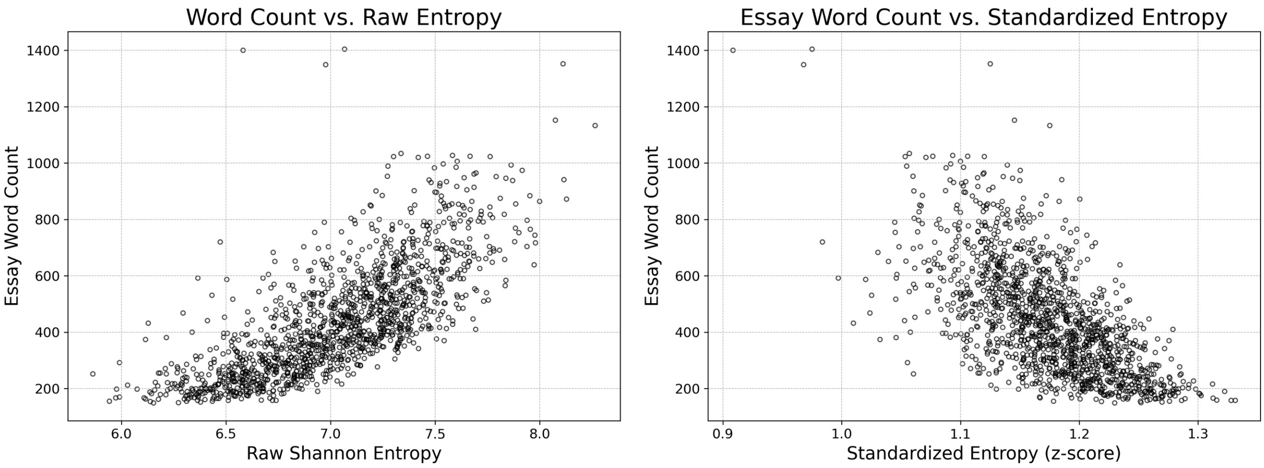


Figure 3. Comparison of Scatterplots of Word Count vs. Raw Shannon Entropy and Standardized Entropy

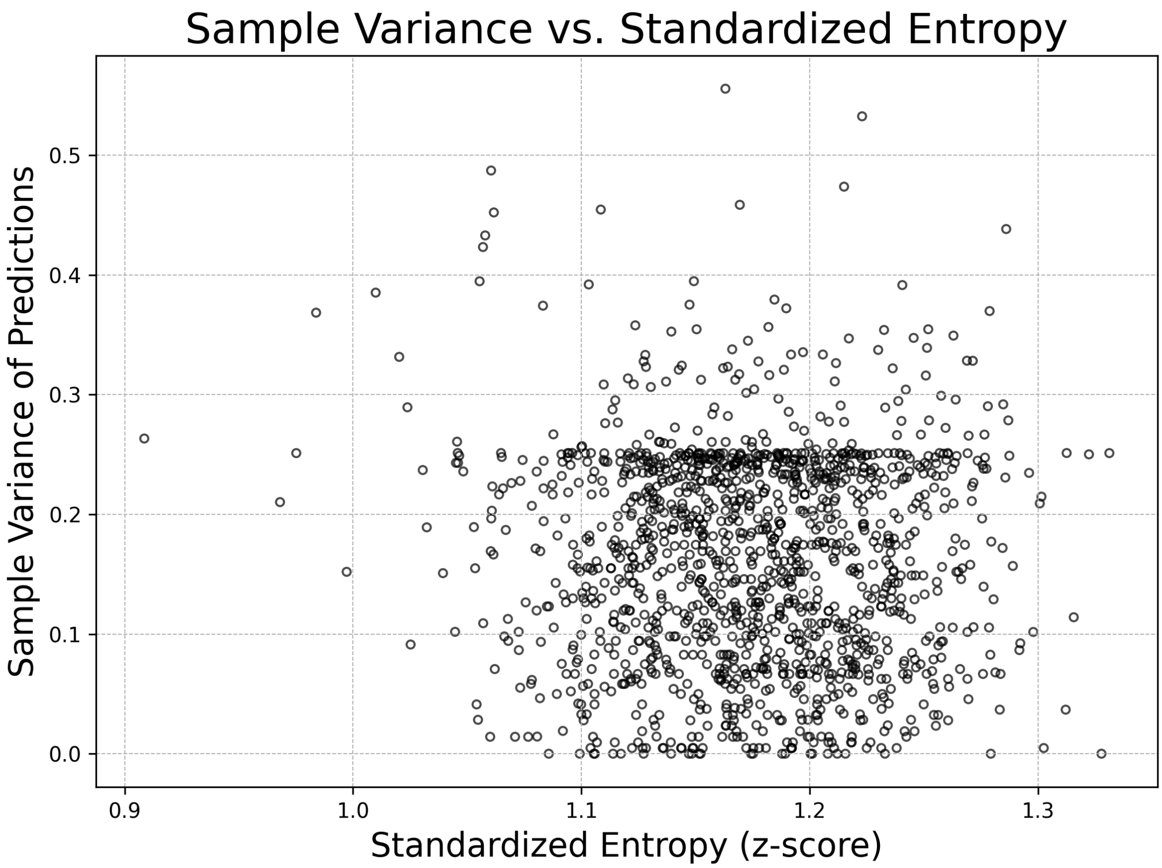


Figure 4. Scatterplot of Score Sample Variance and Standardized Entropy

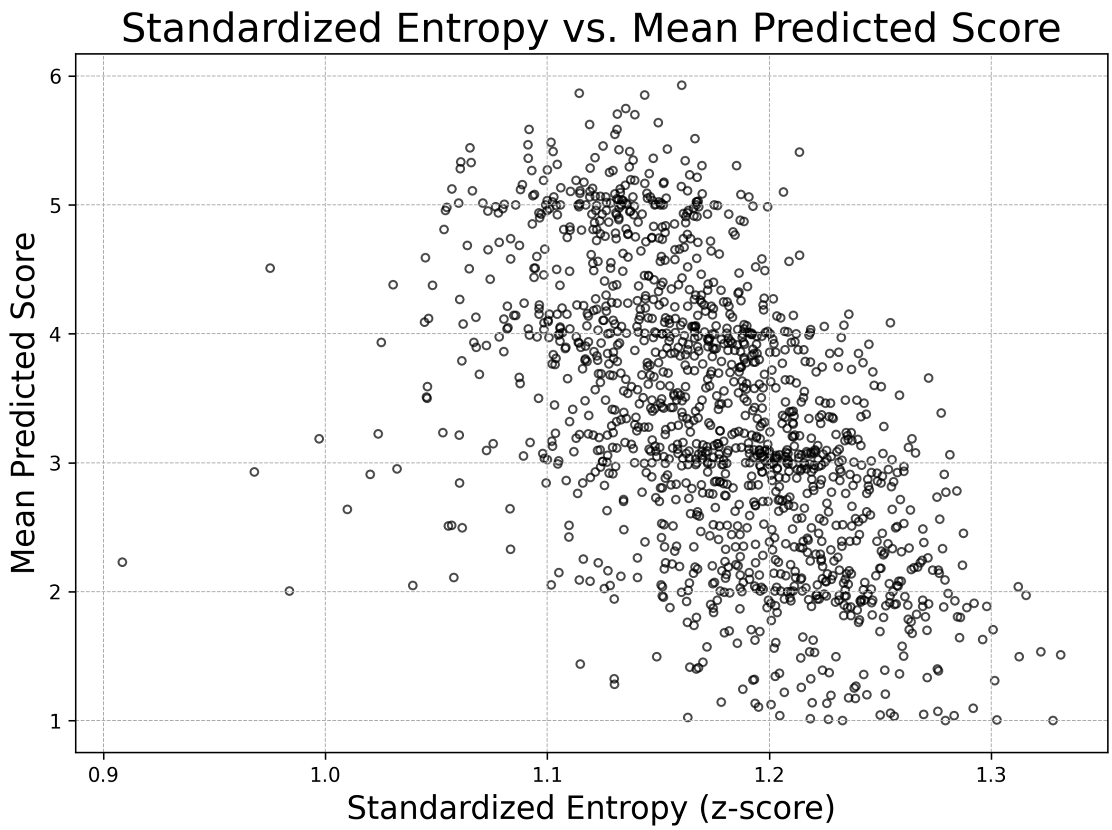


Figure 5. Scatterplot of Standardized Entropy vs. Mean Predicted Scores

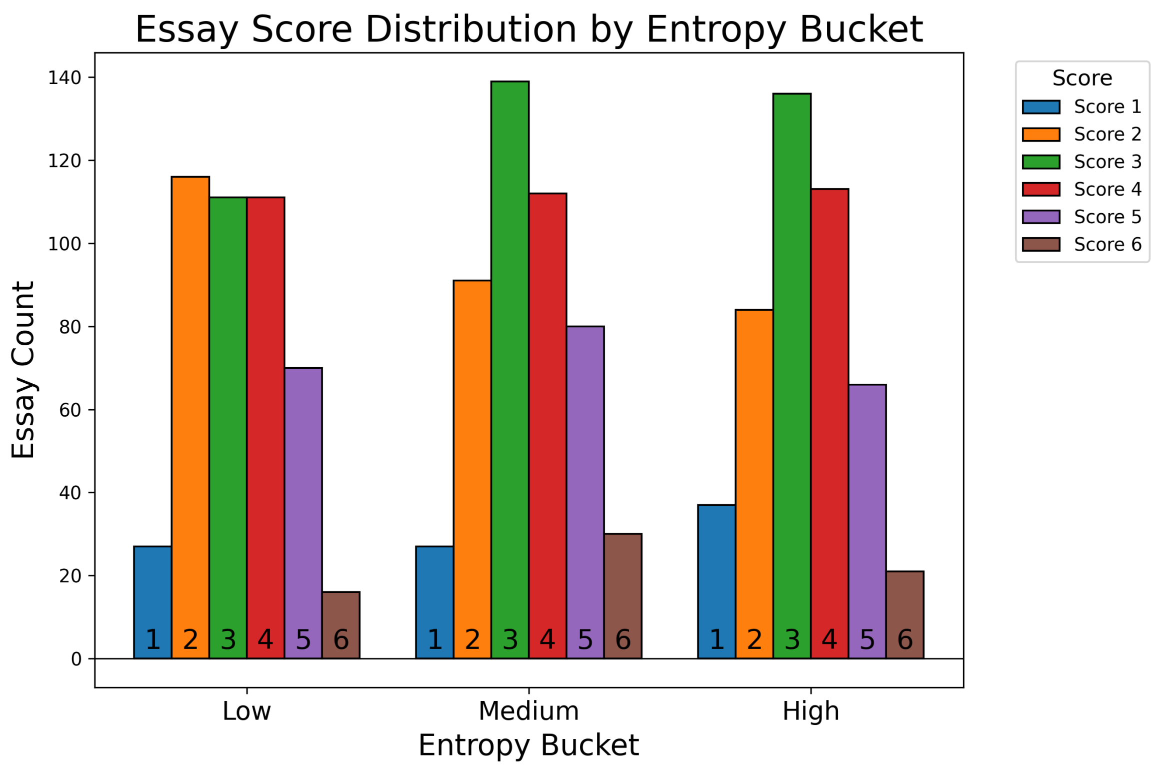


Figure 6. Histogram of Essay Score Distribution by Entropy Bucket

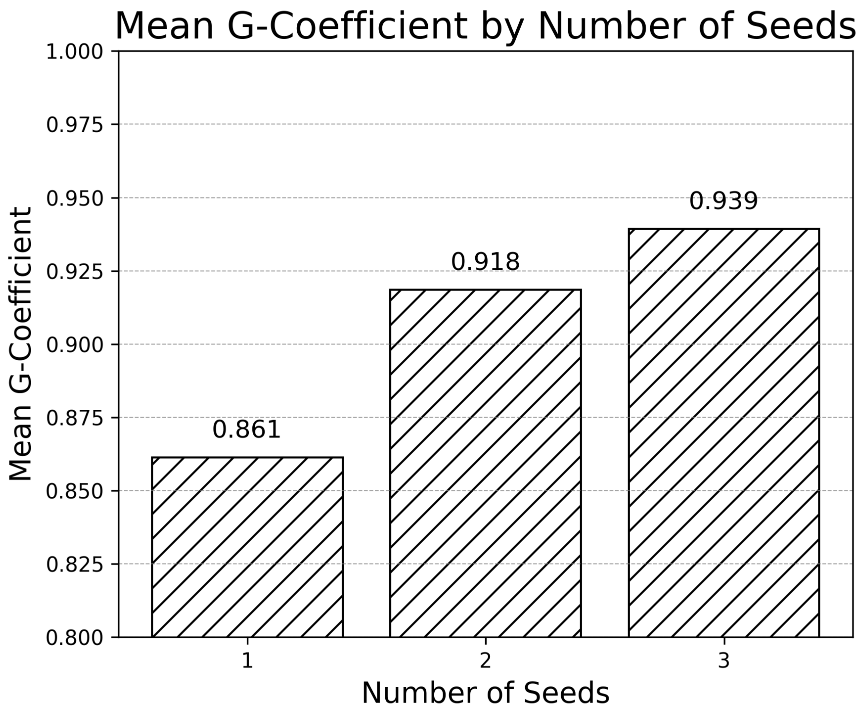


Figure 7. Mean G-coefficients by Number of Seeds in Small-encoder Ensemble Designs

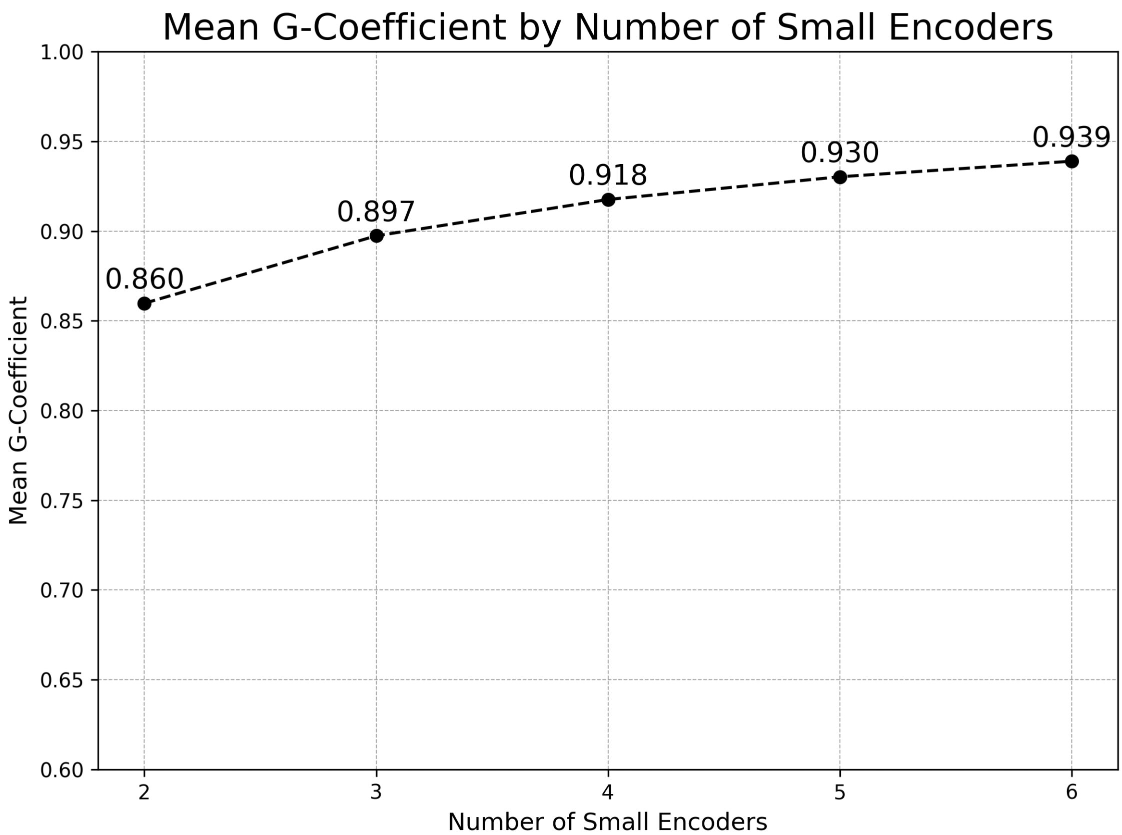


Figure 8. Mean G-coefficients by Number of Small-encoders in Small-encoder Ensemble Designs

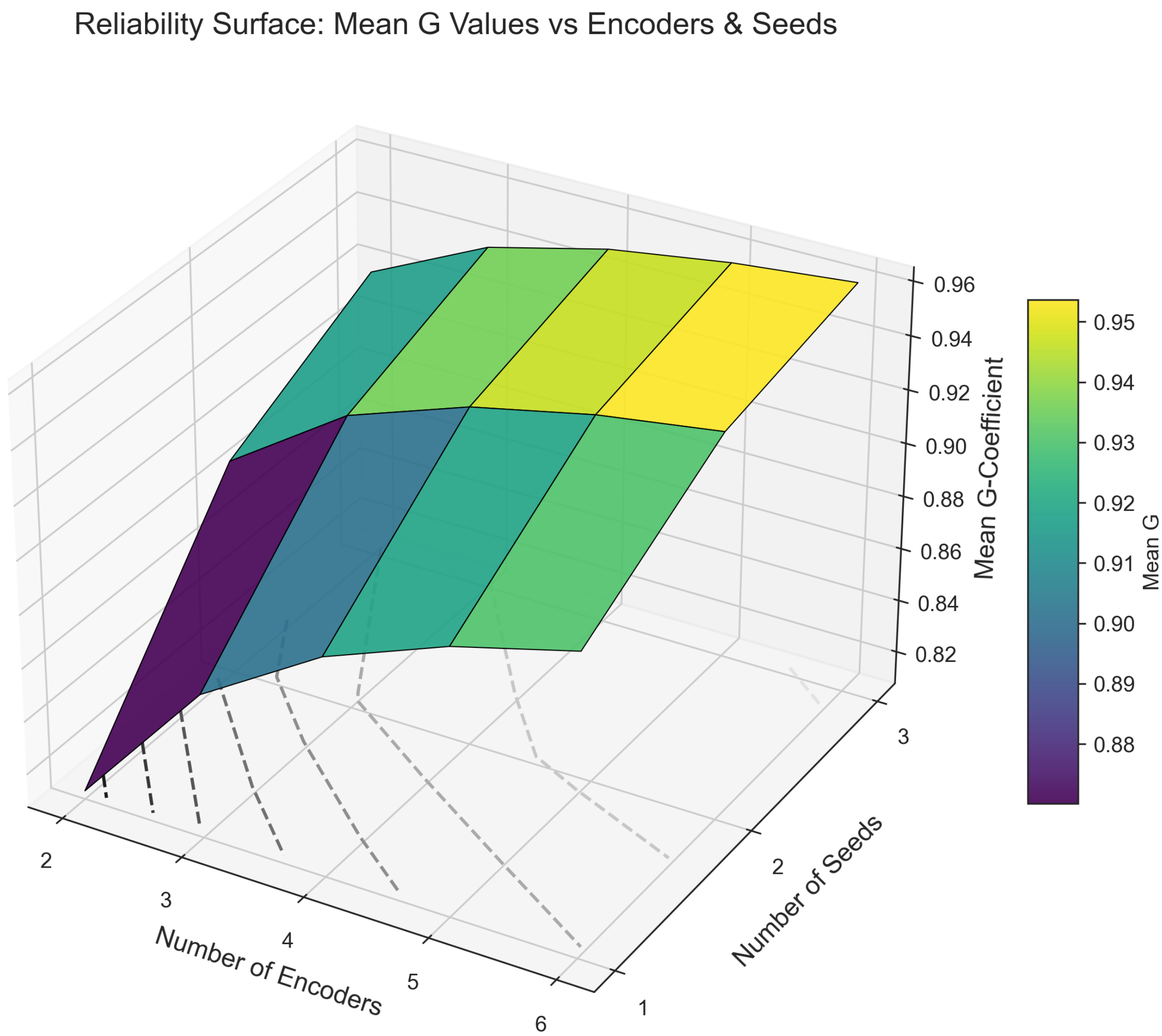


Figure 9. 3-D Plot of Mean C-cefficients vs. Number of Encoders and Seeds

Undergraduate Students' Appraisals and Rationales of AI Fairness in Higher Education

Victoria Delaney, Sunday Stein, Lily Sawi, Katya Hernandez Holliday
Department of Mathematics and Statistics
San Diego State University
San Diego, CA, USA
{vldelaney, sstein9540, lsawi4767, kdrew0853}@sdsu.edu

Abstract

To measure learning with AI, students must be afforded opportunities to use AI consistently across courses. Our interview study of 36 undergraduates revealed that students make independent appraisals of *AI fairness* amid school policies and use AI inconsistently on school assignments. We discuss tensions for measurement raised from students' responses.

1 Introduction

Proficiency with AI tools, particularly generative AI (GenAI), is becoming necessary for job market candidates (Bowen & Watson, 2024; Microsoft, 2024). To develop proficiency, students must be afforded continuous opportunities to learn how to use AI in ways that augment (rather than stymie) their learning and reflect competencies desired in the modern workforce. Some universities have approached this demand by becoming *AI-native*: giving each student access to a chatbot and encouraging AI use (Singer, 2025). AI nativity implies a vision of policy coherence; that students and instructors alike will use AI tools in complementary ways that foster rich, flexible modes of learning, do not undermine each other's goals, and offer consistent ways to measure learning as it relates to students' assessments (and ultimately, the value of their degrees).

This paper contributes a depiction of ethical questions and tensions that arise when various actors within higher education have inconsistent visions of AI in education, and thus, develop diverging ideas about fairness and academic integrity. Drawing from a subset of interview data featuring undergraduate students' uses of AI in problem solving, we found that the vast majority of students believed that fair AI use in school coursework depended on a number of factors, many of which pointed to conceptions of cheating that have become hard to measure when AI is integrated inconsistently into coursework (Lee et al., 2024). Their confusions may create tensions for instructors, who have their own visions of how AI should be used on assignments, and for school administrators, who may expect that students take up AI tools for career and workforce development.

As issues of AI use in decentralized (Weick, 1976) university systems continue to surface (Dabis & Csáki, 2024; Goodier, 2025), we urge educators to pause and consider how such tools force students to reconfigure their judgments of what is fair and how learning is measured. These inconsistencies matter in educational measurement, as how students demonstrate learning on school assignments is in part a function of the tools that they use (Engeström, 2014). We explore these issues and answer:

RQ1: To what extent do students appraise AI use on school assignments as "fair"?

RQ2: What rationales do students give to justify whether AI use on school assignments is fair?

2 Background and Related Work

2.1 AI and Higher Education

The advent of GenAI chatbots (ChatGPT, Gemini, Claude), stirred mixed reactions in higher education. While some institutions initially sought to regulate or ban student access, others encouraged AI for teaching and learning (An et al., 2025). Nonetheless, trends suggest growing employer interest in hiring AI-literate workers (Microsoft, 2024), adding pressure on universities to equip students with AI knowledge and experience that aligns with workforce demand.

To address these demands, many institutions, including the Universities of Oxford, Arizona, Maryland, and Texas at Austin, as well as the entire California State University system (OpenAI, 2024;

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 329–336

CSU, n.d.), have rolled out agreements with OpenAI to provide both students and employees with the advanced capabilities of ChatGPT Edu, implying usage expectations. However, there seems to be no agreement on productive use of AI.

McDonald et al. (2025) analyzed AI policies published by 116 US universities. They found that, while most (N=73, 63%) universities provided guidance for classroom use, encouraging adoption, guidance often focused on writing activities. Students and instructors within STEM fields, therefore, faced absent or vague recommendations at best. These policy inconsistencies may introduce wide variability in GenAI use among both instructors and students across disciplines, raising questions about fairness and educational value.

2.2 AI and Ethics in Higher Education

Although the surge of AI prompted discussions around transparency in data processing, hallucination-induced misinformation, academic fraud, and algorithmic bias (Memarian & Doleck, 2023; Pérez & Mattison, 2025; Zheng, 2024), institutions have yet to directly address the equally important issue of AI fairness. By *AI fairness*, we mean students' abilities to access AI, use it skillfully, and obtain outcomes that reflect their skill (Wang et al., 2024, p. 3). Wang and colleagues describe first, second, and third-order AI-divides that could result if components of their fairness definition are unmet. As students gain widespread access to AI tools, whether it be through personal accounts or institutional licenses, scholars must expand ethical discussion on what constitutes fair and appropriate use among students.

Ethical discussions about fair AI use in education are particularly consequential in cases where instructors integrate AI into their courses differently (Delaney et al., 2025). Some courses may intentionally integrate tools, such as discussion platform PackBack (Lantz et al., 2022), directly into student dashboards. Such tools provide structured opportunities to engage with AI in ways that support writing, self-reflection, and learning. In these cases, AI use is normalized as part of the learning environment. By extension, AI use across students is likely to exhibit less variation (in frequency and types of use) in AI-integrated courses than courses which have AI policies but do not incorporate AI tools directly. These courses may leave decisions about its use to instructor discretion. As a result, students may shoulder a larger ethical burden in courses where the teachers' policies conflict with their own beliefs, values, and ideals about the purpose of higher education.

Instructors for these courses may include AI-use statements in their syllabi either disallowing it or asking students to disclose when and how they use AI tools. Alternatively, an instructor may simply ignore AI, leaving the decision up to students' own judgement. In either case, AI use in such courses is unregulated, unobserved, and often undetectable (Ardito, 2024). This brings up questions about equitable awareness and AI skill development among students (Arum et al., 2025) but also leaves deeper issues of fairness unresolved.

Furthermore, this variation raises ethical questions about institutional consistency and how institutional AI-integration goals may infringe upon instructors' individual pedagogical values or beliefs about how these tools should or should not be used. When AI policies are lacking or unclear, students, instructors, and institutional leaders may hold contradicting views of what ethical use looks like. Moreover, instructors may unintentionally develop inconsistent rules or expectations for AI use among students, further complicating student beliefs about fairness.

Inconsistent AI policies among instructors and AI uses by students complicate measurement and assessment. If, for instance, half of the students in a writing course use an AI chatbot for their final essay, and half do not, scoring the final essay will become internally inconsistent, because it is likely unclear from the grader's perspective (1) who used AI and (2) to what extent and for what purpose users leveraged AI toward the final essay. For half the class, grades are a measurement of knowledge applied and distilled in the essay, and for the other half, grades are a partial measurement of learning and partial measurement of AI savviness. Thus, inconsistent AI policies increase threats to validity and assessment measurement precision (Zheng et al., 2025), and are worth examining in more detail.

3 Methods

3.1 Research Context and Participants

This study was conducted at a large university that serves mostly undergraduate students in the U.S. The university purchased chatbot subscriptions for all students and faculty. Faculty were encouraged to create their own AI course policies and syllabus statements that detailed if and how AI should be used in their courses. Instructors were given flexibility to design AI policies as they pleased and ban AI use if they deemed appropriate.

We recruited 36 undergraduate students from various majors to study AI use during problem solving tasks. We hung flyers around campus in publicly-available locations, solicited open participation calls through email, and encouraged students to sign up using a QR code. The analysis we present in this paper emerged when variations of students' conceptions of AI fairness on school assignments during the study pre-interview arose with a higher frequency than we had anticipated.

3.2 Study Design and Data Collection

We focus on one subcomponent of a larger clinical interview study (diSessa, 2007) that explored how undergraduates used AI while problem solving. We first interviewed each student about their AI use and beliefs. To answer RQ1 and RQ2, we look at participants' responses to question six from our protocol: *"Do you think using AI on school assignments is fair?"* Importantly, members of the research team did not define fairness for the participant. Rather, we responded by deflecting the question (*"What do you think it means?")*

Our data consist of the 36 responses from participants, audiorecorded and transcribed by the research team. For anonymity, we refer to participants as P#, where # represents order (e.g., P4 was the fourth participant). In the event that the researcher did not understand the participant's initial response, or the participant did not appear to answer the question, we asked 1-2 probing questions until we understood their position. For example, P32 initially answered, *"Yeah, I'm kind of conflicted about it, because a lot of my math professors are aware of it, so they make changes."* The interviewer realized that P32 did not give an appraisal, and responded, *"Oh wow, that's interesting, but whether it's fair?"* P32 then said, *"Everyone's aware of it now, so it's becoming more fair…but if not everyone's using it, then I guess not."* We were satisfied that P32's new response gave an appraisal and moved on.

All interview recordings were autotranscribed by otter.ai. The initial error rate was around 9%, typical for recordings taken in consistent, quiet settings (Tran et al., 2023). Three members of the research team cleaned the transcripts to account for cross-talk discrepancies, speaker assignment errors, and errors that arose from abbreviations or acronyms (Matters & Shapiro, 2022).

3.3 Data Analysis

The unit of analysis encompasses students' responses to, "*Do you think using AI on school assignments is fair?",* including follow-up clarifying questions. Three researchers coded the data using an iterative process. During Phase 1, we individually coded participants' appraisals (RQ1) into "yes" "no" and "other." When meeting as a whole group to discuss coding agreements, we realized that a majority of participants did not give a clear yes/no answer. Rather, they gave some form of "it depends," and elaborated on what their appraisal depended on. During Phase 2, we re-coded the data into six appraisal categories based on what we learned in Phase 1: "yes," "yes/it depends," "it depends," "it depends/no," "no," and "unclear" (see Figure 1). We continued coding until we reached internal agreement on the definitions and codes applied (Cornish et al., 2016).

Following coding of appraisals, three researchers inductively (Saldaña, 2021) coded participant rationales (RQ2) in two rounds (Phases 3 and 4). During Phase 3, we induced rationales, converged on the definition of each, and discussed how many rationales should be assigned per participant. We concluded that some participants gave multiple, independent examples of fairness and maintained different rationales per context. For instance, P34 explained that she would not use AI to help her with math problem-solving, but would use it to help her write an essay outline. We therefore decided to apply multiple rationales per participant if the participant introduced a new context or idea and a different rationale than in a previous explanation. In Phase 4, we re-coded the data, resolving disagreements as needed and re-visiting the data to ensure internal consistency.

Figure 2 in the findings shows all appraisal codes, rationales, and code counts.

4 Results

4.1 RQ1: To what extent do students appraise AI use on school assignments as "fair?"

Figure 1 displays the distribution of AI fairness appraisals in our sample. 41.6% of participants (n = 15) responded that AI fairness depended on a number of contextual and organizational factors.

Only six students (16.7%) responded that using AI on school assignments was fair, and five students said that using AI on school assignments was not fair (13.9%). Two students (5.6%) gave unclear answers that did not contain an appraisal of fairness. We coded those instances as "unclear." P4, for instance, described AI in relationship to her ability to learn, but did not address fairness:

> *...It is kind of inevitable at this point…every student is going to be using AI…if you're not, you're just doing twice as much work as all your classmates. But…I don't think it's good. I think at the end of the day I… am glad I made it this far in my education before I was introduced [to GenAI] because if it was introduced earlier, I would never have learned how to think for myself. (P4, 10:52)*

Eight participants (22.2%) gave responses that contained multiple appraisals: they determined AI use on school assignments fell into multiple categories of "yes," "it depends," or "no." As an illustrative example, we observed P34 shift her explanation between "it depends" and "yes" as she considered the practicality of restricting AI:

> *I don't know. I use it sometimes, so I guess I should be saying fair. But I do feel guilty...Not really for math, because I'm still doing all the work. But for writing, I feel like it should be my organic thoughts. Although, at this point, now that it's so available, it's like, especially for young people, it's kind of impossible to tell them not to use it. And so, in that sense, I think a lot of teachers have to understand that they probably are going to use it, and they need to give guidelines for how to use it in a good way. (P34, 12:07)*

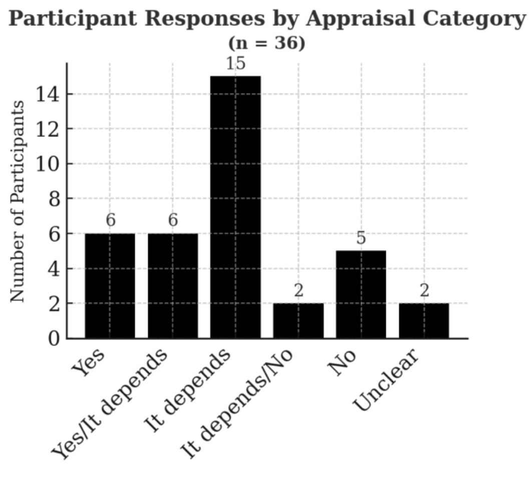


Figure 1: Distribution of undergraduate appraisals

P34, a junior-year civil engineering major, shifted from "it depends," where she reasoned that epistemic differences in school assignments influence her decision to use AI, to "yes" when thinking about the organization of technology in schools. Shifts in student appraisals of fairness suggest that (1) undergraduates may perceive fairness as a continuum between fair and not fair rather than discrete categories, and that (2) fairness appraisals can shift based on situation and context (diSessa et al., 2004).

4.2 RQ2: What rationales do students give to justify whether AI use on school assignments is fair?

Figure 2 shows students' AI fairness appraisals in relation to their rationales. Students asserted "yes," it is fair to use AI, for tutoring (n=2, 5.6%) or learning (n=1, 2.8%), and justified their beliefs by pointing to the ubiquity of AI (n=6, 16.7%) and the endorsement of AI by their institution (n=2, 5.6%). Additionally, some (n=3, 8.3%) expressed confidence that the designs of courses (e.g., in-class assessments, presentations, and projects) would assess students fairly, because assessments are conducted separately from homework assignments. P12 articulated:

> *My friends that were using [AI] were getting such good grades, but they weren't learning anything, or doing [homework] themselves. But when it came to the test, they were doing worse, and I was doing better…So I don't really think it's unfair anymore. (P12, 9:15)*

Students who believed that using AI on school assignments was unfair described issues of unequal access, either between students in their own courses (n=4, 11.1%) or compared to students in previous generations (n=2, 5.6%). Only one student reasoned that AI use was unfair because it hindered learning: *"I wouldn't say it's fair because then everybody's just going to copy paste what's in AI...and not use their mind."* (P18, 14:04). This student indicated earlier in the interview that she participated in higher education before AI (since 2017). It is possible that she compared AI use at present to her experience in 2017, before chatbots.

Most (94%, n=34) participants gave at least one rationale in the "it depends" category, suggesting that judgements of fairness depend on multiple factors including purpose of use, course instruction, and capabilities of free chatbots. Participant P28 offered: *"I would [not say using AI on school assignments is fair] because you get a limited amount of data use, like a phone plan."*

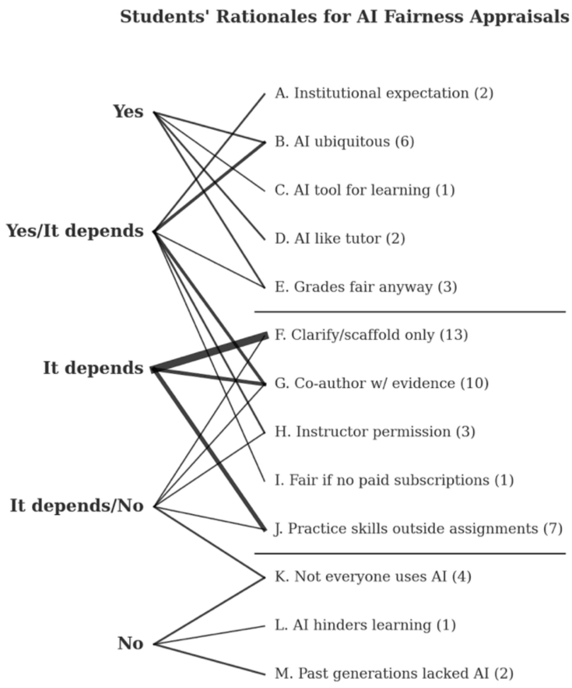


Figure 2: Distribution of rationales for AI fairness

Participants who said "it depends" cited socio-technical tensions in their courses. The first is about when AI is allowed to be used as a tutor: for practicing skills (e.g., learning a foreign language by conversing with AI) versus for tutoring on the content of graded assignments. P26 explained:

> *I think [using AI is] fair to a point where…new topics are introduced…My professor likes to put new topic questions on my homework. Which makes it difficult, because I don't know what's happening ...But personally, I don't really like [AI] too much. Especially writing...teachers say that it removes your voice… I would rather get docked the points.*

A second tension we observed is uncertainty about what forms of assistance AI should give on graded course assignments. Students' descriptions can be categorized by: (1) AI as a clarifier or scaffolder; (2) AI for overcoming impasses (3) AI as a deliverable co-author, and (4) AI as a personal tutor for anything unless their teacher specified a policy otherwise. We give illustrative examples:

> *When you're writing, some people will go give me ideas for writing a thesis on why recycling is bad. And [AI can] give you a couple ideas. And then you can write your own thesis based on some of the ideas or the wording or using it to check for grammar. (P5, a math teaching major, on scaffolding)*

> *I think if you're given a problem set… if you're stuck on a problem, then it's very helpful and I think it's valid to use. But I think if you're using it to write your whole…code program for you, then…it's like, defeating the purpose of learning. (P13, a computer science major, on overcoming an impasse)*

> *What I'll do is write an essay on my own, and then I'll ask [AI] to revise it and make it flow nicely…and make it sound more professional. And I think that's 100% fair. Like, I'd be scared of just putting in the prompt and being like, write an entire essay. But…if you can...make it your own words, then I think that's fair enough. (P11, a marketing major, on AI co-authoring)*

> *It depends how it's used. If the students aren't grasping the information, it's not really helping them in any way. But if it's in a way where they're...using it as a personal tutor, that would be fair. (P8, an engineering major, on AI as a tutor)*

Finally, some participants wished for instructors to clarify expectations on AI use in relation to academic integrity, even in light of a university-wide policy: "*tell the students what they expect to see in the class…if they don't want to see AI use, they should tell the students that.*" However, this same participant later acknowledged that even given clear expectations, some students may not adhere to AI classroom policies: "*it's becoming such a powerful resource that students are going to use it regardless of what a professor says…we should try to incorporate it into classrooms and use it as a good resource. Not trying to ban it, because I don't think that's going to work*." His sentiments seem to reflect university-wide AI integration goals which encourage students and instructors alike to engage with AI for academic work.

5 Discussion

After studying 36 undergraduate students' appraisals and rationales for AI fairness on school assignments, we found a broad range of opinions, use cases, and ethical considerations that formed the foundations of students' judgments. Most were unsure if they should use AI on school assignments and varied in the socio-technical decision parameters they drew from when deciding when and how they would use AI. Although some expressed a desire for guidance from their instructors about what to do, others made determinations based on internal factors, such as whether they personally believed AI would help them learn, or how accessible AI chatbots are becoming as a publicly consumable resource. We hypothesize that AI policy misalignments between the university and individual instructors may have caused students to become confused about AI fairness because they needed to navigate each course policy on a case-by-case basis.

Beyond AI policies, many students in our study drew from their own interpretations of fairness. Their appraisals did not draw from ethical considerations raised about AI by academia (e.g., model bias, data privacy, and data security), but rather their personal beliefs about whether AI tools helped them learn and the appropriateness of using novel technologies to "do school" (Pope, 2001). This suggests that in the wake of a decentralized university-wide policy stance on AI use in school, most students will do what "feels right" to them. This implies that while some students will not use AI at all, some may use it for everything, and some will make independent determinations based on a number of contextual factors (that should be studied in more detail in the future).

"Choose-your-own-adventure"-style AI fairness appraisals made by the undergraduates in our sample have implications for measurement and assessment. Particularly in courses with large student enrollments, instructors do not have time to peruse individual assessments and evaluate if and how AI was used. They likely do not have the resources to evaluate if students who use AI to complete course assessments score differently than students who do not. This implies, on the extreme end, threats to assessment precision and validity (Zheng et al., 2025). That is, a student who uses AI to complete an assessment from end-to-end could receive a better grade than a student who took time to learn the material but did not show mastery on the same assessment. In this case, the knowledge and skill gained (*purpose of learning*) by human students is measured unfairly against the perceived quality of a submitted deliverable (*purpose of work*) by a human extracting information from a large language model trained on most of the corpus of written human texts. Furthermore, if undergraduates can and do complete assessments programs with AI chatbots, it raises questions about the value and purpose of college degrees.

5.1 Limitations

Our study is limited in several ways. First, the broader purpose of our research was not to study fairness in a large population of students. As such, the deductive coding used reflects our intersubjective agreement (Krippendorff, 2018) on fairness in a small sample. Our aim was to illustrate through student narratives what fairness with AI looked like rather than to make claims about statistical power and generalizability (our study has neither of those elements). Nonetheless, larger samples and surveys with statistically validated constructs should be used to study how AI is appraised writ large (e.g., Paik et al., 2025).

Another limitation is that the question, "*Do you think AI use on school assignments is fair?*" was placed sixth in our interview protocol. Participants' answers to the preceding five questions could have influenced their appraisal (diSessa et al., 2004). We asked the first five questions to gain a baseline of students' AI beliefs and uses (e.g., question 1, "*How do you approach problem solving?*"; question 3, "*How, if at all, do you use AI with problem solving?*"). However, sometimes students gave off-topic responses that broached fairness. It is possible that these answers cued *a priori* conceptions of fairness that influenced the answer to the question that we ultimately studied.

5.2 Concluding Recommendations

While AI tools will continue to (rapidly) evolve and pose uncomfortable questions about the nature and fairness of learning, we put forth recommendations from our study. First, universities and instructors should have aligned AI use policies. This stands in opposition to policy flexibility recommended by An and colleagues (2025), and we acknowledge that faculty autonomy will make this difficult to achieve in practice, but our findings show that students rely on internal decision factors in absence of consistent guidance. Second, we recommend that instructors approach every course assessment with the perspective that at least one student will use AI, and ask themselves, "*Will students still achieve my desired learning outcomes with AI? Will their grades fairly reflect the desired learning outcomes?*" If not, they should consider redesigning their curriculum and assessments (Xie et al., 2024). Finally, we recommend that adults working at universities guide students to reflect on and assess their learning in a world with AI.

References

An, Y., Yu, J.H., James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, *22*(10). https://doi.org/10.1186/s41239-025-00507-3

Ardito, C. G. (2024). Generative AI detection in higher education assessments. *New Directions for Teaching and Learning*, 1–18. https://doi.org/10.1002/tl.20624

Arum, R., Calderon Leon, M., Li, X., Lopes, J. (2025). ChatGPT Early Adoption in Higher Education: Variation in Student Usage, Instructional Support, and Educational Equity. *AERA Open*, *11*. https://doi.org/10.1177/23328584251331956

Bowen, J. A., & Watson, C. E. (2024). Teaching with AI: A practical guide to a new era of human learning. *Johns Hopkins University Press*. https://doi.org/10.56021/9781421449227

California State University. (n.d.). *AI tools*. CSU AI Commons. https://genai.calstate.edu/ai-tools

Dabis, A., & Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI. *Humanities and Social Sciences Communications, 11*, Article 1006. https://doi.org/10.1057/s41599-024-03526-z

Delaney, V., Adisa, I. O., Mah, C., & Lee, V. R. (2025). Teaching high school students about generative AI: Cases of teacher lesson design. *The Journal of Educational Research*, 1–16. https://doi.org/10.1080/00220671.2025.2510415

diSessa, A. A. (2007). An example of "how to conceive knowledge" and its fit with the conceptual change literature. In S. Vosniadou, A. Baltas, & X. Vamvakoussi (Eds.), *Re-framing the conceptual change approach in learning and instruction* (pp. 41–67). Elsevier.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*(6), 843–900. https://doi.org/10.1016/j.cogsci.2004.05.003

Engeström, Y. (2014). Learning by expanding: An activity-theoretical approach to developmental research (2nd ed.). *Cambridge University Press*. https://doi.org/10.1017/CBO9781139814744

Goodier, M. (2025, June 15). Revealed: Thousands of UK university students caught cheating using AI. *The Guardian*. https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Lantz, J. L., Liu, J. C., & Basnyat, I. (2022). Piloting artificial intelligence (AI) to facilitate online discussion in large online classes: A case study. In *Cases on innovative and successful uses of digital resources for online learning* (pp. 204-222). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-7998-9004-1.ch009

Lee, V. R., Pope, D., Miles, S., & Zárate, R. C. (2024). *Cheating in the age of generative AI: A high-school survey study of cheating behaviors before and after the release of ChatGPT. Computers and Education: Artificial Intelligence, 7, 100253*. https://doi.org/10.1016/j.caeai.2024.100253

Mathur, A., & Shapiro, B. R. (2022). Interactive transcription techniques for interaction analysis. *In C. A. Chinn, E. Tan, C. K. Chan, & Y. Kali (Eds.), Proceedings of the 16th International Conference of the Learning Sciences (ICLS) 2022* (pp. 19–26). International Society of the Learning Sciences. Available at https://repository.isls.org/bitstream/1/8993/1/ICLS2022_19-26.pdf

Memarian, B., Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*(100152). https://doi.org/10.1016/j.caeai.2023.100152

Microsoft, LinkedIn. (2024). *2024 Annual Work Trend Index*. Microsoft Source. https://news.microsoft.com/annual-wti-2024/

OpenAI. (2024). *Introducing ChatGPT Edu: An affordable offering for universities to responsibly bring AI to campus*. OpenAI For Education. https://openai.com/index/introducing-chatgpt-edu/

Pope, D. C. (2001). *Doing school: How we are creating a generation of stressed out, materialistic, and miseducated students*. Yale University Press.

Pérez, J. M., Mattison, T. S. (2025). Academic Fraud in the Use of Generative Artificial Intelligence (GenAI) for Faculty Promotion and Tenure. *International Journal of Higher Education*, *14*(2). https://doi.org/10.5430/ijhe.v14n2p35

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE Publications.

Singer, N. (2025, June 7). Welcome to campus. Here's your ChatGPT. *The New York Times*. https://www.nytimes.com/2025/06/07/technology/chatgpt-openai-colleges.html

Tran, B. D., Mangu, R., Tai-Seale, M., Lafata, J. E., & Zheng, K. (2023). Automatic speech recognition performance for digital scribes: A performance comparison between general-purpose and specialized models tuned for patient–clinician conversations. In *AMIA Annual Symposium Proceedings 2022 (pp. 1072-1080)*. American Medical Informatics Association. https://pmc.ncbi.nlm.nih.gov/articles/PMC10148344/

Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & de Vreese, C. H. (2024). The artificial intelligence divide: Who is the most vulnerable? *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241232345

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly,* *21*(1), 1–19. https://doi.org/10.2307/2391875

Xie, B., Sarin, P., Wolf, J., Garcia, R. C. C., Delaney, V., Sieh, I., Fuloria, A, Dennison, D. V., Bywater, C., & Lee, V. R. (2024). *Co-designing AI education curriculum with cross-disciplinary high school teachers. Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23146–23154. https://doi.org/10.1609/aaai.v38i21.30360

Zheng, Y., Huang, S., Nydick, S., & Zhang, S. (2025). MxML (Exploring the Relationship between Measurement and Machine Learning): Survey of the Measurement Community. *Chinese/English Journal of Educational Measurement and Evaluation*. https://doi.org/10.59863/GVZE8492

AI-Generated Formative Practice and Feedback: Performance Benchmarks and Applications in Higher Education

Rachel Van Campenhout, Michelle Clark, Jeffrey S. Dittel, Bill Jerome, Nick Brown, and Benny G. Johnson
VitalSource

Abstract

Integrating formative practice questions with text content is a highly effective learning method. Millions of AI-generated formative practice questions, embedded in thousands of publisher e-textbooks, are now available to students in higher education. This paper reviews findings from a multi-year research program to synthesize performance benchmarks for automatically generated questions and feedback derived from large-scale student interaction data. In addition, we report classroom-based applications that demonstrate how these questions can support learning when integrated into instruction. A central contribution of this review is to identify barriers to effectively scaling student engagement with formative practice, identifying both the successes of automatic question generation systems and the persistent challenges that must be addressed to maximize their potential for classroom impact.

1 Introduction

Formative practice has long been known to be highly effective for learning for students of all ages, but especially struggling students [1, 2]. Research studying the relationship between integrating formative practice with expository content and learning outcomes in digital learning environments found that doing practice was an average of six times more effective for learning than just reading [3, 4]. Called the doer effect, this learning science principle was also shown to have a causal impact on learning [4, 5]. Studies replicating the doer effect in different learning environments confirmed generalizability of this learning by doing approach [6, 7]; however, bringing this method to more students was a persistent challenge. Artificial intelligence presented a solution to this challenge as tools became robust enough to develop an automatic question generation (AQG) pipeline capable of generating millions of practice questions in very little time. The primary objective of the AQG system was to generate formative practice and feedback to be placed alongside textbook content in an ereader platform for use by students in higher education contexts. After the release of the automatically generated (AG) questions, years of research looking at millions of student-question interactions contributed to setting performance metric benchmarks for AG questions and revealed new insights into student behaviors and learning [8-12].

The aim of this paper is twofold. First, we synthesize findings from our multi-year program of research on AI-generated formative practice questions, highlighting both the technical performance benchmarks and their impact in classroom contexts. Second, we reflect on the persistent challenges of effectively scaling student engagement with formative practice, setting out a forward-looking vision for integrating these tools into everyday learning. By combining a review of empirical results with an analysis of practical barriers, we seek to show not only that AI-generated practice can achieve comparable quality to human-authored questions, but also how these systems can maximize learning potential when thoughtfully embedded into teaching and learning environments.

In line with this dual focus, the paper is organized to address both performance at scale and applications in authentic classrooms. Performance metrics drawn from millions of student-question interactions establish validity and reliability of AG questions, while classroom-based studies demonstrate how instructor course policies and student use patterns influence outcomes. Together,

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 337–344
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

these complementary perspectives underscore how AQG contributes to learning when implemented in real-world educational settings and highlight the remaining obstacles to broader adoption.

2 AQG Methods

The AQG system is designed to support students with formative practice while engaging with textbook material, and so to ensure the questions are closely aligned to the source content, the AQG system uses the textbook as the corpus for natural language processing. Kurdi et al. [13] recommended describing the system according to level of understanding and procedure of transformation. In this system, the level of understanding includes both syntactic and semantic information, and the procedure of transformation is primarily rule-based.

Natural language processing tasks are executed using the spacy library [14], employing its CPU-optimized large language model (en_core_web_lg). Question generation relies on both syntactic and semantic understanding of the text. For cloze question types, this dual-level analysis enables two central operations: identifying the sentences from which questions will be generated and selecting the term(s) to be removed as answers. Syntactic information, including part-of-speech (POS) tagging and dependency structure, informs both content sentence selection and answer word identification. Additionally, semantic information contributes to recognizing conceptually important material. The transformation process that converts sentences into questions follows a rule-based approach developed by experts.

To identify high-value sentences, the textbook is segmented into logical sections of roughly 1,500 words, following the major organizational structure of the textbook such as chapters and their primary headings; sections exceeding this length are further subdivided. Within each section, sentence importance is assessed using the TextRank algorithm [15]. TextRank evaluates similarity between sentences by computing their vector embeddings, the effectiveness of which depends on the embedding technique employed. Our implementation uses a word2vec-based model [16] within spacy, which forms sentence embeddings by averaging the token vectors in each sentence. Prior to embedding, the AQG system filters out stop words and non-alphabetic tokens (e.g., punctuation, numerals). Sentences that are overly short (under 5 words) or long (over 40 words) are also excluded, as they are generally less appropriate for question formation. TextRank is applied to the remaining sentences in each section.

A second core aspect of cloze question creation involves selecting the appropriate answer word(s) from the previously identified sentences. Our system accounts for multiple factors in this process, such as word frequency within the corpus and whether a term appears in the textbook's glossary. However, the most critical factor is part of speech: only nouns and adjectives are considered viable candidates for answer blanks. Research from authentic learning contexts supports this focus—questions that target these parts of speech tend to receive better evaluations from learners than those using verbs or other word types [17]. As such, POS tagging is a fundamental component of AQG, as it is in many NLP applications.

Multiple choice or glossary term compare-and-contrast questions rely on the existence of a textbook glossary, but are created using similar methods.

This AQG approach is designed for broad applicability across academic disciplines but is not suitable for all subject areas; notably, it is not effective for mathematics or language instruction.

Feedback is provided using textbook sentences that are related to the one from which the question stem was created—either a different sentence containing the same answer term (example in Figure 1) or neighboring sentences that provide added context. Outcome-based feedback (correct/incorrect) is always presented.

While the system does not attempt to calibrate question difficulty during generation, student response data collected after deployment is used to monitor difficulty levels. Questions identified as excessively difficult for formative purposes are automatically replaced [17]. Paraphrasing or rewording of textbook content is intentionally avoided to ensure terminology consistency between questions and the source material. The resulting questions with integrated feedback are delivered in clusters that open alongside the relevant textbook section and allow students to get immediate feedback, retry or reveal answers, and rate questions (Figure 1).

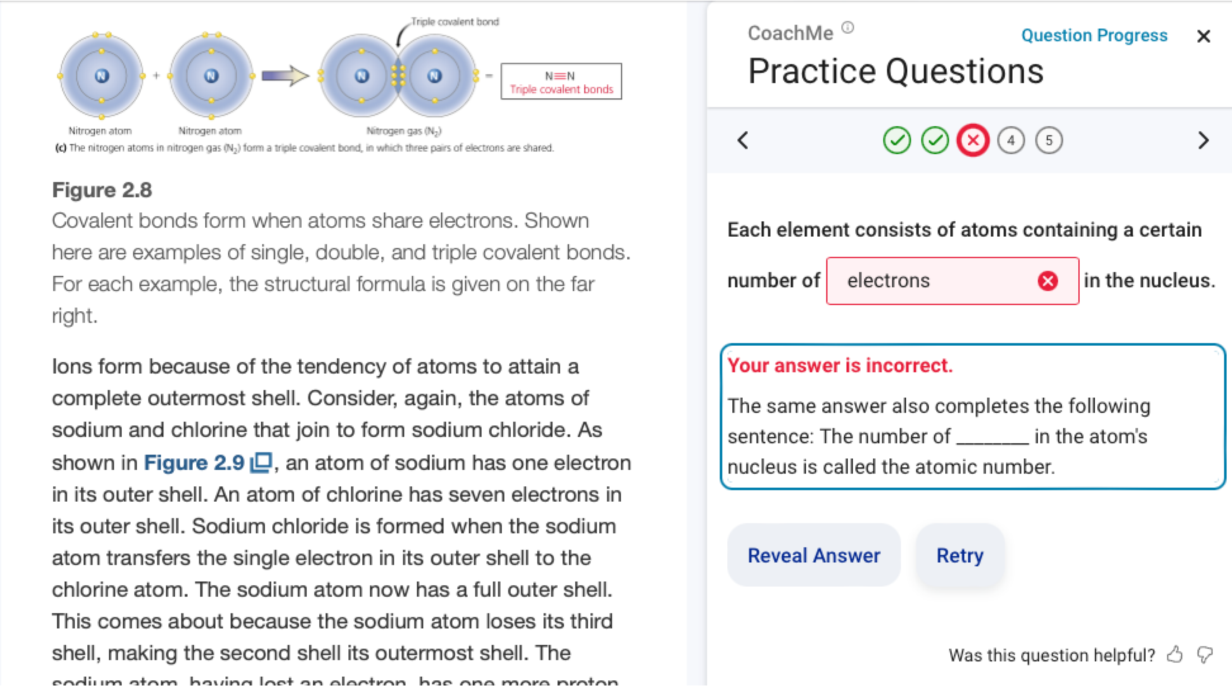


Figure 1. A fill-in-the-blank question open next to the textbook content

The original AQG system was developed without the use of large language models (LLMs) for two key reasons. First, LLMs lacked sufficient reliability at the time of the pipeline's development. Second, their potential to introduce factual inaccuracies posed an ethical concern, especially given the vast number of questions being generated—making human oversight unfeasible at scale. However, LLMs have key strengths that could potentially be harnessed for specific tasks within the existing AQG pipeline [18] or providing error-specific feedback on open-ended questions [19]. While crafting open-ended questions is relatively straightforward, offering targeted feedback is significantly more complex. Intelligent tutoring systems are known for delivering highly effective, individualized feedback that addresses student errors, making them among the most impactful forms of computer-based learning [20, 21]. Historically, scaling this type of feedback has been a major limitation. However, the proficiency of LLMs in text comparison may offer a viable path forward in addressing this challenge.

In the autumn of 2024, two new types of open-ended questions were introduced alongside the existing AG formative question types: a glossary term compare-and-contrast prompt and a "write your own exam question" task. These additions were chosen specifically to engage learners in advanced cognitive process dimensions [22]. To support these questions, an LLM is employed to analyze student responses by comparing them to the corresponding textbook sections and generating constructive, personalized feedback. Although the rule-based AQG pipeline had the capacity to formulate such open-ended prompts previously, deploying them without the ability to provide feedback risked leaving students unsure about the accuracy of their answers—potentially reinforcing misconceptions. As a result, implementing these question types necessitated the inclusion of a mechanism for tailored feedback.

3 Performance Metric Benchmarks

A benefit of digital learning environments is their ability to collect enormous quantities of high-quality data [23]. These microlevel clickstream data allow us to investigate old questions with novel data and gain a finer-grained understanding of student learning processes [24, 25]. The microlevel data collected by the ereader platform are valuable for investigating both the performance of AG questions and student behaviors. The platform records each student interaction with a timestamp and unique numeric identifier for the student. Student-question sessions are formed by grouping all interactions of a single student on a single question. No personally identifiable information is collected by the platform. These data are then used to evaluate several different performance metrics, including:

- Difficulty index: Percentage of sessions in which the student's first answer attempt was correct (lower values correspond to more difficult questions).
- Persistence rate: Among sessions in which the first attempt was incorrect, the percentage in which the student continued until submitting a correct answer.
- Thumbs up rate: Number of thumbs up ratings per 1,000 student-question sessions (one rating opportunity per session).
- Thumbs down rate: Number of thumbs down ratings per 1,000 student-question sessions.

The initial release of AG questions for student use was in a courseware environment where AG questions were intermixed with human-authored questions and placed intermittently with short content lessons. This first research found no difference in how students use AI-generated versus human-authored questions. Comparing automatically generated questions to human-authored questions in the same course using a mixed-effects logistic regression model found they were similar on engagement, difficulty, persistence, and discrimination [8, 9]. The most notable difference was in the cognitive process dimension of the questions: recall types and recognition types grouped together on performance metrics—regardless of method of creation.

With satisfactory performance in a courseware environment, the AG questions were then delivered as a free study feature (CoachMe) in the Bookshelf ereader, deploying millions of questions across thousands of textbooks. Analysis of over 7 million student-question interactions confirms these performance metric benchmarks at scale—recognition-type questions are generally easier than recall-type questions and have higher persistence. Investigating student answers revealed insight into behaviors: only about 12% of students-question interactions had a "non-genuine" input to the fill-in-the-blank, and nearly half of those students persist in answering until they get the correct response, indicating non-genuine responses were part of a strategy for many students [26]. Tracing interaction patterns also revealed the type of question impacted how students engaged with them [27]. The scale of this release made human monitoring of question performance impossible, so a content improvement service (CIS) was developed. The CIS is a platform-level adaptive system that monitors every student-question interaction in real time and deploys tools such as Bayesian evaluation of difficulty metrics and student ratings (thumbs down specifically) to determine if questions need to be removed and replaced [28]. Across a total of 3,594,408 question sessions, the overall thumbs down rate observed was 1.94 [29].

To provide an updated set of aggregated performance metrics, all student-question interaction events were retrieved starting from the feature's launch on January 1, 2022, to June 11, 2025. The resulting dataset consisted of 16,645,791 sessions across 2,485,201 unique questions, 822,678 students, and 14,371 textbooks, with a total of 26,169,711 interaction events. Table 1 summarizes these performance metrics by question type.

Compared to the performance metrics from [26] in 2023, the overall trends by the cognitive process

	Answered	Mean Difficulty	Persistence	Thumbs Up Rate	Thumbs Down Rate
Matching	4,028,835	80.3	72.8	3.56	1.00
Self-Graded Submit & Compare	526,080	86.8	NA	4.64	2.37
FITB	11,912,905	61.4	62.1	3.28	1.73
Multiple Choice	205,774	74.1	76.1	3.68	2.10

Table 1. Performance metrics by question type.

dimension of the question types remain the same. The recognition-type matching and multiple choice questions are easier and have higher persistence than the recall FITB type. However, we see some interesting changes. In 2023, the FITB had a difficulty of 54.7 and persistence of 58.5. The most recent data show an increase for both metrics to a difficulty of 61.4 and persistence of 62.1. This increase is overall positive, and potentially was impacted by improvements made to the AQG pipeline and question placement in December of 2023. The only other large difference is persistence for multiple choice, which fell from 93.6 to 76.1—potentially related to the nearly tenfold increase in data collected on this question type since 2023.

In addition to monitoring performance benchmarks of the AG questions themselves, we investigated AG feedback. The type of feedback used for formative practice matters. Scaffolding feedback that provides another context (Figure 2) was most effective for increasing student persistence in answering until correct as well as decreasing the time it took to get to the correct answer [11]. Additionally, the advances in large language models (LLMs) made it possible to scale personalized, error-specific feedback for open-ended question types—a hallmark feature of intelligent tutoring systems [19]. Introducing LLM-based error-specific feedback for open-ended questions produced by this AQG pipeline provided experience with an LLM-based feature that could replicate the hallmark personalized feedback of intelligent tutoring systems but required careful development to minimize potential LLM failures [17].

4 Data from the Classroom

The millions of questions available for analysis provide valuable performance benchmark metrics for AG questions. However, the large aggregated dataset includes all learners in all learning contexts—even those who only answered a few questions. Therefore, it was also valuable to engage in classroom-based research to determine how instructor course policies impacted student engagement with the practice and how the AG formative practice might impact learning. Studying 19 course sections where faculty assigned these AG questions as a participation homework assignment showcased how classroom contexts and course policies increased student engagement and impacted performance metrics [29]. Nearly all students answered 100% of the questions, even when only 80% was required to receive credit. Across all courses, the matching questions had a mean difficulty of 82.8% and a persistence of 96.7%. The FITB questions had a mean difficulty of 82.7% and a persistence of 94.0%. The higher difficulty index and persistence for questions in the classroom setting indicates students put more effort into their first attempt at the question and were motivated to continue answering until they input

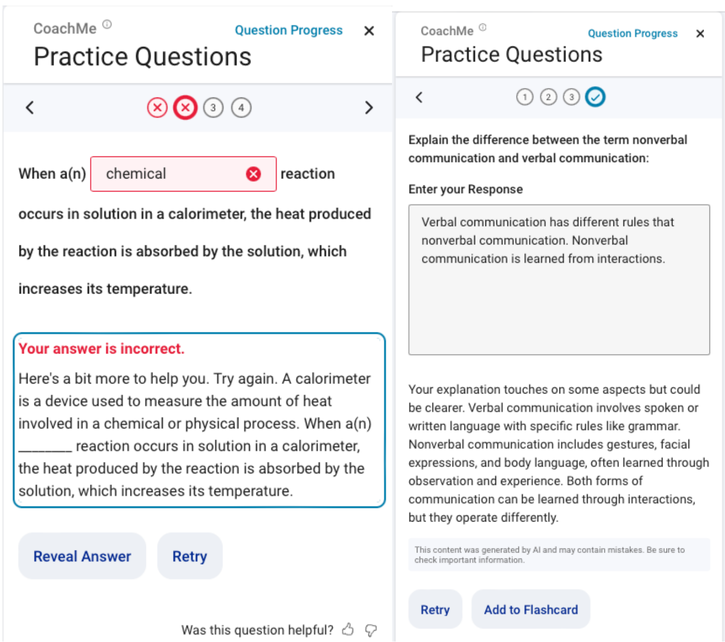


Figure 2. Scaffolding feedback for FITB questions (left) and LLM-based error-specific feedback for open-ended questions (right).

the correct response. The non-genuine responses for FITB ranged widely between courses, but remarkably, 12 of 19 courses had persistence over 99% for non-genuine responses. Faculty observed increased preparedness for classroom discussions and higher quality written assignments and projects and students anonymously reported finding the practice helpful for both learning and accountability on course evaluations [12].

In two semesters of a large cognitive psychology course, a change in faculty policy similarly shifted students from doing practice at the end of the course when it would not be helpful for the exams to prior to the related exam [30]. This change led to a statistically significant increase in exam scores (particularly meaningful for struggling students at the 25th and 50th percentile). Additionally, a post hoc analysis replicating Koedinger et al.'s doer effect analysis found results consistent with the literature. This first analysis of AI-generated questions eliciting the same doer effect principle in the classroom confirms the utility of AI for question generation at scale [30].

5 Recommendations, Challenges, Future Work

A key contribution of this review is to identify not only what the AQG pipeline has achieved in terms of question quality and learning outcomes, but also the persistent barriers that hinder scaling student engagement with formative practice. Each individual research study conducted on this AQG system since its initial release in 2019 investigates specific components in detail, such as performance metrics, student perceptions, feedback, student engagement patterns, textbook reading, learning outcomes, etc. Together, this rigorous evaluation of nearly every aspect of question performance and student behaviors and learning is essential to a comprehensive overview of the efficacy of AI-generated questions for formative practice at scale.

While our analyses confirm that AG questions perform well across multiple metrics and can replicate the doer effect in classroom settings, two persistent barriers emerge. First, faculty awareness and adoption remain uneven—many instructors are not fully informed about the availability of AG questions embedded in their textbooks. Second, student engagement is highly dependent on course structures; voluntary use of AG practice is typically low unless supported by meaningful course incentives or policies. These barriers illustrate that successful application of AQG in classrooms is not a purely technical challenge but an educational and organizational one. Addressing these barriers is essential to realizing the potential of formative practice: maximizing learning through classroom application. Without meaningful faculty engagement, voluntary student use of the questions will remain low. Instructors remain the most meaningful agents of change in the classroom and helping to inform and educate instructors as key partners in implementation will remain the focus of future efforts.

Future work will always include iterative improvement to the AQG pipeline. The analysis of the questions showcases their validity, yet continued refinement can further improve question quality. We have evidence of the importance of this improvement cycle, as changes made to sentence selection and placement within the text in the winter of 2023 resulted in a reduction of thumbs down ratings from 1.95 to 1.39 per thousand. While the thumbs down rate is very low, decreasing it by more than 25% indicates an effective improvement that could influence student perceptions of the questions. While LLMs were not used in the existing AQG pipeline, we have conducted promising research on how introducing LLMs at key steps in the pipeline could further increase question quality [18].

Taken together, the results of this research establish clear performance benchmarks for AI-generated formative practice questions, demonstrating that they perform comparably to human-authored questions across difficulty, persistence, and engagement metrics at scale. Classroom-based implementations further confirm that when these questions are embedded into instruction, they not only support higher persistence and accuracy but also contribute to measurable gains in exam performance and student preparedness. These findings underscore that AI-generated formative practice is both valid and impactful when used in authentic educational settings.

Looking ahead, the continued refinement of AQG pipelines, coupled with thoughtful integration of LLM-based personalized feedback and stronger faculty engagement strategies, points toward a future in which textbooks function as interactive, learning-by-doing environments that reliably maximize student learning potential.

References

[1] Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81–90. https://doi.org/10.1177/003172171009200119

[2] Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., & Willingham, D. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

[3] Koedinger, K. R., Kim, J., Jia, J., McLaughlin, E., & Bier, N. (2015, March). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. *Proceedings of the Second ACM Conference on Learning@Scale (L@S'15),* pp. 111–120. https://doi.org/10.1145/2724660.2724681

[4] Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016, April). Is the doer effect a causal relationship? How can we tell and why it's important. *Proceedings of the Sixth International Learning Analytics & Knowledge Conference (LAK '16),* pp. 388–397. http://dx.doi.org/10.1145/2883851.2883957

[5] Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the doer effect robust across multiple data sets? In *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 369–375).

[6] Van Campenhout, R., Johnson, B. G., & Olsen, J. A. (2021, July 18–22). The doer effect: Replicating findings that doing causes learning. In *The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning* (*eLmL 2021)* (pp. 1–6). IARIA. https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001

[7] Van Campenhout, R., Jerome, B., & Johnson, B. G. (2023). The doer effect at scale: Investigating correlation and causation across seven courses. *Proceedings of the 13th International Learning Analytics & Knowledge Conference (LAK '23),* pp. 357–365. https://doi.org/10.1145/3576050.3576103

[8] Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation. In *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education CEUR Workshop Proceedings*, (pp. 1–12). https://ceur-ws.org/Vol-2895/paper06.pdf

[9] Johnson, B. G., Dittel, J. S., Van Campenhout, R., & Jerome, B. (2022). Discrimination of automatically generated questions used as formative practice. *Proceedings of the Ninth ACM Conference on Learning@Scale (L@S'22),* pp. 325–329. https://doi.org/10.1145/3491140.3528323

[10] Van Campenhout, R., & Hubertz, M. (2023). Context and considerations for investigating the impact of learning by doing on student equity. *Workshop on Equity, Diversity, and Inclusion in Educational Technology, Artificial Intelligence in Education (AIED)* (pp. 1–5). https://doi.org/10.5281/zenodo.8208452

[11] Van Campenhout, R., Kimball, M., Clark, M., Dittel, J. S., Jerome, B., & Johnson, B. G. (2024). An investigation of automatically generated feedback on student behavior and learning. *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK'24)*, pp. 850–856. https://doi.org/10.1145/3636555.3636901

[12] Van Campenhout, R., Johnson, B. G., Clark, M., Deininger, M., Harper, S., Odenweller, K., & Wilgenbusch, E. (2025). AI-generated questions in context: A contextualized investigation using platform data, student feedback, and faculty observations. *Journal of Communications Software and Systems, 21*(2), 178–188. https://doi.org/10.24138/jcomss-2024-0120

[13] Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

[14] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

[15] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* pp. 404–411. https://aclanthology.org/W04-3252

[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR) 2013. Workshop proceedings.* https://doi.org/10.48550/arXiv.1301.3781

[17] Jerome, B., Van Campenhout, R., Dittel, J. S., Benton, R., & Johnson, B. G. (2023). Iterative improvement of automatically generated practice with the Content Improvement Service. In R. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems*. *HCII 2023* (Lecture Notes in

Computer Science, pp. 312–324). Springer. https://doi.org/10.1007/978-3-031-34735-1_22

[18] Dittel, J. S., Van Campenhout, R., & Johnson, B. G. (2025). Refining sentence selection for automatic cloze question generation with large language models. *The Twelfth ACM Conference on Learning @ Scale (L@S'25)*. https://doi.org/10.1145/3698205.3733926

[19] Van Campenhout, R., Dittel, J. S., & Johnson, B. G. (2026). Scaling effective characteristics of ITSs: A preliminary analysis of LLM-based personalized feedback. In S. Graf & A. Markos (Eds.), *Generative systems and intelligent tutoring systems. ITS 2025. Lecture Notes in Computer Science* (Vol. 15723). Springer. https://doi.org/10.1007/978-3-031-98281-1_13

[20] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, *46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

[21] Kulik, J. A., & Fletcher, J. D. (2016).Effectiveness of intelligent tutoring systems:A meta-analytic review. *Review of Educational Research, 86*(1), 42-78.

[22] Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). Longman.

[23] Goldstein, P. J., & Katz, R. N. (2005). Academic analytics: The uses of management information and technology in higher education. Educause. https://library.educause.edu/-/media/files/library/2005/12/ers0508w-pdf.pdf

[24] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

[25] McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open, 7*(1), 1–12. https://doi.org/10.1177/23328584211052055

[26] Van Campenhout, R., Clark, M., Jerome, B., Dittel, J. S., & Johnson, B. G. (2023). Advancing intelligent textbooks with automatically generated practice: A large-scale analysis of student data. *5th Workshop on Intelligent Textbooks*. *The 24th International Conference on Artificial Intelligence in Education* (pp. 15–28). https://intextbooks.science.uu.nl/workshop2023/files/itb23_s1p2.pdf

[27] Van Campenhout, R., Clark, M., Dittel, J. S., Brown, N., Benton, R., & Johnson, B. G. (2023). Exploring student persistence with automatically generated practice using interaction patterns. In *Proceedings of 2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)* (pp. 1–6). https://doi.org/10.23919/SoftCOM58365.2023.10271578

[28] Jerome, B., Van Campenhout, R., Dittel, J. S., Benton, R., Greenberg, S., & Johnson, B. G. (2022). The Content Improvement Service: An adaptive system for continuous improvement at scale. In Meiselwitz, et al., *Interaction in New Media, Learning and Games. HCII 2022 (*Lecture Notes in Computer Science, Vol 13517, pp. 286–296). Springer, https://doi.org/10.1007/978-3-031-22131-6_22

[29] Van Campenhout, R., Clark, M., Johnson, B. G., Deininger, M., Harper, S., Odenweller, K., & Wilgenbusch, E. (2024). Automatically generated practice in the classroom: Exploring performance and impact across courses. In *Proceedings of the 32nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2024)* (pp. 1–6). https://doi.org/10.23919/SoftCOM62040.2024.10721828

[30] Van Campenhout, R., Autry, K., Clark, M. W., & Johnson, B. G. (2025). Scaling the doer effect: A replication analysis using AI-generated questions. *Proceedings of the Twelfth ACM Conference on Learning@Scale (L@S'25)*. https://doi.org/10.1145/3698205.3729545

Beyond Agreement: Rethinking Ground Truth in Educational AI Annotation

Danielle R. Thomas
Carnegie Mellon University
Pittsburgh, PA, USA
drthomas@cmu.edu

Conrad Borchers
Carnegie Mellon University
Pittsburgh, PA, USA
cborchers@cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
Pittsburgh, PA, USA
koedinger@cmu.edu

Abstract

Humans can be notoriously imperfect evaluators. They are often biased, unreliable, and unfit to define "ground truth." Yet, given the surging need to produce large amounts of training data in educational applications using AI, traditional inter-rater reliability (IRR) metrics like Cohen's kappa remain central to validating labeled data. IRR remains a cornerstone of many machine learning pipelines for educational data. Take, for example, the classification of tutors' moves in dialogues or labeling open responses in machine-graded assessments. This position paper argues that overreliance on human IRR as a gatekeeper for annotation quality hampers progress in classifying data in ways that are valid and predictive in relation to improving learning. To address this issue, we highlight five examples of complementary evaluation methods, such as multi-label annotation schemes, expert-based approaches, and close-the-loop validity. We argue that these approaches are in a better position to produce training data and subsequent models that produce improved student learning and more actionable insights than IRR approaches alone. We also emphasize the importance of external validity, for example, by establishing a procedure of validating tutor moves and demonstrating that it works across many categories of tutor actions (e.g., providing hints). We call on the field to rethink annotation quality and ground truth—prioritizing validity and educational impact over consensus alone.

1 Introduction

In the field of educational measurement, researchers increasingly rely on automated assessment methods to enable scalable, real-time evaluation of learning (Messer et al., 2024a; Thomas et al., 2025b). These methods typically include artificial intelligence (AI)-based models, such as the BLEURT (Sellam et al., 2020) metric for machine translation, and recent rubric-based prompting strategies with large language models (LLMs) (Hashemi et al., 2024). Yet, even in such scenarios, the ultimate source of truth remains tied to human annotations, either as a source of supervision during model training or as a post-hoc verification of model outputs. As the quality of any evaluation protocol hinges on the robustness of its underlying "ground truth," human annotation schemes must not be taken at face value, but instead be rigorously assessed for pertinence and alignment with educational evaluation objectives.

This work focuses on the limitations of inter-rater reliability (IRR) and recommends alternatives. IRR is a statistical measure used to determine the degree of agreement among multiple human annotators (Gwet, 2021) and is often used as a basis for validation in educational evaluation and modeling. Although IRR provides a convenient quantitative proxy for annotation quality, it is fundamentally limited by the subjective, biased, and often inconsistent nature of human judgment (Gwet, 2021).

The concerns surrounding IRR in the field of educational AI are not new. Researchers have long challenged the sufficiency of traditional IRR metrics such as Cohen's kappa and Krippendorff's α in capturing the complexity of human evaluative tasks (Ando and Zhang, 2005; Gwet, 2021; Doewes et al., 2023; Plank, 2022). The data labeling industry echoes this concern, "There is a general belief in the data labeling industry that high inter-rater reliability indicates high-quality data... However, this is not always the case" (Toloka AI, 2023). Yet, educational applications are also different from this prior research: As we go on to argue, the impact of assessment on learning and its generalizability across learning contexts poses distinct opportunities to move beyond IRR-only evaluations. The present study's primary contribution is to provide a comprehensive understanding of the problem of IRR-based validation and the paths to move beyond it in educational applications.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 345–351
October 27-29, 2025

1.1 Understanding the Problem

In education and beyond, the reliability of human-annotated training data as the ultimate "gold standard" for evaluating learners' responses has come under increasing scrutiny (Chen et al., 2024; Messer et al., 2024b). While these issues—such as inconsistency, subjective interpretation, and bias—are well documented, we argue they are increasingly overlooked in the rush to use capable AI models to scale assessments quickly. IRR offers a tempting sense of rigor, often mistaken for objectivity, especially in inherently subjective tasks like grading essays or labeling open responses. However, high IRR can mask annotation shallowness, promote premature consensus, and ignore valid alternative interpretations. Overreliance on IRR is common and problematic, reinforced by decades of practice but increasingly out of step with the complexity of contemporary assessment tasks and AI capabilities (Ando and Zhang, 2005; Doewes et al., 2023; Gwet, 2021). The need for more robust, nuanced, and task-sensitive evaluation strategies is now more urgent than ever, especially as new educational AI models are rapidly being deployed.

Next to the intrinsic limitations of AI systems—such as hallucinations in LLMs or the low interpretability of many ML architectures (Huang et al., 2025)—there are deep-seated problems stemming from the training data itself. This includes both the human-annotated datasets and the open-web corpora typically used to train foundation models (e.g., LLMs are known to include biases they soaked up from these training data (Chen et al., 2024). Our work focuses specifically on the former: when training data for assessment models is generated by humans, IRR is often used to validate annotation rubrics and establish the "gold standard" against which models are evaluated. However, IRR does not always reflect annotation correctness or task difficulty. High agreement can obscure flawed or superficial annotations, while disagreement may indicate productive ambiguity or meaningful variation in interpretation. This paper highlights alternative approaches to supplement and strengthen traditional methods, highlighting a multidimensional approach that can support the demands of scalable, equitable, and informative educational assessment.

1.2 Aims

The stakes are high. Presently, 86% of students and 60% of teachers report using AI tools, as the education market is expected to exceed $88 billion by 2032 (Digital Education Council, 2025). AI has become ubiquitous within classrooms, interventions, and high-stakes testing, with the need to responsibly define and evaluate "ground truth" increasingly urgent. If we continue to equate consensus with correctness, we risk optimizing models not for pedagogical utility or learning outcomes, but for compliance with flawed human standards. We argue that the field of educational AI must move beyond narrow agreement metrics and embrace more flexible, validity-driven approaches to annotation that ensure effectiveness and impact of AI tools.

This work contributes examples of alternative approaches proposed by researchers in the field and their results, such as multi-label annotations that reflect interpretative diversity and close-the-loop validity measures that tie labels to learning outcomes. Together, these approaches provide a richer, more responsible framework for defining and validating ground truth in educational AI. However, these alternative approaches alone are not the solution but provide some examples of how other researchers have attempted to varying degrees of success to overcome the challenges of sole reliance on IRR alone. In showcasing other approaches, we emphasize the lack of external validity in educational AI, such as tutoring systems. We hope to increase awareness as educational AI systems are encroaching on learning as we know it.

The aims of this work are three-fold:

- **Challenge the field's overreliance on inter-rater reliability (IRR)** as the primary validator of annotation quality in educational AI, arguing that consensus alone is insufficient for modeling complex, subjective data
- **Introduce and illustrate alternative or supplemental frameworks** that support a more multidimensional, validity-centered approach to defining "ground truth" in assessment
- **Call attention to the lack of external validity in educational AI** and propose a challenge of demonstrating examples in the field, such as a generalizable tutoring model across a diverse range of datasets.

2 Case Applications in Educational AI

2.1 Comparative Judgment

Reported Use Case: Using comparative judgment to assess students' reading comprehension and flu-

ency in open responses. Henkel and Hills (2023) present a compelling alternative to traditional IRR approaches by implementing comparative judgment as a method for labeling educational data in the form of middle school students' open response to math problems. The researchers identify the limitations of relying on expert raters and rigid categorical rubrics for scoring student responses, especially for complex or open-ended tasks, and propose comparative judgment as a more scalable, accessible alternative. Comparative judgment requires raters to determine which of two student responses is better, rather than assigning an absolute score. This approach is cognitively easier for raters, particularly non-experts, and aligns with reinforcement learning from human feedback methods used in AI (Christiano et al., 2017). In two experiments involving short-answer reading comprehension and oral reading fluency, the study compares traditional categorical judgment with comparative judgment. Results show that comparative judgment substantially improved both accuracy and inter-rater reliability. For short-answer tasks, Krippendorff's α improved from 0.66 to 0.80, and accuracy increased by 13%. For oral fluency, α improved from 0.70 to 0.78. These gains were statistically significant.

Henkel and Hills (2023) argue that comparative judgment not only improves labeling quality but also challenges the primacy of IRR as the sole measure of annotation quality. They demonstrate that comparative approaches can match or exceed expert-level consistency, even when crowdworkers are used. This study makes a significant contribution by showing that comparative judgment can be an effective alternative or supplement to IRR in educational data annotation, with practical implications for scaling data labeling efforts in educational research and AI development.

2.2 Multi-label Annotation

Reported Use Case: Identifying toxic or offensive text in chat messages. Arhin et al. (2021) proposed a multi-label annotation strategy to address the challenges of subjectivity and inconsistency in toxic text classification. Although this use case is not directly applied to an education context, per se, identifying possibly harmful language is important in all aspects of educational AI. Recognizing that language is deeply contextual and that annotator judgments often vary due to differing backgrounds, values, and interpretations, the authors rejected the traditional assumption of a singular, definitive ground truth label. To operationalize their approach, they re-annotated three toxic text datasets using three context-based label types: **strict label** (based on the presence of offensive words, regardless of context); **relaxed label** (a more lenient judgment allowing for interpretive variability); **inferred group label** (based on how a statement might be perceived if uttered by a member of the referenced group). This multi-labeling scheme captured nuanced perspectives and better represented the diversity of valid interpretations. While the approach did not always improve inter-annotator agreement, it produced annotations with higher alignment to external machine-learning classifiers (e.g., Detoxify and Perspective API), thus offering enhanced dataset quality and potential model generalizability. Researchers emphasized that multi-labeling reflects the inherent ambiguity in human language more faithfully than forced consensus and serves as a more robust foundation for training and evaluating AI systems (Arhin et al., 2021).

2.3 Expert-Based Labeling Approaches

Reported Use Case: Evaluating annotator quality through expert-grounded benchmarks in dialogues. While traditional IRR assumes agreement among annotators as the gold standard, recent work challenges this assumption by leveraging expert-labeled data as a more principled benchmark for evaluating annotation quality. We provide two examples of leveraging expert-based approaches.

First, Wang et al. (2024b) examined annotator characteristics, where researchers compared individual annotator judgments against expert-provided labels rather than against other annotators, arguing that consensus does not always equate to correctness. Their findings showed that high inter-annotator agreement can sometimes mask low-quality or superficial judgments, particularly when annotators share common biases or lack subject matter expertise. Researchers further introduced a predictive modeling approach to identify reliable annotators in advance, using background traits (e.g., education, domain familiarity) and behavioral features (e.g., response time). By combining expert-aligned accuracy with annotator profiling, this work offers a novel lens for establishing annotation reliability independent of peer agreement.

Second, Nahum et al. (2024) established a high-confidence ground truth and rigorously addressed IRR challenges. They implemented an expert-based re-annotation protocol in conjunction with

LLM-driven error detection. They used 11 datasets on a number of binary tasks, such as identifying (or not) hallucinations in LLM-generated text and fact verification. Specifically, researchers focused expert efforts on examples where LLM ensemble predictions disagreed with the original labels, regardless of model confidence. Each of these examples was independently reviewed by two expert annotators—who were familiar with task definitions and annotation guidelines. During the initial annotation phase, both experts rated examples on a binary scale, classifying them as either factually consistent or inconsistent. To ensure impartiality, examples were presented in randomized order without exposing the original or LLM-provided labels. Following this phase, a reconciliation process was undertaken for all examples where the annotators disagreed. The experts discussed each disagreement to reach consensus, allowing for a refined label set used throughout the remainder of our analysis. This reconciliation improved the quality of annotations and significantly enhanced IRR, with Fleiss's κ for expert annotations increasing from 0.486 to 0.851 after reconciliation.

For comparison, researchers computed IRR statistics across all annotator types, including crowd-sourced workers, individual LLMs (with prompt variations), and an ensemble of LLM models. The results, summarized in Table 1, reveal that while GPT-4 and PaLM2 achieved high κ values (0.706 and 0.750, respectively), closely matching human expert reliability, crowd-sourced annotations from MTurk exhibited near-random agreement (κ = 0.074). Furthermore, the ensemble approach of combining multiple LLM models and prompts yielded a moderate κ of 0.521, suggesting that assembling not only improves the label quality, but also stabilizes the level of agreement across annotations. These findings highlight the value of expert reconciliation in overcoming IRR limitations and underscore the relative strengths of LLMs as scalable annotation tools when expert supervision is applied selectively.

2.4 Predictive Validity

Reported Use Case: Mapping open responses to MCQs on the same learning objectives. Thomas et al. (2025a) used predictive validity as an alternative method to IRR in validating the effectiveness of their large language models (LLM) on assessing tutor learners open responses while engaging in scenario-based training. In this context, to establish predictive validity, the researchers use multiple-choice questions (MCQs) that assess the same learning objectives as open-response questions (Thomas et al., 2025a). Figure 1 displays an open response and corresponding multiple-choice question assessing the same learning objective of how to effectively respond to a student who has just made a math error.

Predictive validity refers to the extent to which an assessment accurately forecasts or correlates with future performance on a related measure (Trochim et al., 2016). In this case, predictive validity evaluates whether performance on MCQs can reliably predict outcomes on open-response questions or other measures of learner understanding. Much work has found MCQs can be as effective, and more efficient, than open response tasks when instructional time is limited (Thomas et al., 2025b; Butler, 2018). This approach enhances objectivity by reducing subjective biases inherent in human scoring while ensuring that MCQs serve as effective proxies for more complex assessments.

To assess the predictive validity of LLM-generated scores, namely GPT-4o, Gemini 1.5 pro, and LearnLM, on open responses in relation to MCQ scores, Thomas et al. (2025a) computed the

Table 1: Reliability across annotation sources demonstrating excellent agreement among experts after reconciliation. Originally published in Nahum et al. (2024).

Source	Fleiss's κ	Interpretation
Experts (Post-Recon.)	0.851	Excellent
GPT-4	0.706	Close to experts
PaLM2	0.750	High reliability
MTurk Workers	0.074	Near random

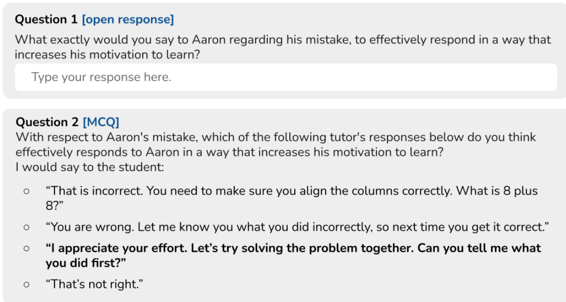


Figure 1: An open response and corresponding MCQ (correct selection in bold) that assess a tutor learner on the same learning objective (effectively responding to a student who has made a math error). By using predictive validity, human grading is not the sole source of ground truth. Adapted and modified from Thomas et al. (2025a,b).

correlation of participants' MCQ scores and their LLM scores on open responses. The analysis revealed a significant, positive correlation between MCQ scores and human-graded open-response scores, $r(86) = 0.421, p < .001$. While this correlation was statistically significant, it was not particularly large. Potential contributors to the moderate correlation size can be attributed to the few test items and inherent ambiguity in judging the correctness of open responses, even for human graders. Given they were striving to find alternative methods to human grading, this is not the best direct comparison. Thus, they computed the correlation between MCQ and LLM scores, correlating MCQ scores with GPT-4o-scored open responses, yielding $r(86) = 0.406, p < .001$; and MCQ scores with LearnLM-scored open responses, yielding $r(86) = 0.477, p < .001$. They found that LLM scores have significant predictive validity and this validity can be determined without open response grading by humans.

Notably, the LearnLM correlation of 0.477 is 0.056 higher than the human-scored correlation of 0.421. In other words, with this tighter comparison, we find that the predictive validity of the LLM scoring is comparable to that of human scoring. This result supports predictive validity as a complementary method for evaluating model performance, though in this particular case, it should be combined with additional measures to ensure a more comprehensive assessment. Other alternative approaches to using human scores as "ground truth" mentioned that align with LLMs-as-a judge include: using the average LLM score among several models, including adversarial models, e.g., LearnLM vs GPT-4o, to establish reliability, rather than comparing human judgments; and applying LLM self-consistency measures by comparing evaluations across varying prompts to check robustness.

2.5 Close-the-loop Validity

Reported Use Case: Mapping tutor move classifications to student performance and learning outcomes. From a measurement point of view, where we have critiqued the use of human-derived annotation schemes and codes as limited, close-the-loop validity can also be used to qualify the predictive capabilities of a measurement or label. Specifically, close-the-loop validity ensures that a given assessment or model produces improved learning in line with its theoretical underpinning or coding scheme. A recent example of this is Wang et al. (2024a) who demonstrated that tutors with access to Tutor CoPilot—an AI-powered system trained to reflect specific expert pedagogical strategies and reasoning—were more likely to use strategies aligned with high-quality teaching (e.g., asking guiding questions), and that these differences in tutor behavior translated into significant gains in student mastery, particularly for students taught by lower-rated tutors. This illustrates how the measurement of pedagogical quality and its operationalization in tutor strategy recommendations, supported by AI-based classifiers, can exhibit close-the-loop (and internal) validity by linking to the learning outcomes of the students within the system. Crucially, by linking tutor practices to student learning gains internally, in addition to validating classifiers, the researchers closed the loop on their classification scheme and demonstrated that their evidence-based taxonomy of tutor strategy recommendations correlates meaningfully with better learning. However, the study was also not without limitations: in particular, they did not correlate the specific occurrence or frequency of strategies (e.g., identified in tutor transcripts) to differences in learning gains. As we go on, such correlations of individual labels (and their dosage) with learning may pose an even stronger form of close-the-loop validity.

3 Towards a Multidimensional Ground Truth and Future Directions

Establishing internal validity among AI classifiers, such as the case of Tutor CoPilot, is uncommon, and desperately needed in the field of educational AI. But what is even more uncommon is external validity. ***We could not find a single example use case demonstrating external validity within educational AI.*** Establishing external validity is rare in educational AI. External validity is a type of validity, which broadly refers to how well a measurement or study captures what it intends to measure and supports the conclusions drawn from it (Trochim et al., 2016). Specifically, external validity is about generalizability—whether findings from one context apply to other people, settings, or times. In the case of tutoring, external validity may ask whether assessments of tutoring skills during upskilling (such as in structured training tasks or simulations) accurately reflect how tutors will perform in real-world practice. It also concerns whether observed effectiveness in one context (e.g., producing an effective response during training) can be expected in others. Without external validity, we cannot assume that skills shown in training or outcomes observed in one setting will transfer to different, authentic tutoring environments.

Part of the reason for the lack of external validity is that real-world evaluation studies are costly and resource-intensive. It requires the time of teachers, often school permissions, and other overhead that purely algorithmic evaluations of assessment models do not reach. This is perhaps also why they are so important. While we can make use of evidence-based practices to grade tutor moves and create taxonomies that we know will likely make a difference in practice (e.g., prioritizing self-explanation in tutoring, which is known to enhance learning in lab experiments (Berthold et al., 2009)), such practices are not guaranteed to lead to improved learning in authentic classroom contexts. We argue that these forms of external validity do matter and can reveal gaps between technical innovation and real-world impact in education. We believe that generalizability matters.

In 2019, Baker presented a list of six challenges within the field of learning analytics (Baker, 2019). The description of these challenges included the evidence needed to demonstrate that the challenge was solved. In a similar fashion, we propose establishing generalizability of different AI tutoring classifiers across datasets as a challenge.

The details of the challenge are described as follows:

1) Build AI classifiers to identify or detect tutor moves;

2) Apply these classifiers across tutoring datasets of diverse tutor-student populations and varying tutoring modalities and implementations;

3) Provide evidence by demonstrating that the classifiers work across datasets, e.g., with degradation of quality under 0.1 (AUC ROC, Pearson/Spearman correlation, and remaining better than chance).

We leave this challenge to researchers and developers of tutoring models within educational AI.

4 Summary and Conclusion

Automated assessment methods and AI require increasingly large amounts of human-annotated training data in education. As these AI systems increasingly shape how learning is assessed and supported, the validity of their training data becomes ever more critical. Though tempting for straightforward validation and quantifiable metrics of fidelity, relying solely on IRR to define "ground truth" risks reinforcing flawed human judgments rather than optimizing for meaningful educational outcomes.

This paper challenges the field's prominent reliance on IRR as the primary standard for validating annotations in educational AI. We argue that this reliance often overlooks the complexity, subjectivity, and pedagogical significance of human responses, especially in open-ended or dialogic tasks. By showcasing supplemental frameworks, such as multi-label annotation, expert-based reconciliation, predictive validity, and close-the-loop experimentation—we demonstrate that richer, more reliable forms of "ground truth" are possible.

Moving forward, educational AI should prioritize multidimensional and validity-centered approaches, striving for external validity, to ensure its tools are not only scalable but also meaningful, effective, and grounded in authentic learning outcomes.

Acknowledgments

The authors thank Ralph Abboud for his thoughtful insights and constructive feedback. This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings, and conclusions expressed in this material are those of the authors.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-truth, whose truth?–examining the challenges with annotating toxic text datasets. *arXiv preprint arXiv:2112.03529*.

Ryan S Baker. 2019. Challenges for the future of educational data mining: The baker learning analytics prizes. *J. of educational data mining*, 11(1):1–17.

Kirsten Berthold, Tessa HS Eysink, and Alexander Renkl. 2009. Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, 37:345–363.

Andrew C Butler. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, 7(3):323–331.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Digital Education Council. 2025. Digital education council global ai student survey. https://www.digitaleducationcouncil.com/form/global-ai-student-survey-2024.

Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023*, pages 103–113.

Kilem L Gwet. 2021. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13806–13834. Association for Computational Linguistics.

Owen Henkel and Libby Hills. 2023. Leveraging human feedback to scale educational datasets: Combining crowdworkers and comparative judgement. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 411–415.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Marcus Messer, Neil CC Brown, Michael Kölling, and Miaojing Shi. 2024a. Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1):1–43.

Marcus Messer, Neil CC Brown, Michael Kölling, and Miaojing Shi. 2024b. How consistent are humans when grading programming assignments? *arXiv preprint arXiv:2409.12967*.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2024. Are llms better than reported? detecting label errors and mitigating their effect on model performance. *arXiv preprint arXiv:2410.18889*.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Danielle R Thomas, Conrad Borchers, Shambhavi Bushan, Sanjit Kakarla, Alex Houk, Erin Gatz, and Kenneth R Koedinger. 2025a. Improving open-response assessment with learnlm. *Proceedings in the Inter. Conference on AI in Education*.

Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. 2025b. Does multiple choice have a future in the age of generative ai? a posttest-only rct. In *Proceedings of the Inter. LAK Conference*, pages 494–504.

Toloka AI. 2023. What is inter-rater reliability and why it matters. https://toloka.ai/blog/inter-rater-reliablility/.

William MK Trochim, James P Donnelly, and Kanika Arora. 2016. Research methods: The essential knowledge base.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024a. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024b. Human-llm collaborative annotation through effective verification of llm labels. In *2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Automated Search Algorithm for Optimal Generalized Linear Mixed Models (GLMMs)

Miryeong Koo[1] · Jinming Zhang[1]
[1]University of Illinois at Urbana-Champaign

Abstract

Only a limited number of predictors can be included in a generalized linear mixed model (GLMM) due to estimation algorithm divergence. This study aims to propose a machine learning based algorithm (e.g., mixed-effects random forest) that can consider all predictors without the convergence issue and automatically searches for optimal GLMMs.

1 Introduction

Educational data typically have a hierarchical structure (Bryk & Raudenbush, 1989; Woltman et al., 2012) due to their sampling scheme. For example, schools in a nation are first selected, and then students in the schools are sampled. As a result, students are nested in schools, schools are nested in nations. Consequently, students from the same school tend to be correlated among themselves (school effect or random effect). A generalized linear mixed model (GLMM) is typically used in such a data set.

GLMMs estimate both fixed and random effects, leading to accommodating school effects. However, GLMMs frequently fail to converge when they need to consider many predictors and their interactions (Bates et al., 2015). To solve the convergence issue, previous research typically considered only a small subset of predictors based on literature review or applying regularization techniques such as Lasso (Tibshirani, 1996). Nevertheless, both approaches have limitations. The former case may exclude some predictors that have a large influence on the outcome, while the latter does not account for random effects.

To address the issues, this study aims to develop an algorithm that can automatically rank all predictors of statistical importance based on the whole dataset without convergence problems and then search for optimal GLMMs according to the ranking. The algorithm applies a machine learning method, called mixed-effects random forest (MERF; Hajjem et al., 2014), to rank all predictors according to their statistical importance in a mixed-effects model. Then, the algorithm searches for a random intercept model with significant predictors sequentially based on the ranking provided by MERF. Next, it searches for significant interaction terms. Lastly, random slopes are explored and possibly added to the models.

Although the proposed algorithm does not directly account for substantive meaning of each predictor, it does provide candidate GLMMs recommended. Thus, the proposed algorithm has the potential to reduce the time and effort otherwise required by researchers to identify optimal GLMMs.

2 Theoretical framework

2.1 Generalized Linear Mixed Models (GLMMs)

The generalized linear model (GLM) (McCullagh & Nelder, 1989) assumes that all observations are independent, while the generalized linear mixed model (GLMM) (Breslow & Clayton, 1993) allows dependency among subjects in the same group. GLMM can handle data hierarchy by including random effects for group dependency (e.g., school effects. The linear mixed model (LMM), also known as hierarchical linear modeling (Raudenbush & Bryk, 2002) or multilevel modeling (Goldstein, 2011), is a special case of the GLMM, where the response variable is continuous.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 352–358
October 27-29, 2025

The LMM estimates both fixed effects and random effects and has the form as shown in Equation (1),

$$Y = X\beta + Zb + \varepsilon, \quad (1)$$

where $\varepsilon \sim N(0, \sigma^2 I)$, $b \sim N(0, G)$, and ε and b are independent from each other. Here G is a block-diagonal covariance matrix.

The first two terms of the right-hand side of Equation (1) represent fixed- and random-effects parts, respectively. X is the $n \times (K + 1)$ matrix of fixed effects from predictors (here n is the number of observations and K is the number of predictors), β is the $(K + 1)$ dimensional fixed effect parameter vector, Z is the design matrix of J groups (schools), b is random effect vector, and ε is the (level-1) residual vector. The vector b for a random intercept model is J×1 vector of random intercept (b_{0j}) for each group, where $b_{0j} \sim N(0, \tau^2)$. For a random intercept and a random slope model, vector b is a J×2 vector of random effects for each group, where $b_j = \begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim N(0, G)$, and Z is an n×2J block-diagonal matrix, for each subject, including 1 for their group's random intercept (b_{0j}) and Z for their group's random slope (b_{1j}). It can be represented as a conditional model, as illustrated in Equation (2),

$$\mu = E(Y|b) = X\beta + Zb. \quad (2)$$

The generalized linear mixed model (GLMM) extends LMM to accommodate non-continuous responses, such as binary or categorical responses and the models are denoted as Equation (3),

$$g(\mu) = X\beta + Zb, \quad (3)$$

where $g()$ is a monotonic increasing and differentiable link function. For example, the logit function widely serves as a link function for binary responses.

The correlation between individuals (students) within the same group (school) is called the intraclass correlation coefficient (ICC) which measures the similarity of within-group individuals (Raudenbush & Bryk, 2002). In a random intercept model, the ICC is calculated by between-group variance (τ^2) divided by total variance, which is the sum of between-group variance (τ^2) and within-group variance (σ^2). Since the ICC is the proportion of total variance due to group differences, higher ICC implies larger group difference.

2.2 Machine learning method: MERF

A machine learning method, RF is a tree-based ensemble method that aggregates a cluster of random decision trees. Unlike standard RF (Breiman, 2001) that considers fixed effects only, MERF (Hajjem et al., 2014) is also capable of taking random effects into account, as shown in Equation (4),

$$Y = f(X) + Zb + \varepsilon, \quad (4)$$

where $f(X)$ is a general and unspecified fixed-effects part. MERF is applied as follows: After calculating the fixed part for the predictors with initial value for $\hat{\beta}_j$, $\hat{\sigma}$, and $\hat{G}$, the algorithm takes bootstrap samples from the training set to build a forest of trees. The predicted fixed part for observation i in group j, $\hat{f}(X_{ij})$, is obtained with the training set of trees in the forest. Next, it computes $\hat{b}_j$ with the updated estimate of the random part of Equation (4) and updates the variance components $\hat{\sigma}$ and $\hat{G}$. The algorithm keeps repeating those steps until convergence. See Hajjem et al. (2014) for detailed explanation.

RF-based methods rank all predictors by their importance in prediction. Specifically, the importance function of LongituRF package (Capitaine, 2020) in R prints two measures of variable importance: the mean decrease of prediction accuracy when a given variable is permutated (permutation-based importance) and the total decrease in node impurity that results from splits over that variable, averaged over all trees (node impurity-based importance) (James et al., 2013). The permutation-based importance criterion is applied to rank predictors to avoid overfitting.

In the proposed algorithm, the ranking of all predictors is utilized to support predictor selection, which serves as a basis of optimal model selection. By sequentially adding a top-ranked predictor to the provisionary model, the algorithm performs predictor selection. Then, based on the predictors selected, the optimal model is finally identified in the last step. Since educational large-scale assessment (LSA) data typically have hierarchical structures in common, MERF is applied for supporting predictor selection from LSA data. The detailed explanation of each step is followed in the next section.

3 Automated search algorithm for optimal GLMM

In this study, we developed an algorithm to automatically search for optimal GLMMs for any large data set with hierarchical structures, especially LSA data. It mimics how experienced researchers identify the best-fitting GLMM. This algorithm utilizes forward selection (Hastie et al., 2017), which begins with a model containing no predictors, and then adds predictors to the model, one at a time, until complex model with the newly added predictor is not significantly different from simpler model. Unlike traditional forward selection that adds the predictor that gives the greatest additional improvement to the fit to the model, we select the predictor based on the ranking of their importance in prediction sorted by MERF. The proposed algorithm contains three main steps, preparation, predictor selection, and model selection, as demonstrated in Figure 1.

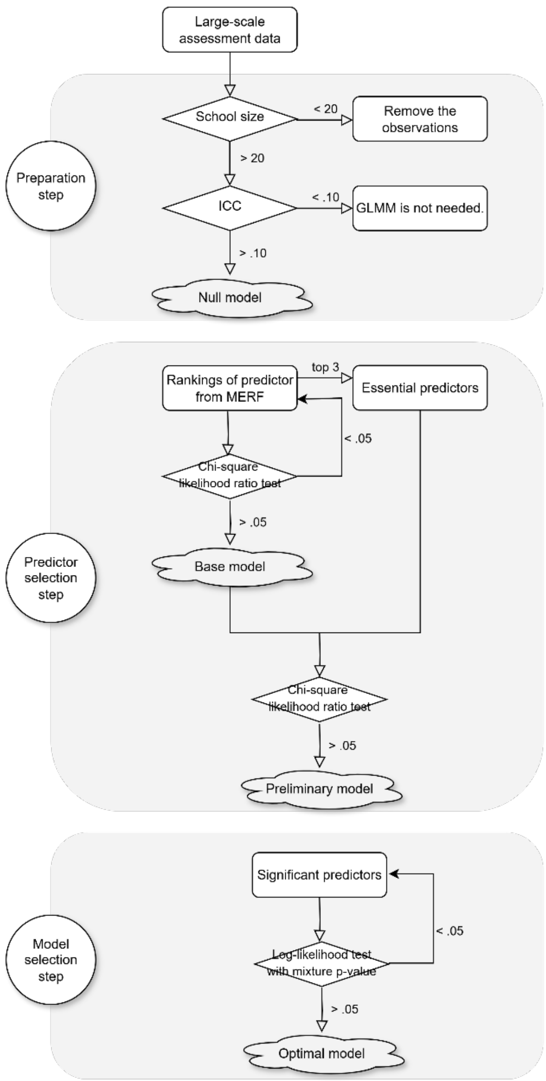


Figure 1: Flowchart of the automated search algorithm.

3.1 Preparation

In the preparation step, data from schools with fewer than 20 students are removed. Next, the size of school effects (i.e., ICC) is measured to decide whether GLMM is needed or not. If the ICC is less than .1, the GLMM is not needed. If the ICC is larger than .1, a random intercept model with no predictors, called the null model, is built. Both the minimum school size and the magnitude of the ICC are tentatively decided, yet they can be set by user.

3.2 Predictor selection

In the predictor selection stage, MERF is applied to rank all predictors of statistical importance, specifically mean decrease of prediction accuracy, from the highest to the lowest. Especially, the three highest ranked predictors are denoted essential predictors, *V1*, *V2*, and *V3*, and further utilized to the selection of possible interaction term(s).

Secondly, the top-ranked predictor is selected among all predictors that are not in the provisional model and added to the model. If the model doesn't converge, the algorithm chooses the next top-ranked predictor instead. Next, a log-likelihood ratio test is performed to determine whether the complex model with the newly added predictor is significantly different from the simpler model. Note that we use forward selection, a greedy algorithm, producing a nested sequence of models. If the complex model with the added predictor is significantly different from the simpler model, the predictor is added to the provisionary model; otherwise, stop predictor selection and fit the GLMM, a random intercept model with all significant predictors. This model is referred to as a base model.

Then, the algorithm searches for significant interaction terms. Three interaction terms of the essential predictors (i.e., *V1*:*V2*, *V1*:*V3*, and *V2*:*V3*) are sequentially added to the provisionary model and tested their significance. If the complex model with the newly added interaction term is significantly different from the simpler model, add the term to the model; otherwise, stop the procedure and identify the current model as the preliminary model.

3.3 Model selection

In the model selection step, the proposed algorithm aims to identify the optimal GLMM. First, it systematically tests each of the significant predictors in the base model as a random slope. To test the number of random effects, e.g., whether a random intercept and random slope model is significantly different from a random intercept model, a likelihood ratio test with mixture *p*-value (Self & Liang, 1987; Stram & Lee, 1994) is conducted.

To be specific, the top-ranked significant predictor is selected and a random slope for the predictor is added to the preliminary model. Then, whether the complex model with the newly added random slope is significantly different from the simpler model (preliminary model) is tested. If it is significant, the random slope is added to the preliminary model. Adding the next ranked significant predictors to the model and testing its significance are repeated until a newly added random slope is not significant; otherwise, stop and identify the current model as the optimal GLMM.

The processes of optimal model selection are summarized as follows.

1. Start with the null model.
2. Rank all predictors based on their importance using machine learning methods (e.g., MERF).
3. Select the top-ranked predictor among all predictors that are not in the provisional model.
4. Add this predictor to the (provisional) model. If the model doesn't converge, choose the next top-ranked predictor, and so on. Conduct a chi-square difference test to determine whether the model with the newly added predictor is significantly different from the simpler model.
5. If there is a significant difference, add the predictor to the current model; otherwise, remove the predictor.
6. Repeat steps 3 to 5 until a newly added predictor is not significant.
7. Add an interaction term of the top-3 ranked predictors, called essential predictors, to the model one-at-a time and test their significance.
8. If there is a significant difference, add the interaction to the model; otherwise, remove the interaction term.
9. Repeat steps 7 to 8 until a newly added interaction term is not significant.
10. Select the top-ranked significant predictor and add a random slope for the predictor to the model. Conduct a likelihood ratio test with mixture *p*-value whether the model with the newly added random slope is significantly different from the simpler model.
11. If there is a significant difference, add the slope to the current model; otherwise, remove the slope.
12. Repeat steps 10 and 11 until a newly added random slope is not significant.
13. Identify the current model as the optimal GLMM.

4. A real data example

We illustrate how the proposed algorithm can be applied to explore optimal GLMMs using real LSA data, the Trends in International Mathematics and Science Study (TIMSS). Specifically, the U.S. eighth grade student data collected in 2019 is utilized to explore optimal GLMM on their achievement scores in mathematics. Note that this section does not intend to compare the proposed algorithm's performance with other existing methods, but to illustrate how the algorithm automatically searches for optimal GLMMs from LSA data with hierarchy step by step. As far as we know, there are no existing methods to perform such a thing so far. Data was already cleaned (e.g., missing data imputation, etc.).

4.1 Preparation

Starting from 232 predictors (independent variables) of 8,698 students, the algorithm examines the schools with less than the minimum number of students and calculates the ICC value to decide whether GLMM is necessary. The minimum number of students and the ICC cut-off values are temporarily set to 20 and 0.1, respectively. We find that there are 44 schools with fewer than 20 students, leading that 600 observations being deleted. The ICC value is 0.44, implying that 44% of total variance in students' achievement scores in mathematics is explained by school differences. Thus, GLMM is needed.

Next, the algorithm fits a random intercept model without any predictors, also denoted as null model (Equation 5),

$$y_{ij} = 515 + b_{0j} + \varepsilon_{ij}. \quad (5)$$

$i = 1, 2, \dots, I_j, J = 229$, $b_{0j} \sim N(0, 4{,}027)$, $\varepsilon_{ij} \sim N(0, 5{,}239)$, where y_{ij} referring to students' achievement scores in mathematics.

4.2 Predictor selection

In this step, the algorithm aims to search for the best base model, which is a random intercept model including all significant predictors. First, a machine learning method, MERF, ranks all 232 predictors based on their importance. We set hyper-parameters: the number of trees (*ntree*) is set to 2,000, and the number of variables randomly sampled as candidates at each split (*mtry*) is to 73, the floor of the total number of predictors divided by 3, as recommended by Breiman (2001). Note that as *ntree* gets larger, the more stable predictive error can be obtained, at the expense of computational efficiency.

The importance plot of the top-30 ranked predictors obtained by MERF is illustrated in Figure 2. Their names are abbreviated as *V1* to *V30* by their ranks and the original names are presented in Table 1.

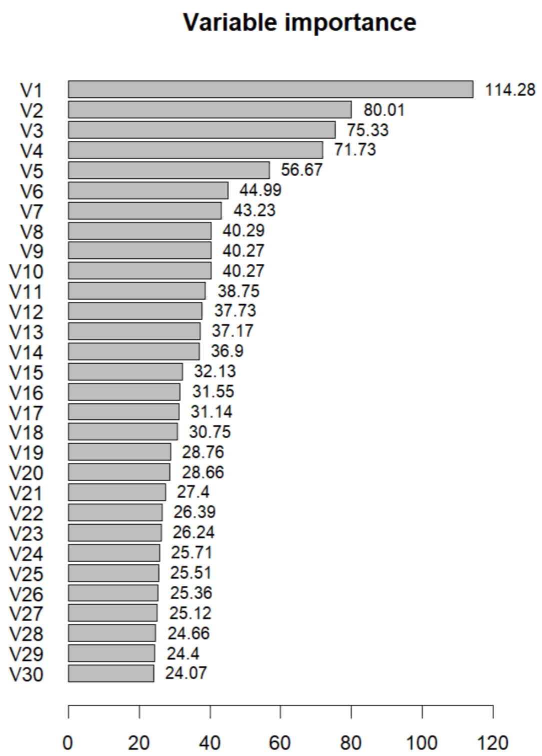


Figure 2: Predictor importance plot.

Then, all the predictors are tested for their significance one-at-a time. The algorithm fits the base model with 18 significant predictors (*V1* to *V18*), including two school-level predictors. A likelihood ratio test is conducted to compare the null model with the base model. The result indicates that the addition of the predictors significantly improved model fit, $\Delta\chi^2(18) = 3366.70$, $p < .001$ (see, Table 2). The model's R_1^2 and R_2^2 are 0.33 and 0.67, respectively. Note that *V9* is no longer significant in the base model.

No.	Variable	No.	Variable
V1	BSBG04	V16	BSBM18F
V2	BSBM19C	V17	BSBM18E
V3	BCBG03A	V18	BSBM15
V4	BSBM19A	V19	BSBS24C
V5	BSBG07	V20	BCBG06B
V6	BSBM19B	V21	BCDGTIHY
V7	BSBM19F	V22	BCBG03B
V8	BSBM19H	V23	BSBM26AA
V9	BCDGSBC	V24	BSBM16I
V10	BSDGEDUP	V25	BSBS27BB
V11	BSBM18C	V26	BSBE03E
V12	BSBM19D	V27	BSBS43BB
V13	BSDAGE	V28	BSBE03F
V14	BSBS24B	V29	BSBM18D
V15	BSBS24G	V30	BSBE01A

Table 1: Predictor names.

Model	*df*	AIC	logLik	$\Delta\chi^2$	*p*
Null	3	93095	-46544		
Base	21	89764	-44861	3367	<.001
Prelim-	22	89759	-44858	7	<.01

Table 2: Model comparison (step 2).

Next, the algorithm sequentially adds the interaction terms of the essential predictors. The result of likelihood test shows that an interaction term, *V1:V2*, significantly improved model fit, $\Delta\chi^2(1) = 6.77, p < .01$. We denote the model with the added interaction term as the preliminary model. The models' R^2 at both levels are also slightly improved.

4.3 Model selection

Based on the preliminary model in the previous step, the algorithm searches for the optimal random slopes. As a result, a random slope for *V1* is added to the preliminary model. The likelihood ratio test with mixture *p*-value indicates that the complex (called intermediate) model with the newly added slope is significantly different from the simpler (preliminary) model, $\Delta\chi^2(2) = 56.69, p$ (mixture)

Model	*df*	AIC	logLik	$\Delta\chi^2$	p^*
Prelim-	22	89759	-44858	7	
Interim-	24	89707	-44829	57	<.001
Optimal	27	89672	-44809	41	<.001

Table 3: Model comparison (step 3).

< .001 (see, Table 3). Note that the interaction term (*V1*:*V2*) is no longer significant.

Then, a random slope for *V2* is newly added to intermediate model and the likelihood test with mixture *p*-value also demonstrates that the newly added slope significantly improved model fit, $\Delta\chi^2(3) = 41.00, p$ (mixture) < .001 (see, Table 3). Since the random slope for *V3* is not significant, we stop model searching and the provisionary model is identified as the optimal GLMM.

The optimal GLMM has a random intercept, 18 predictors (two school-level predictors, 16 student-level predictors), an interaction term, and two random slopes for *V1* and *V2*, as shown in Equation (6),

$$y_{ij} = 603 + 9(V1)_{ij} + 6(V2)_{ij} - 17(V3)_j - 12(V4)_{ij} + 7(V5)_{ij} + 5(V6)_{ij} - 3(V7)_{ij} + 5(V8)_{ij} - 4(V9)_j + (V10)_{ij} + 5(V11)_{ij} - 3(V12)_{ij} - 11(V13)_{ij} + 5(V14)_{ij} + 6(V15)_{ij} + 3(V16)_{ij} + 2(V17)_{ij} - 8(V18)_{ij} + (V1)_{ij} \times (V2)_{ij} + b_{1j}(V1)_{ij} + b_{2j}(V2)_{ij} + b_{0j} + \varepsilon_{ij}. \quad (6)$$

$i = 1, 2, \ldots, I_j, J = 229, \varepsilon_{ij} \sim N(0, 3381)$,

$$\begin{pmatrix} b_{0j} \\ b_{1j} \\ b_{2j} \end{pmatrix} = b_j \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, G = \begin{pmatrix} 2080 & - & - \\ -178 & 52 & - \\ -152 & 28 & 48 \end{pmatrix}\right).$$

Note that the interaction term (*V1*:*V2*) is no longer significant in the optimal GLMM. The proposed algorithm is likely to find slightly different optimal GLMM, due to randomness of predictors' importance ranking obtained by MERF. A user can ensure consistent results across runs by setting a random seed.

5. Scientific Importance

It is crucial for researchers to build an adequate optimal model to make valid statistical inferences. Identifying the best-fitting GLMM is more time-consuming and complex than finding the best generalized linear model (GLM), because GLMM also includes random effects. This algorithm automatically evaluates a large number of models during the process of building an optimal GLMM model. One of the major components is a machine learning approach (e.g., MERF), which is applied to sort all predictors based on their importance, allowing for efficient predictor selection.

In addition, all available predictors from LSA data can be utilized in searching for optimal GLMMs without convergence problems using the algorithm developed here. It also provides a systematic and transparent process that can be produced by others, for example, a random intercept model is fitted, interaction terms of essential predictors are searched, and then random slopes are sequentially added to the model. The proposed algorithm has the potential to reduce the time and effort required by researchers and to provide guidelines for exploring the optimal GLMMs. We further update this algorithm by taking more considerations into account to explore best-fitting GLMMs.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv*:1506.04967. https://arxiv.org/pdf/1506.04967

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25. https://doi.org/10.2307/2290687

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65–108. http://www.jstor.org/stable/1084940

Capitaine, L. (2020). *LongituRF: random forests for longitudinal data*. R package version 0.9.

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84*(6), 1313-1328. https://doi.org/10.1080/00949655.2012.741599

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv*. https://doi.org/10.48550/arXiv.1707.08692

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Routledge.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605–610. https://doi.org/10.1080/01621459.1987.10478472

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*, 1171–1177. https://doi.org/10.2307/2533455

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69. https://modir3-3.ir/article-english/article-m2568.pdf

Exploring the Psychometric Validity of AI-Generated Student Responses: A Study on Virtual Personas' Learning Motivation

Huanxiao Wang
Division of Human Development & Quant Methods
University of Pennsylvania
Philadelphia, PA, USA
Contact: huanxiao0210@gmail.com

Abstract

This study explores whether large language models (LLMs) can simulate valid student responses for educational measurement. Using GPT-4o, 2000 virtual student personas were generated. Each persona completed the Academic Motivation Scale (AMS). Factor analyses(EFA and CFA) and clustering showed GPT-4o reproduced the AMS structure and distinct motivational subgroups.

1 Introduction

In psychometric research, collecting real human responses is essential for developing and validating psychological scales. Traditional item development requires large datasets to estimate item parameters, assess factor structures, and revise items based on empirical evidence. However, collecting human data is time-consuming, costly, and sometimes limited by ethical or logistical concerns.

With the rise of generative AI models, some researchers have explored using Large Language Models (LLMs) to simulate response data for psychological research. These models can generate both participant profiles (personas) and their responses to standardized instruments. For example, De Winter et al. (2024) used ChatGPT-4 to generate 2000 virtual personas who completed multiple personality tests. Other studies have applied similar methods in areas such as personality assessment (Argyle et al., 2023), item generation (Bhandari et al., 2024), and clinical diagnosis (Cook et al., 2024). These early findings suggest that LLMs may reproduce some aspects of human psychological variability, but the extent of their validity remains an open question. Liu et al. (2025) found that LLM-generated responses cannot fully substitute human respondents in all aspects of item-level psychometric performance.

In the field of educational psychology, relatively few studies have examined the use of generative AI models to simulate student motivation data. Learning motivation is commonly assessed through self-report instruments such as the Academic Motivation Scale (AMS; Vallerand et al., 1992) and the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991). Among these, AMS is widely used to measure motivation based on Self-Determination Theory (SDT; Deci & Ryan, 2000), covering intrinsic motivation, extrinsic regulation, and amotivation dimensions. The AMS has been validated in many populations and contexts (Bacanlı & Sahinkaya, 2011; Barkoukis et al., 2008; Fairchild et al., 2004; Guay et al., 2014; Stover et al., 2012; Utvær & Haugan, 2016), but it remains unclear whether AI-generated data can replicate its established factor structure.

This study investigates whether GPT-4o can simulate plausible student responses on the AMS, and whether the generated data exhibit acceptable psychometric properties. Specifically, this study asks:

Q1: Can GPT-generated responses reproduce the expected 7-factor structure of AMS, including IMTK, IMTA, IMES, EMID, EMIN, EMEX, and AMOT subscales?

Q2: Can GPT-generated persona descriptions be clustered into meaningful subgroups, and do these subgroups show distinct response patterns across AMS subscales?

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 359–366

2 Related Works

2.1 Measuring Learning Motivation

Student motivation has been widely studied in educational psychology. Several validated scales are used to measure motivation constructs. The Academic Motivation Scale (AMS; Vallerand et al., 1992) is one of the most widely used instruments, grounded in Self-Determination Theory (SDT; Deci & Ryan, 2000). The AMS assesses seven subtypes of academic motivation, including intrinsic motivation to know (IMTK), intrinsic motivation to accomplish (IMTA), intrinsic motivation to experience stimulation (IMES), identified regulation (EMID), introjected regulation (EMIN), external regulation (EMEX), and amotivation (AMOT). This structure has been validated across different populations and cultural contexts (Bacanlı & Sahinkaya, 2011; Barkoukis et al., 2008; Fairchild et al., 2004; Guay et al., 2014; Stover et al., 2012; Utvær & Haugan, 2016).

However, even for well-established scales like AMS, researchers typically require large-scale human response data to examine factor structure, reliability, and construct validity. This process can be resource-intensive, and difficult to replicate across diverse samples.

2.2 AI-Generated Psychometric Data

With the development of generative AI models, researchers have begun to explore using large language models to simulate human response data. De Winter et al. (2024) demonstrated that ChatGPT-4 can generate thousands of virtual personas who complete various personality inventories. Similar methods have been applied to simulate item-level responses in personality assessment (Argyle et al., 2023), clinical measurement (Cook et al., 2024), and item generation for educational testing (Bhandari et al., 2024). These studies suggest that LLMs may capture certain aspects of psychological variability.

However, the validity of AI-generated psychometric data remains uncertain. Liu et al. (2025) found that while LLM-generated respondents may mimic some human response patterns, they cannot fully substitute for human data, particularly when analyzing fine-grained item functioning. Moreover, most prior studies have focused on personality or clinical scales. Applications of LLMs in simulating student motivation data remain scarce.

2.3 Research Gap

Although early work suggests LLMs have some capacity to generate psychologically meaningful data, few studies have systematically tested whether AI-generated responses can reproduce complex factor structures of educational motivation instruments like AMS. In addition, the combination of LLM-based persona generation and psychometric analysis has not been fully explored in this domain. This study addresses this gap by evaluating whether GPT-4o can simulate realistic AMS responses, and whether persona embeddings can reveal subgroup patterns consistent with motivation theory

3 Methods

3.1 Participants and Data Generation

In this study, no real human participants were recruited. Instead, all data were generated through simulated personas using the GPT-4o model. The generation process included two stages: persona creation and questionnaire response simulation.

3.2 Personas Generation

The design of the persona prompt was informed by prior work using ChatGPT to create fictional respondents for psychological surveys (De Winter et al., 2024). The process was not a strict iterative protocol, but the final version was tested informally to ensure that the personas were coherent and diverse.

The prompt structure also reflects the RISE framework of prompt engineering. It defined a clear *Role* for the model (generate fictional students), specified the *Input* (age, gender, and a short profile), outlined the *Steps* (produce three descriptive sentences), and set the *Expectations* (concise one-line outputs). This helped create consistent personas while still allowing variation across individuals.

First, 2000 virtual student personas were generated using GPT-4o (temperature = 1). Each persona included three elements: age (18-25), gender, and a short description (3 sentences) summarizing their academic personality, learning style, and motivational tendencies. The generation prompt was structured to ensure diversity across motivational profiles while maintaining coherence within each persona. The personas were returned as text files for further processing. The prompts used in this stage are shown below:

Generate 20 fictional student personas. Each should include:

- Age (18–25)

- Gender

- A 3-sentence description of their academic personality, learning style, and motivation.

Each persona should be on one line, like:

0001. 20, Female - Loves collaborative learning; often uses concept maps to organize her thoughts; tends to get anxious during exams.

Only return the 20 personas, nothing else.

To generate the full dataset, GPT-4o was called repeatedly in batches of 20 personas per request. In total, 100 batches were generated to produce 2000 unique persona descriptions.

It is important to clarify that the inclusion of "learning styles" in the persona descriptions does not reflect a theoretical endorsement of this concept. The idea that students learn best in their preferred style has been widely challenged in the literature (Nancekivell et al., 2019). In this study, learning style phrases were used only to enrich the variety and naturalness of the personas. They were not analyzed as variables and did not influence the psychometric results.

3.3 AMS Responses Generation

After generating the personas, each simulated student was asked to complete the Academic Motivation Scale (AMS), which consists of 28 items rated on a 7-point Likert scale (1 = Does not correspond at all, 7 = Corresponds exactly). For response generation, GPT-4o was instructed to simulate AMS item-level answers based on the persona descriptions. To minimize randomness in item response generation, the temperature was set to 0. The model returned raw item responses as a list of integers for each persona. Here is the prompt used in this stage.

Imagine the following student: (personas),

This student is now responding to the Academic Motivation Scale (AMS).

There are 28 items, each rated from 1 (Does not correspond at all) to 7 (Corresponds exactly).

(28 full items),

Please return exactly 28 integers separated only by commas. No explanation, no labels. Just the numbers.

The full item texts were embedded into the prompt to ensure that GPT-4o received the exact questionnaire content for each response simulation.

3.4 Psychometric Analysis (EFA and CFA)

Exploratory factor analysis (EFA) was conducted to examine whether the GPT-generated AMS responses reproduced the expected factor structure. All 28 AMS items were included in the analysis. The factor extraction was performed using principal axis factoring. The number of factors was determined based on parallel analysis and scree plot inspection. Promax rotation was applied to allow correlated factors. Factor loadings were evaluated to assess whether items loaded onto the intended subscales.

Confirmatory factor analysis (CFA) was used to test the fit of the established seven-factor structure of AMS. Each item was assigned to its corresponding subscale based on the original AMS theoretical model (Vallerand et al., 1992). The model included the following factors: IMTK, IMTA, IMES, EMID, EMIN, EMEX, and AMOT. CFA was conducted using maximum likelihood estimation. Model fit was evaluated using common fit indices: Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Factor loadings were examined to assess item performance within each factor.

3.5 Semantic Clustering of GPT-Generated Personas

In addition to analyzing AMS responses, persona descriptions generated by GPT-4o were analyzed to identify motivational subgroups. The persona texts were vectorized using GPT-4o embedding models (text-embedding-3-small). The resulting embeddings represented the semantic information contained in each persona description. K-means clustering was applied to the embeddings to partition the personas into three clusters (k = 3). The choice of three clusters was based on initial exploratory analysis and interpretability considerations.

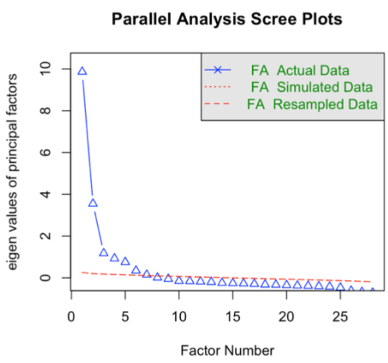


Figure 1: Parallel Analysis Scree Plot for Factor Extraction.

Factor	Item	Standardized Loading
IMTK (Intrinsic Motivation - To Know)	AMS_Q2	0.805
	AMS_Q9	0.938
	AMS_Q16	0.953
	AMS_Q23	0.857
IMTA (Intrinsic Motivation - Toward Accomplishment)	AMS_Q6	0.791
	AMS_Q13	0.582
	AMS_Q20	0.818
	AMS_Q27	0.778
IMES (Intrinsic Motivation - Experience Stimulation)	AMS_Q4	0.417
	AMS_Q11	0.926
	AMS_Q18	0.980
	AMS_Q25	0.899
EMID (Extrinsic Motivation - Identified Regulation)	AMS_Q3	0.756
	AMS_Q10	0.938
	AMS_Q17	0.896
	AMS_Q24	0.891
EMIN (Extrinsic Motivation - Introjected Regulation)	AMS_Q7	0.852
	AMS_Q14	0.814
	AMS_Q21	0.851
	AMS_Q28	0.807
EMEX(Extrinsic Motivation - External Regulation)	AMS_Q1	0.758
	AMS_Q8	0.891
	AMS_Q15	0.863
	AMS_Q22	0.978
AMOT (Amotivation)	AMS_Q5	0.731
	AMS_Q12	0.352
	AMS_Q19	0.833
	AMS_Q26	0.322

Table 1: Confirmatory Factor Analysis Results for GPT-Generated AMS Responses.

3.6 Subgroup Responses Analysis

After clustering, subscale scores for each of the seven AMS factors were calculated for each persona. Subgroup differences across clusters were analyzed to examine whether the clustering structure aligned with meaningful motivational patterns. Boxplots were used to visualize subscale score distributions across clusters. Non-parametric tests (e.g., Kruskal-Wallis tests) were performed to evaluate the statistical significance of subgroup differences on each AMS subscale.

All analyses were conducted in R (version 4.2.3; R Core Team, 2023) using RStudio (version 2025.05.1+513; Posit Software, PBC).

4 Results and Discussions

4.1 EFA and CFA

A parallel analysis was conducted to determine the appropriate number of factors. The analysis suggested that seven factors should be extracted, fully consistent with the original structure of the AMS. This result indicates that the GPT-generated responses preserved the intended dimensional structure of academic motivation as specified by Self-Determination Theory. Figure 1 shows parallel analysis scree plot for factor extraction.

The scree plot suggested a dominant first factor, followed by a sharp decline after the second factor. This result reflects a potential tendency of GPT-generated data to compress variance into fewer principal components, possibly due to the semantic coherence of AI-simulated responses.

Factor	Item	Standardized Loading
IMTK (Intrinsic Motivation - To Know)	AMS_Q2	0.805
	AMS_Q9	0.938
	AMS_Q16	0.953
	AMS_Q23	0.857
IMTA (Intrinsic Motivation - Toward Accomplishment)	AMS_Q6	0.791
	AMS_Q13	0.582
	AMS_Q20	0.818
	AMS_Q27	0.778
IMES (Intrinsic Motivation - Experience Stimulation)	AMS_Q4	0.417
	AMS_Q11	0.926
	AMS_Q18	0.980
	AMS_Q25	0.899
	AMS_Q3	0.756

Factor	Item	Standardized Loading
EMID (Extrinsic Motivation - Identified Regulation)	AMS_Q10	0.938
	AMS_Q17	0.896
	AMS_Q24	0.891
EMIN (Extrinsic Motivation - Introjected Regulation)	AMS_Q7	0.852
	AMS_Q14	0.814
	AMS_Q21	0.851
	AMS_Q28	0.807
EMEX(Extrinsic Motivation - External Regulation)	AMS_Q1	0.758
	AMS_Q8	0.891
	AMS_Q15	0.863
	AMS_Q22	0.978
AMOT (Amotivation)	AMS_Q5	0.731
	AMS_Q12	0.352
	AMS_Q19	0.833
	AMS_Q26	0.322

Table 1 specifies what font sizes and styles must be used for each type of text in the manuscript.

A confirmatory factor analysis (CFA) was conducted to test the fit of the theoretical seven-factor structure of AMS. The CFA model included all seven subscales: IMTK, IMTA, IMES, EMID, EMIN, EMEX, and AMOT. The model demonstrated acceptable fit: CFI = 0.908, TLI = 0.894, RMSEA = 0.082, and SRMR = 0.065. These indices suggest that the GPT-generated responses generally reproduced the expected factor structure of AMS, although the fit was not perfect.

Standardized factor loadings were strong for most items, particularly for intrinsic and identified motivation subscales. For example, IMTK items loaded between 0.81 and 0.95, and EMID items loaded between 0.76 and 0.94. In contrast, several AMOT items displayed lower or unstable loadings (e.g., AMS_Q12 = 0.35; AMS_Q26 = 0.32). This pattern suggests that GPT-4o simulated positively valenced motivation dimensions more consistently than disengagement states such as amotivation.

The EFA and CFA results suggest that GPT-4o can partially reproduce the established factor structure of the AMS. While the model fit is not perfect, the seven-factor structure generally holds, particularly for intrinsic and identified motivation subscales. The relatively weaker performance on amotivation items may reflect GPT-4o's default bias toward goal-directed, coherent outputs, which may limit its ability to fully simulate psychological disengagement. Similar limitations have been noted in previous LLM-based simulation studies (Liu et al., 2025). Overall, these results provide preliminary evidence that LLM-generated response data may capture key aspects of psychological constructs but may require further refinement when modeling negative or conflictual motivational states.

4.2 Semantic Clustering of GPT-Generated Personas

To explore potential subgroups in the GPT-generated student personas, semantic clustering was conducted based on persona descriptions. Each persona text was vectorized using GPT-4o embeddings (text-embedding-3-small), and k-means clustering was applied. The number of clusters was set to k = 3 based on interpretability and preliminary exploration.

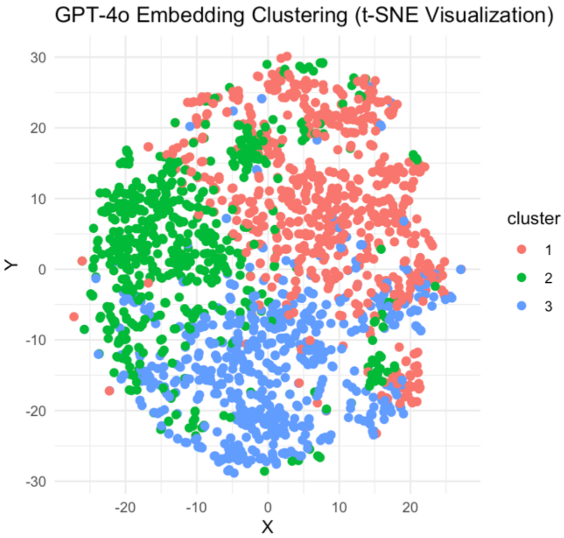


Figure 2: Semantic Clustering of GPT-Generated Personas Using t-SNE Visualization

A t-SNE visualization was generated to display the cluster separation in two dimensions. The clusters were well-separated, suggesting that GPT-generated persona descriptions contained distinct semantic features that could differentiate students into subgroups.

Subscale scores for the seven AMS factors were calculated for each persona. Boxplots were created to compare AMS subscale scores across clusters. Results showed that cluster membership was associated with different motivational profiles.

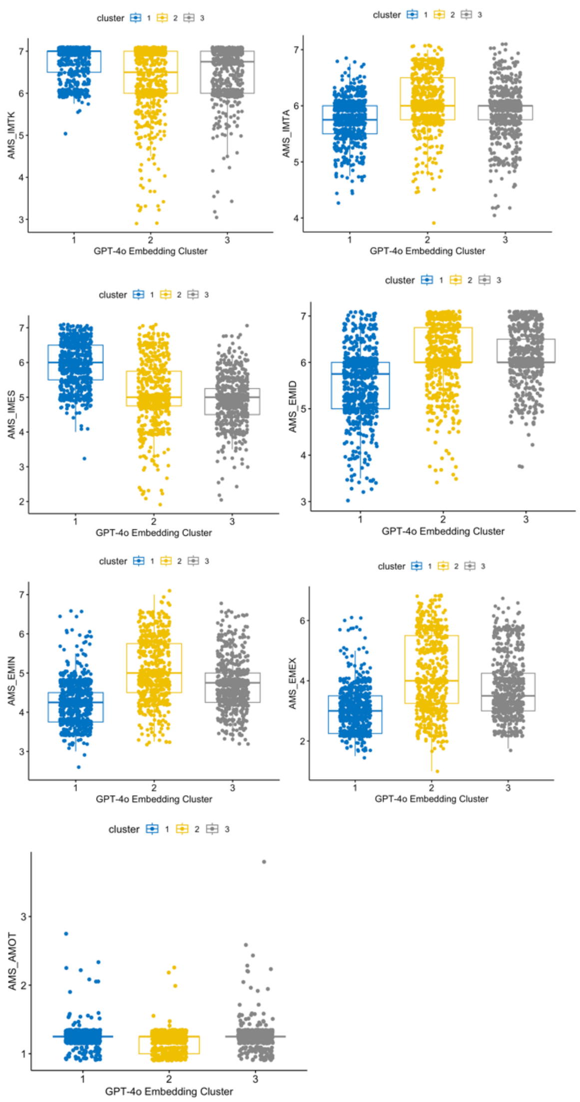


Figure 3: Subscale Score Differences Across GPT-4o Semantic Clusters

Cluster 1 displayed higher scores on intrinsic motivation subscales (IMTK, IMTA, IMES), suggesting students with strong intrinsic academic interests. Cluster 2 showed moderate levels of intrinsic and extrinsic motivation. Cluster 3 demonstrated slightly elevated external regulation (EMEX) and lower intrinsic motivation, suggesting more externally driven or performance-oriented students.

Kruskal-Wallis tests confirmed significant differences across clusters for most AMS subscales ($p < .001$ for IMTK, IMTA, EMEX, AMOT; $p < .01$ for IMES, EMID, EMIN), indicating that semantic clustering based on persona descriptions corresponded meaningfully with simulated questionnaire responses.

The clustering results suggest that GPT-4o not only generated item-level responses consistent with AMS factor structure, but also produced semantically rich persona profiles that reflected distinct motivational orientations. Subgroups identified through semantic embeddings showed systematic differences across AMS subscales, supporting the convergent validity between generated persona characteristics and questionnaire outcomes.

These findings demonstrate that large language models may capture latent psychological patterns even before formal scale administration, purely based on persona-level text descriptions. This capability may have potential applications for early-phase scale development, where synthetic data may help evaluate item functioning across diverse hypothetical profiles prior to human data collection. Researchers can use this approach to screen for problematic items, evaluate whether expected factor structures emerge, and explore subgroup patterns across hypothetical profiles. Such applications may reduce costs, speed up validation cycles, and support scale adaptation in new contexts. Theoretically, this work also contributes to ongoing debates about the extent to which LLMs can approximate latent psychological constructs. It shows both the potential and the current limits of AI personas in capturing human-like motivational patterns.

However, it should be noted that GPT-generated clusters may reflect idealized or overly coherent motivational types, as the model tends to generate consistent and goal-oriented outputs. Additional validation with real human data is needed to fully assess the generalizability of these subgroup structures.

5 Limitations and Future Directions

While the present study demonstrates the promising potential of large language models (LLMs) like GPT-4o in generating psychologically plausible student response data, several limitations should be acknowledged.

First, although the generated responses reproduced the theoretical factor structure of AMS reasonably well, the confirmatory factor analysis still yielded only moderate model fit (e.g., RMSEA = 0.082). This suggests that LLM-generated data may not fully replicate the nuanced variance found in real human populations. In particular, the amotivation (AMOT) subscale consistently showed weaker or unstable loadings, which may reflect GPT-4o's inherent difficulty in simulating disengaged or conflicted psychological states. This aligns with prior observations that LLMs tend to default toward coherent, goal-oriented, and positively valenced outputs (e.g., Liu et al., 2025).

Second, the current study focused on only one questionnaire (AMS) and one LLM model (GPT-4o). The generalizability of these findings to other constructs, instruments, or LLM architectures remains unclear. Expanding this approach to include additional validated scales (e.g., MSLQ, AEQ) and cross-model comparisons would help clarify whether the observed psychometric patterns are robust across different psychological domains.

Third, the clustering analysis relied solely on text embeddings of GPT-generated persona descriptions. While meaningful subgroups were identified, these clusters are not directly validated against real-world student samples. Future work should compare LLM-generated subgroup structures to empirical cluster solutions obtained from human data to assess alignment and potential biases.

Finally, this study examined LLM-generated responses under controlled prompting conditions, using fixed temperature settings and instruction formats. Prompt engineering decisions likely play a crucial role in shaping response variability and latent structure reproduction. Future research should systematically investigate how prompt design, randomness parameters, and persona context framing influence the psychometric properties of generated data.

Another point worth noting is that the use of learning style descriptors in the persona prompts should not be read as a validation of the learning styles hypothesis. Like we mentioned before, the concept has been widely debated and is not supported by strong empirical evidence (Nancekivell et al., 2019). Here it served only as a descriptive element to make the personas sound more realistic, and it had no bearing on the psychometric findings.

Despite these limitations, this study offers a novel empirical demonstration of how generative AI can contribute to early-stage scale development and psychometric exploration. As LLM capabilities continue to evolve, careful validation studies combining both real and simulated data will be critical for evaluating the responsible integration of AI tools in psychological measurement. The findings have both theoretical and practical implications. Theoretically, they suggest that large language models can reproduce complex motivational structures like those in AMS, although with biases toward positive and coherent states. This adds to current discussions in psychometrics about whether AI can model latent constructs. Practically, the study points to the value of AI personas as a tool for instrument testing and development. With further refinement, such simulations may help researchers reduce the cost and time of scale validation, while also expanding opportunities to explore item functioning across diverse cultural or contextual settings.

Acknowledgments

The author declares no conflict of interest. The datasets generated during this study are available from the corresponding author upon reasonable request. This paper was completed as part of an independent study during the author's master's program; all views and any errors are the author's own and do not represent the University of Pennsylvania. The author also used OpenAI's ChatGPT to assist in polishing the language of the manuscript.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bacanlı, H., & Sahinkaya, O. (2011). The Adaptation Study of Academic Motivation Scale into Turkish. *Procedia - Social and Behavioral Sciences, 12*, 562–567. https://doi.org/10.1016/j.sbspro.2011.02.068

Barkoukis, V., Tsorbatzoudis, H., Grouios, G., & Sideridis, G. (2008). The assessment of intrinsic and extrinsic motivation and amotivation: Validity and reliability of the Greek version of the Academic Motivation Scale. *Assessment in Education Principles Policy and Practice, 15*(1), 39–55. https://doi.org/10.1080/09695940701876128

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education Artificial Intelligence, 7*, 100284. https://doi.org/10.1016/j.caeai.2024.100284

Cook, D. A., Overgaard, J., Pankratz, V. S., Del Fiol, G., & Aakre, C. A. (2024). Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialog with feedback (Preprint). *Journal of Medical Internet Research.* https://doi.org/10.2196/68486

De Winter, J. C., Driessen, T., & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences, 228,* 112729. https://doi.org/10.1016/j.paid.2024.112729

Fairchild, A. J., Horst, S. J., Finney, S. J., & Barron, K. E. (2004). Evaluating existing and new validity evidence for the Academic Motivation Scale. *Contemporary Educational Psychology, 30*(3), 331–358. https://doi.org/10.1016/j.cedpsych.2004.11.001

Guay, F., Morin, A. J. S., Litalien, D., Valois, P., & Vallerand, R. J. (2014). Application of exploratory structural equation modeling to evaluate the academic motivation scale. *The Journal of Experimental Education, 83*(1), 51–82. https://doi.org/10.1080/00220973.2013.876231

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025b). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13570

Nancekivell, S. E., Shah, P., & Gelman, S. A. (2019). Maybe they're born with it, or maybe it's experience: Toward a deeper understanding of the learning style myth. *Journal of Educational Psychology*, *112*(2), 221–235. https://doi.org/10.1037/edu0000366

Pintrich, P & Smith, D. & Duncan, Teresa & Mckeachie, Wilbert. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). Ann Arbor. Michigan. 48109. 1259.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American *Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066x.55.1.68

Stover, J., De La Iglesia, N., Rial, A., & Liporace, N. F. (2012). Academic Motivation Scale: adaptation and psychometric analyses for high school and college students. *Psychology Research and Behavior Management,* 71. https://doi.org/10.2147/prbm.s33188

Utvær, B. K. S., & Haugan, G. (2016). The Academic Motivation Scale: dimensionality, reliability, and construct validity among vocational students. Nordic *Journal of Vocational Education and Training,* 17–45. https://doi.org/10.3384/njvet.2242-458x.166217

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: a measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement, 52*(4), 1003–1017. https://doi.org/10.1177/0013164492052004025

Measuring Teaching with LLMs

Michael Hardy
Stanford University
hardym@stanford.edu

Abstract

Objective and scalable measurement of teaching quality is a persistent challenge in education. While Large Language Models (LLMs) offer potential, general-purpose models have struggled to reliably apply complex, authentic classroom observation instruments. This paper uses custom LLMs built on sentence-level embeddings, an architecture better suited for the long-form, interpretive nature of classroom transcripts than conventional subword tokenization. We systematically evaluate five different sentence embeddings under a data-efficient training regime designed to prevent overfitting. Our results demonstrate that these specialized models can achieve human-level and even super-human performance with expert human ratings above 0.65 and surpassing the average human-human rater correlation. Further, through analysis of annotation context windows, we find that more advanced models—those better aligned with human judgments—attribute a larger share of score variation to lesson-level features rather than isolated utterances, challenging the sufficiency of single-turn annotation paradigms. Finally, to assess external validity, we find that aggregate model scores align with teacher value-added measures, indicating they are capturing features relevant to student learning. However, this trend does not hold at the individual item level, suggesting that while the models learn useful signals, they have not yet achieved full generalization. This work establishes a viable and powerful new methodology for AI-driven instructional measurement, offering a path toward providing scalable, reliable, and valid feedback for educator development.[1]

1 Introduction

Measuring teaching quality is hard (Jurenka et al., 2024; Kane and Staiger, 2012; Ho and Kane, 2013). Despite their ubiquity as the primary form of teacher development and evaluation, ratings of instructional quality, even when conducted by trained experts, have low reliability, unknown accuracy, and are very expensive to conduct (Ho and Kane, 2013; Kane et al., 2015; Kane and Staiger, 2012; Glaese et al., 2022; Whitehill and LoCasale-Crouch, 2024; Whitehurst et al., 2014; Jurenka et al., 2024; Tack et al., 2023; Grissom et al., 2013; Liu and Cohen, 2021). Recent work has sought to reduce the costs of these evaluations using large language models (LLMs) to annotate spoken discourse in classrooms to support such evaluations on actual instruments used with educators, but such models have not yet shown the ability to help with these complex tasks (Wang and Demszky, 2023; Whitehill and LoCasale-Crouch, 2024; Xu et al., 2024; Hardy, 2024). This study builds on these studies, answering calls to do more to evaluate the capacity of LLMs for classroom tasks (Casabianca et al., 2013; Liu and Cohen, 2021).

This study investigates whether pretrained contextual embeddings at the sentence level can meaningfully capture classroom dialogue for automated assessment of instructional quality. We systematically evaluate multitask encoder models trained on fixed sentence-level embeddings to predict expert human ratings of teaching effectiveness across 25 distinct instructional dimensions. We achieve state-of-the-art performance on this task, surpassing existing human benchmark correlations while providing novel insights into the training dynamics of multitask models applied to educational discourse.

Our findings have significant implications for educational assessment and the broader application of NLP methods to specialized domains. We show that shared-weight multitask architectures initially learn general representations of teaching quality that align well with student outcomes, but continued training may lead to overfitting to noisy human

[1] https://github.com/hardy-education/measuring$_teaching_encoders$

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 367–384
October 27-29, 2025

annotations on individual constructs rather than the underlying pedagogical constructs of interest. These insights suggest new directions for developing robust AI systems for classroom analysis and highlight fundamental challenges in aligning automated assessments with meaningful educational outcomes.

1.1 Primary Contributions

1. **A longitudinal analysis of representation learning for teaching quality**. We provide the first systematic evidence of how different sentence-embedding LLMs develop an understanding of effective instruction throughout their training process.

2. **A new benchmark for automated instructional rating**. We demonstrate state-of-the-art performance, outperforming existing models by achieving the highest human-expert correlation across 25 distinct instructional dimensions.

3. **A critique of single-turn evaluation**. Through the first analysis of score stability over time, we show that static, single-turn evaluations are insufficient and introduce a more robust, temporally-aware method for assessing LLM-based ratings.

4. **A validation framework linking ratings to student achievement**. We introduce the first methodology to directly measure the alignment between an LLM's ratings of teaching and externally-validated measures of student learning gains (teacher value-added).

2 Related Work and Motivation

2.1 Sentence Embeddings

The robustness of pre-trained language models to out-of-distribution text remains an open question, particularly for specialized domains such as educational settings where child speech patterns and pedagogical discourse structures predominate. To address this challenge, we investigate whether sentence-level embeddings (Reimers and Gurevych, 2019) can provide more stable representations of classroom language than traditional subword tokenization approaches, despite the potential loss of fine-grained semantic information. We test the large versions of each of the following pre-trained sentence embeddings: Unsupervised SimCSE and Supervised SimCSE (Gao et al., 2022), E5 (Wang et al., 2024a), Multilingual E5 (Wang et al., 2024b), GTE (Li et al., 2023), and a contrast fine-tuned RoBERTa model released with sentence-transformers (Liu et al., 2019; Reimers and Gurevych, 2019). This diversity in embedding approaches also allows us to investigate potential biases in model interpretation across different linguistic communities and teaching contexts.

2.2 Teacher Development and Evaluation

School leaders working with teachers to improve the quality of instruction typically evaluate the teacher's proficiency in a range of competencies (typically measured during in-class observation and evaluation on a teaching rubric; (Aguilar, 2013; Bambrick-Santoyo, 2016, 2018)), a process that is often time-consuming and produces ratings (labels) that are unreliable (Kane and Staiger, 2012; Blazar, 2018; Kane et al., 2013; Casabianca et al., 2013). Without accurate classifications, it is challenging for practitioners to prioritize instructional needs and aligned practices from among the many elements of good teaching (Saphier et al., 2008; Darling-Hammond, 2014; Hammond, 2015; Lemov and Atkins, 2015; Lemov, 2021; Liljedahl et al., 2021; Darling-Hammond et al., 2020; Schwartz et al., 2016) and for researchers to empirically quantify the impact of good teaching practices (Pianta and Hamre, 2009; Charalambous and Delaney, 2019; Blazar and Pollard, 2022; Jurenka et al., 2024). Thus, this work provides a bridge to research seeking to improve teaching quality by providing feedback to teachers on various instructional techniques (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Demszky et al., 2021; Suresh et al., 2022; Jacobs et al., 2022; Alic et al., 2022; Demszky and Liu, 2023; Demszky et al., 2024, 2023).

Automated Evaluation Several studies have investigated automated evaluation (Whitehill and LoCasale-Crouch, 2024; Wang and Demszky, 2023; Xu et al., 2024; Hardy, 2024). This study builds on these studies, and replicates the encoder model constructions described by Hardy.[2].

3 Methods

For each method, we also display the results at the item-level for better understanding of the learning process.

[2]A more complete description of the model architecture is in (Hardy, 2024)

3.1 Human Expert Rating Correlation

To find a generalized measure of correlation across items similar to CLASS and MQI, we use a multi-level partial Spearman's correlation with inference based on item-level random effects. This accounts for the hierarchical structure while providing a robust, rank-based measure of association that generalizes beyond the specific items sampled.

$$\rho_{\text{part}} = \text{Corr}(\tilde{R}_{ij}^{(1)}, \tilde{R}_{ij}^{(2)} \mid \mathbf{J}_{ij})$$

3.2 LLM Rating Stability via Variance Decomposition

We employed a generalizability theory framework (Brennan, 2001) to decompose variance in automated LLM scoring across six nested hierarchical levels: sentences (X) within utterances (U) within chapters (C) within lesson stages (S) within lessons (L) within teachers (T), denoted as $X : U : C : S : L : T$. This design enables quantification of context dependency as models evolve during training. The proportion of variance attributable to each level, $h \in \{T, L, S, C, U, X, e\}$, is:

$$\rho_h = \frac{\sigma_h^2}{\sigma_T^2 + \sigma_L^2 + \sigma_S^2 + \sigma_C^2 + \sigma_U^2 + \sigma_X^2 + \sigma_e^2}$$

.

3.3 External Validity via τ-Canonical Correlation Analysis

We employ canonical correlation (Hotelling, 1936) analysis (CCA) with a Kendall's tau (Kendall, 1938, 1945) kernel to measure alignment between teacher value-added measures (VAMs) and classroom instructional ratings. In this case, we need a metric that captures the directional alignment only, as differences in scales and ranks may not be meaningful.[3] We briefly reconstruct the τ kernel for creation of scatter matrices here to motivate it as a highly robust (Bishara and Hittner, 2017) measure of alignment between LLM ratings, **x**, and another metric, **y** and as having a straightforward alignment interpretation. We translate the following statement into the desired kernel: *"LLM X rates lesson A as better than lesson B: $[x_A > x_B]$. Does the order align with student learning results Y associated with each lesson, $[y_A > y_B]$?"* Thus, for any two lessons, indexed by i and j and with brackets as indicator functions, we construct this relationship for each:

$$x_{ij} = [j > i] - [j < i], \quad y_{ij} = [j > i] - [j < i]$$

We construct Gram matrices $\mathbf{K}_X$ and $\mathbf{K}_Y$ where $[\mathbf{K}_X]_{ij} = K_\tau(\mathbf{x}_i, \mathbf{x}_j)$ and analogously for K_Y, which are used to solve the eigenvalue problem required for canonical correlation. In our case, all matrices were positive semi-definite and no smoothing was needed. The results are robust to typical transformations for the calculation.[4]

4 Data and Experiment

4.1 Data

The original classroom lessons used in this study are from the National Center for Teacher Effectiveness (NCTE) Main Study (Kane et al., 2015),[5] which contains 3 years of data collection and observations of math instruction in approximately 50 schools and three hundred (4th and 5th grade) mathematics classrooms across 4 school districts in the United States. This rich dataset contains **multiple measures of teaching effectiveness**, including expert ratings of the lessons classrooms, and **multiple high-quality measures student learning gains**.

Classroom Lessons and Text Human raters watched videos classrooms, and the transcripts[6] of these same videos (Demszky and Hill, 2022) are used by LLMs for the same task, where the class discourse is equipartitioned across words to align the text with human labels in the absence of timestamps, following (Hardy, 2024).

Observation Instruments Our approach encompasses two complementary observation frameworks: a 12-item general teaching practices instrument and a 13-item mathematics-specific teaching assessment (Bacher-Hicks et al., 2017, 2019): the CLASS framework (Pianta et al., 2008) for general instructional practice and the content-specific Mathematical Quality of Instruction (MQI) (Hill

[3]For example, if two raters give Lesson A the same score of 7, but give Lesson B different ratings, 3 and 4, we would not have evidence to support the notion that LLMs (or even humans) apply an instructional rubric precisely enough for such differences in interval ranks to be practically meaningful when measuring alignment.

[4]We also compute the generalized eigenvalue problem using the methods put forward by (Yoon et al., 2020) with the results in the appendix.

[5]https://www.icpsr.umich.edu/web/ICPSR/studies/36095/datadocumentation

[6]https://github.com/ddemszky/classroom-transcript-analysis

et al., 2008). The 63 MQI raters and the 19 external CLASS raters attended biweekly calibration meetings to ensure standardization of scoring procedures. Both frameworks are composed of multiple items that represent distinct instructional dimensions to be evaluated (Hill et al., 2008; Hardy, 2024; Hill et al., 2012; Kane and Staiger, 2012). The MQI and CLASS also represent two types of task for LLMs–detection and summarization, respectively–a distinction that is also clearly illustrated by the distributions of scores coming from the human raters (see Figure 2). Human rating experts watched videos and provided ratings on all MQI and CLASS items at regular intervals throughout the class, resulting in 779,107 unique numeric ratings provided in 1,762 lessons[7] delivered by 317 teachers to more than 10,000 students in 53 schools.

Value-added measures (VAMs) are the current gold standard for estimating teacher effects on student achievement gains (Kane and Staiger, 2012; Bacher-Hicks et al., 2017, 2019). VAMs use prior student achievement data and other covariates to measure whether a student's end-of-year performance was above or below the student's expected performance on a standardized exam. The teacher-level VAM is an estimate of the combeined deviations from expected performance for their students, offering a rigorous aggregated estimate of a teacher's contribution to student learning gains over a school year. As far as we are aware, this is the first study to test LLMs using standardized and value-added measures of student learning. Rare for education datasets, the data of the present study have multiple VAM measurements, which we will use together as a random vector for canonical correlation (and stacked at the year level (Bacher-Hicks et al., 2019) for item-level comparisons).

4.2 Encoder Model Construction

We develop custom encoder architectures based on sentence-level embeddings to address four key research questions: (1) embedding efficiency in model training, (2) performance relative to human raters, (3) score variation across different temporal contexts, and (4) alignment with rigorous measures of student learning outcomes.

Why Sentence Embeddings? A More Interpretable Architecture. Analyzing lengthy classroom discourse poses a challenge for standard LLMs, whose subword tokens are often too granular and computationally intensive for such long-form text. To overcome this, we chose sentence-level embeddings as our foundation. Sentences are natural, interpretable semantic units, allowing our models to efficiently process entire lessons and directly map what a teacher says to established pedagogical frameworks. We evaluated five pretrained sentence-embedding models within a multi-task architecture that learns to score 25 distinct teaching dimensions simultaneously. The model's core weights are shared across all tasks, reflecting the pedagogical theory that effective teaching is a cohesive skill set. Only the final output layers are specialized for each dimension, allowing the model to learn both general and specific features of instructional quality.

Structuring Data for Meaningful Analysis. To capture the natural rhythm of a class, we organized transcripts into a three-phase structure (beginning, middle, end) common in elementary math. Each phase was then divided into chapters of a fixed duration (7.5 minutes for MQI, 15 for CLASS). We processed teacher speech into individual sentences, creating uniform inputs and enabling the model to update its assessment with each new utterance. We further augmented our dataset using a sliding-window technique to generate additional training samples from the discourse.

An Efficient and Robust Training Protocol. Given the challenge of collecting high-quality observation data, our training protocol prioritized efficiency and generalization. All models were trained for five epochs. We first determined the optimal optimizer by testing AdamW and Adamax; Adamax proved more stable for four of the five models under the strong regularization needed to prevent overfitting. The fifth model (GTE) performed better with AdamW. We retained all models to ensure a comprehensive evaluation. Additional training details follow the protocol in (Hardy, 2024).

5 Discussion

5.1 Some embeddings are better than others

Not all contextual embeddings capture the same semantic information. The RoBERTa, Unsupervised SimCSE, and E5 Multilingual models consistently outperformed other embedding models in the present study. When correlations are measured by

[7] Transcripts are available for 1,600 of the lessons (Demszky and Hill, 2022)

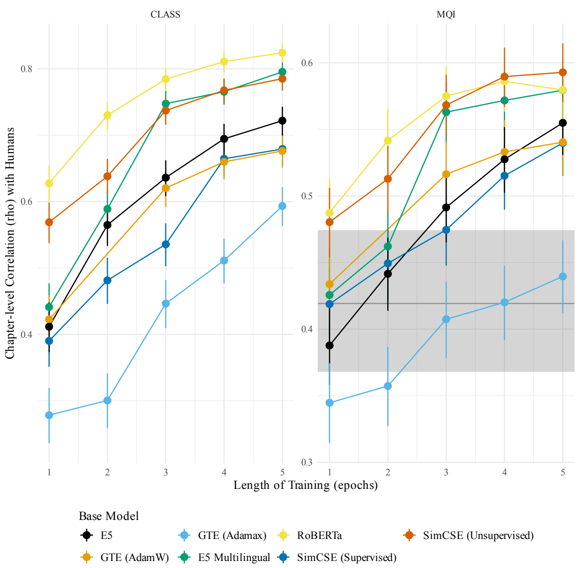


Figure 1: Correlation with human experts across training epochs. The MQI instrument had at least two human raters per lesson, and the mean and interquartile range of all 63 human MQI raters correlated across the other raters are represented by the gray line and shaded region in the figure.

aggregating to the lesson-level, all models' ratings would be in the top quartile of human raters. The item level performance at the chapter level can be investigated more deeply in Fig 6.

Implications Future work should explore fine-tuning SimCSE embeddings or similar self-supervised fine-tuning techniques on related classroom transcripts to investigate the extent to which providing domain-specific contrastive learning could further capture the most relevant semantic information from classroom discourse. Sentence-level embeddings also provide a pathway to interpretability with feature attribution via integrated gradients. Usefulness of feature attribution methods rely on how interpretable each input feature is. In the case of classroom instruction, a sentence spoken is a meaningful unit of discourse, whereas more common methods of creating features for LLMs typically rely on subword tokenization, producing a pixelated semanticity much harder for humans to interpret.

5.2 Mature models show score stability across longer time windows

Analysis across training epochs revealed a systematic shift in variance attribution: early-epoch models exhibited substantial utterance-level variation (ρ_U), while mature models demonstrated increased lesson-level variation (ρ_L) with reduced local context dependency. This finding challenges prevailing single-turn annotation paradigms in LLM evaluation, suggesting that as models improve, they capture broader pedagogical patterns rather than utterance-specific features. Consequently, evaluation frameworks must incorporate extended conversational contexts and hierarchical sampling strategies to accurately assess model performance in educational applications.

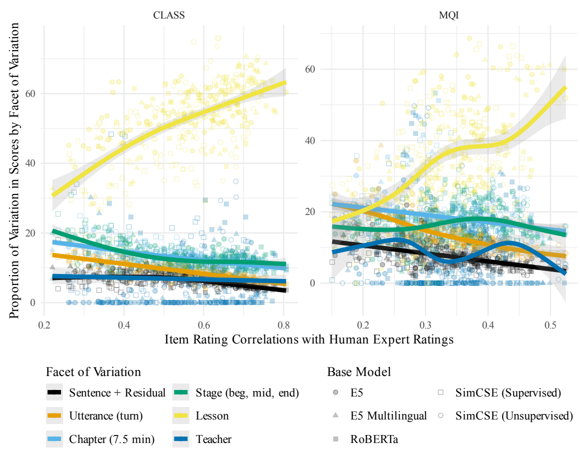


Figure 2: Proportion of variation explained as related to a model's alignment to human expert ratings.

We find that the variation in scores for models that are more aligned with human ratings, tends to be less sensitive to smaller changes in time. For the CLASS rubric in particular, more human-aligned models maintain more score stability across an entire lesson, suggesting that the scores are more representative of persistent differences in lessons.

5.3 External Validity

In order to assess for overfitting to this particular subset of human raters, we measured the alignment of the chapter-level multi-task scores against the value-added to learning for the students in the class. To date, there is no study of which we are aware that connect LLM measures of teaching to the actual value-added in student learning external validity.

We find that for is increasing alignment between VAM and classroom ratings as models mature and become more aligned with humans.

We demonstrate that while individual task performance generally improves with training, the correlation between model predictions and student achievement outcomes follows a non-monotonic trajectory. This phenomenon reveals important ten-

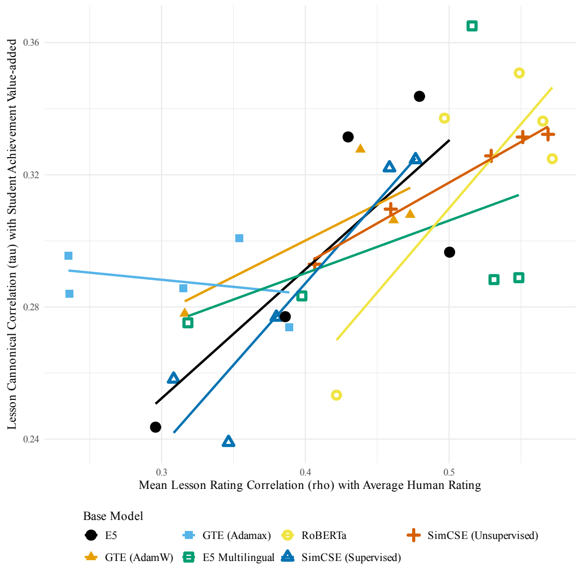


Figure 3: τ-canonical correlation between classroom observation ratings and value-added measures as a function of model alignment to human expert ratings.

sions in multitask learning: as models become more specialized at distinguishing between specific instructional skills, their ability to capture the general teaching effectiveness that correlates with student learning gains may paradoxically diminish.

To assess whether our models exhibit overfitting to this particular subset of human raters, we examined the relationship between chapter-level multitask scores and value-added measures (VAM) of student learning outcomes. We employed τ-canonical correlation analysis to quantify the strength of association between these two sets of variables while accounting for their multivariate nature.

This analysis addresses a critical gap in the literature: to our knowledge, no prior study has investigated the connection between LLM-derived measures of teaching quality and externally validated VAMs. Such validation is essential for establishing the practical utility of automated classroom assessment tools (Figure 3).

While the τ-canonical correlation between model predictions and student achievement outcomes generally improves with model sophistication, individual item performance can follow an inverted trajectory performance (Figure 9. This phenomenon illuminates an important tension inherent in multitask learning for educational assessment. As models become increasingly specialized at distinguishing between specific instructional skills that human raters prioritize, their capacity to capture the broader dimensions of teaching effectiveness that correlate with actual student learning gains may paradoxically diminish. This suggests that perfect alignment with human expert ratings may not constitute the optimal objective for developing classroom observation tools intended to predict student outcomes.

6 Conclusion

Rating classroom teaching quality remains a persistent challenge, with both human evaluators and large language models (LLMs) struggling to effectively utilize authentic classroom observation instruments. While general-purpose GPT models have shown limited promise for this task, we developed custom LLMs based on sentence embeddings to overcome the interpretability and scalability limitations of traditional subword tokenization approaches when processing lengthy classroom transcripts. We systematically evaluated five pretrained sentence embedding models using identical training regimes designed to maximize efficiency and minimize overfitting given the scarcity of authentic classroom data. We assessed their ability to capture pedagogically relevant information using established observation frameworks. Our results demonstrate that three embedding models achieved human-level performance, with correlations exceeding 0.65 for CLASS and surpassing human averages for MQI. More mature models increasingly attribute variation to differences at lesson-level rather than utterance-specific features. This finding challenges prevailing single-turn evaluation paradigms in LLM assessment and development, suggesting that improved models capture broader pedagogical patterns in long-context classroom dynamics. Validity analysis using teacher value-added measures revealed that while models achieving better human alignment also showed stronger alignment with learning outcomes in aggregate, this relationship did not hold at the item level. These results indicate that although models learn pedagogically meaningful features, evidence for generalization remains limited, highlighting important directions for future development of automated teaching quality assessment systems.

Limitations

The findings of this study should be considered in light of several limitations related to the data, models, and readiness for practical application. We position this work as a proof of concept, and the following factors must be addressed before these methods can be considered for real-world implementation.

Scope of Data and Generalizability The primary limitation of this study is the specificity of the dataset, which consists of transcripts from fourth and fifth-grade mathematics classrooms in the United States. This narrow scope means our models lack proven generalizability to other grade levels, subject areas (e.g., literacy, science), or international school systems. While the underlying methods may hold broader potential, our specific findings are bound to the context of U.S. elementary mathematics education. Expanding the applicability of these models is contingent upon the availability of more varied public data.

Task and Model Specificity Our approach is limited by both the evaluation task and the model architecture. We focused on a subset of rating items from the MQI rubric, which may not fully represent the complexity of the universal task of instructional rating. Additionally, the inherent imperfections of observational rubrics, even for human experts, are a constraint on the ground truth data. Furthermore, our encoder models were custom-built for this task. While this specialized design allows a single model to score 25 distinct measures, it is not designed to generalize to new domains or contexts without substantial changes to its architecture or the introduction of new training data.

Considerations for Practical Application and Deployment Despite achieving high performance on several metrics, the models in their current state are not ready for high-stakes deployment. Substantial research and validation are necessary to ensure their reliability and to understand potential failure modes. Even when used with a human-in-the-loop, more work is needed to align the models' capabilities with the potential assumptions of end-users. Crucially, this study should not be interpreted as an endorsement for using general-purpose "GPT-style" large language models for similar evaluative tasks. The challenges inherent in this domain require specialized, carefully validated solutions rather than the application of general-purpose technologies.

References

Elena Aguilar. 2013. Developing a Work Plan: How Do I Determine What to Do? In *The art of coaching: effective strategies for school transformation*, pages 119–144. Jossey-Bass, A Wiley Brand, San Francisco.

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. *arXiv preprint*. ArXiv:2208.04715 [cs].

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2017. An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys.

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73:101919.

Paul Bambrick-Santoyo. 2016. *Get better faster: a 90-day plan for coaching new teachers*. Jossey-Bass, A Wiley Brand, San Francisco, CA.

Paul Bambrick-Santoyo. 2018. *Leverage leadership 2.0: a practical guide to building exceptional schools*. Jossey-Bass, San Francisco, CA.

Anthony J. Bishara and James B. Hittner. 2017. Confidence intervals for correlations when data are not normal. *Behavior Research Methods*, 49(1):294–309.

David Blazar. 2018. Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students. *Education Finance and Policy*, 13(3):281–309.

David Blazar and Cynthia Pollard. 2022. Challenges and Tradeoffs of "Good" Teaching: The Pursuit of Multiple Educational Outcomes. Technical report, Annenberg Institute at Brown University. Publication Title: EdWorkingPapers.com.

Robert L. Brennan. 2001. *Generalizability Theory*. Springer, New York, NY.

Jodi M. Casabianca, Daniel F. McCaffrey, Drew H. Gitomer, Courtney A. Bell, Bridget K. Hamre, and Robert C. Pianta. 2013. Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, 73(5):757–783. Publisher: SAGE Publications Inc.

Charalambos Y. Charalambous and Seán Delaney. 2019. 13 Mathematics Teaching Practices and Practice-Based Pedagogies. Brill. Section: International Handbook of Mathematics Teacher Education: Volume 1.

Linda Darling-Hammond. 2014. What Can PISA Tell Us about U.S. Education Policy? *New England Journal of Public Policy*, 26(1).

Linda Darling-Hammond, Lisa Flook, Channa Cook-Harvey, Brigid Barron, and David Osher. 2020. Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2):97–140. Publisher: Routledge _eprint: https://doi.org/10.1080/10888691.2018.1537791.

Dorottya Demszky and Heather Hill. 2022. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. Publisher: arXiv Version Number: 1.

Dorottya Demszky and Jing Liu. 2023. M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, pages 59–69, New York, NY, USA. Association for Computing Machinery. Event-place: Copenhagen, Denmark.

Dorottya Demszky, Jing Liu, Heather C. Hill, Shyamoli Sanghi, and Ariel Chung. 2023. Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms. Technical report, Annenberg Institute at Brown University. Publication Title: EdWorkingPapers.com.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Dorottya Demszky, Rose Wang, Sean Geraghty, and Carol Yu. 2024. Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, pages 632–644, New York, NY, USA. Association for Computing Machinery.

Patrick J. Donnelly, Nathaniel Blanchard, Andrew M. Olney, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 218–227, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint*. ArXiv:2104.08821 [cs].

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, and 15 others. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint*. ArXiv:2209.14375.

Jason Grissom, Susanna Loeb, and Benjamin Master. 2013. Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals. *Educational Researcher*, 42(8)(42(8)):433.

Zaretta Hammond. 2015. *Culturally responsive teaching and the brain: promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin, a SAGE company, Thousand Oaks, California. OCLC: ocn889185083.

Michael Hardy. 2024. "All that Glitters": Approaches to Evaluations with Unreliable Model and Human Annotations. *arXiv preprint*. ArXiv:2411.15634 [cs].

Michael Hardy. 2025. "All that Glitters": Techniques for Evaluations with Unreliable Model and Human Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2250–2278, Albuquerque, New Mexico. Association for Computational Linguistics.

Heather C. Hill, Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, 26(4):430–511. Publisher: Taylor & Francis, Ltd.

Heather C. Hill, Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, 41(2):56–64. Publisher: American Educational Research Association.

Andrew D. Ho and Thomas J. Kane. 2013. The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540957.

Harold Hotelling. 1936. Relations Between Two Sets of Variates*. *Biometrika*, 28(3-4):321–377.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, and 51 others. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach.

Thomas Kane, Heather Hill, and Douglas Staiger. 2015. National Center for Teacher Effectiveness Main Study: Version 4.

Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540959.

Thomas J. Kane and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540960.

Sean Kelly, Andrew M. Olney, Patrick Donnelly, Martin Nystrand, and Sidney K. D'Mello. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*, 47(7):451–464. Publisher: American Educational Research Association.

Maurice George Kendall. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Maurice George Kendall. 1945. THE TREATMENT OF TIES IN RANKING PROBLEMS. *Biometrika*, 33(3):239–251.

Doug Lemov. 2021. *Teach like a champion 3.0: 63 techniques that put students on the path to college*, third edition edition. Jossey-Bass, a Wiley imprint, Hoboken, NJ.

Doug Lemov and Norman Atkins. 2015. *Teach like a champion 2.0: 62 techniques that put students on the path to college*, second edition edition. Jossey-Bass, San Francisco, CA.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv preprint*. ArXiv:2308.03281 [cs].

Peter Liljedahl, Tracy Johnston Zager, and Laura Wheeler. 2021. *Building thinking classrooms in mathematics: 14 teaching practices for enhancing learning: Grades K-12*. Corwin Mathematics. Corwin, Thousand Oaks, California London New Delhi Singapore.

Jing Liu and Julie Cohen. 2021. Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, 43(4):587–614. Publisher: American Educational Research Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Robert C. Pianta, Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. 2008. Classroom Effects on Children's Achievement Trajectories in Elementary School. *American Educational Research Journal*, 45(2):365–397. Publisher: American Educational Research Association.

Robert C. Pianta and Bridget K. Hamre. 2009. Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, 38(2):109–119. Publisher: American Educational Research Association.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Borhan Samei, Andrew M. Olney, Sean Kelly, Martin Nystrand, Sidney D'Mello, Nathan Blanchard, Xiaoyi Sun, Marcy Glaus, and Art Graesser. 2014. Domain Independent Assessment of Dialogic Properties of Classroom Discourse. Technical report. Publication Title: Grantee Submission ERIC Number: ED566380.

Jon Saphier, Mary Ann Haley-Speca, and Robert Gower. 2008. *The skillful teacher: building your teaching skills*, 6th ed edition. Research for Better Teaching, Acton, Mass.

Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*, first edition edition. Norton books in education. W.W. Norton & Company, New York.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. *arXiv preprint*. ArXiv:2204.09652 [cs].

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. *arXiv preprint*. ArXiv:2306.06941.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and

Furu Wei. 2024a. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint*. ArXiv:2212.03533 [cs].

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*. ArXiv:2402.05672 [cs].

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated Evaluation of Classroom Instructional Support with LLMs and BoWs: Connecting Global Predictions to Specific Feedback. *arXiv preprint*. ArXiv:2310.01132 [cs].

Grover J. Whitehurst, Matthew M. Chingos, and Katharine M. Lindquist. 2014. Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts. Technical report, Brookings Institution. Publication Title: Brookings Institution ERIC Number: ED553815.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education. *arXiv preprint*. ArXiv:2404.02444 [cs].

Grace Yoon, Raymond J Carroll, and Irina Gaynanova. 2020. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3):609–625.

A Item Information

B Figures

B.1 Human Expert Score Distributions

B.2 Test Set Distributions

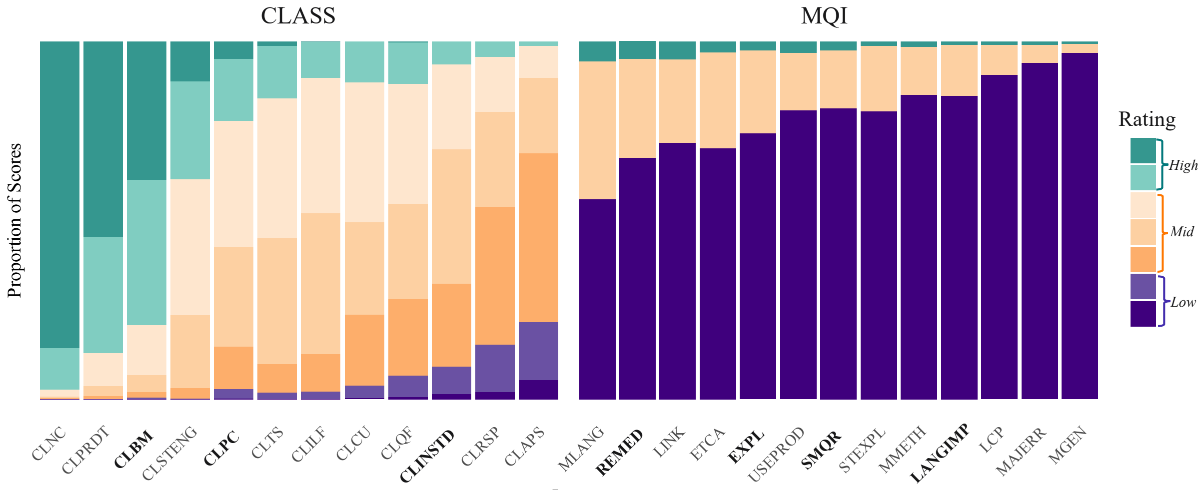


Figure 4: **Human Expert Score Distributions**. These are the score distributions from human experts. The distinct rating patterns highlight the underlying qualitative differences in the constructs being rated. Previous studies have focused on a limited range of items (bolded, (Wang and Demszky, 2023))

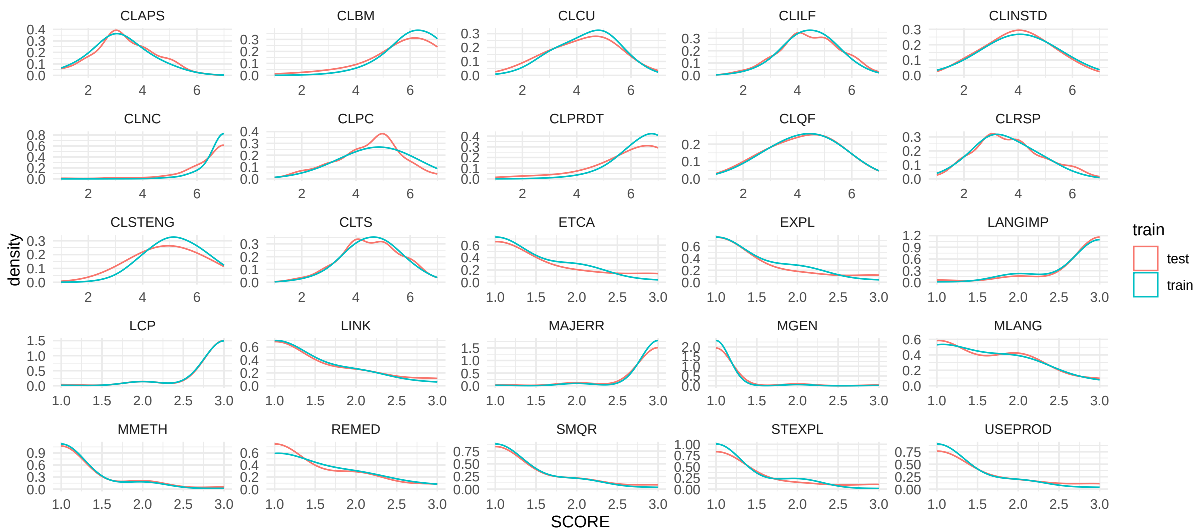


Figure 5: **Held-out Test Set Distributions**. These are comparative score distributions from human experts for the items in the held-out test set and the remaining sample. No differences were statistically significant.

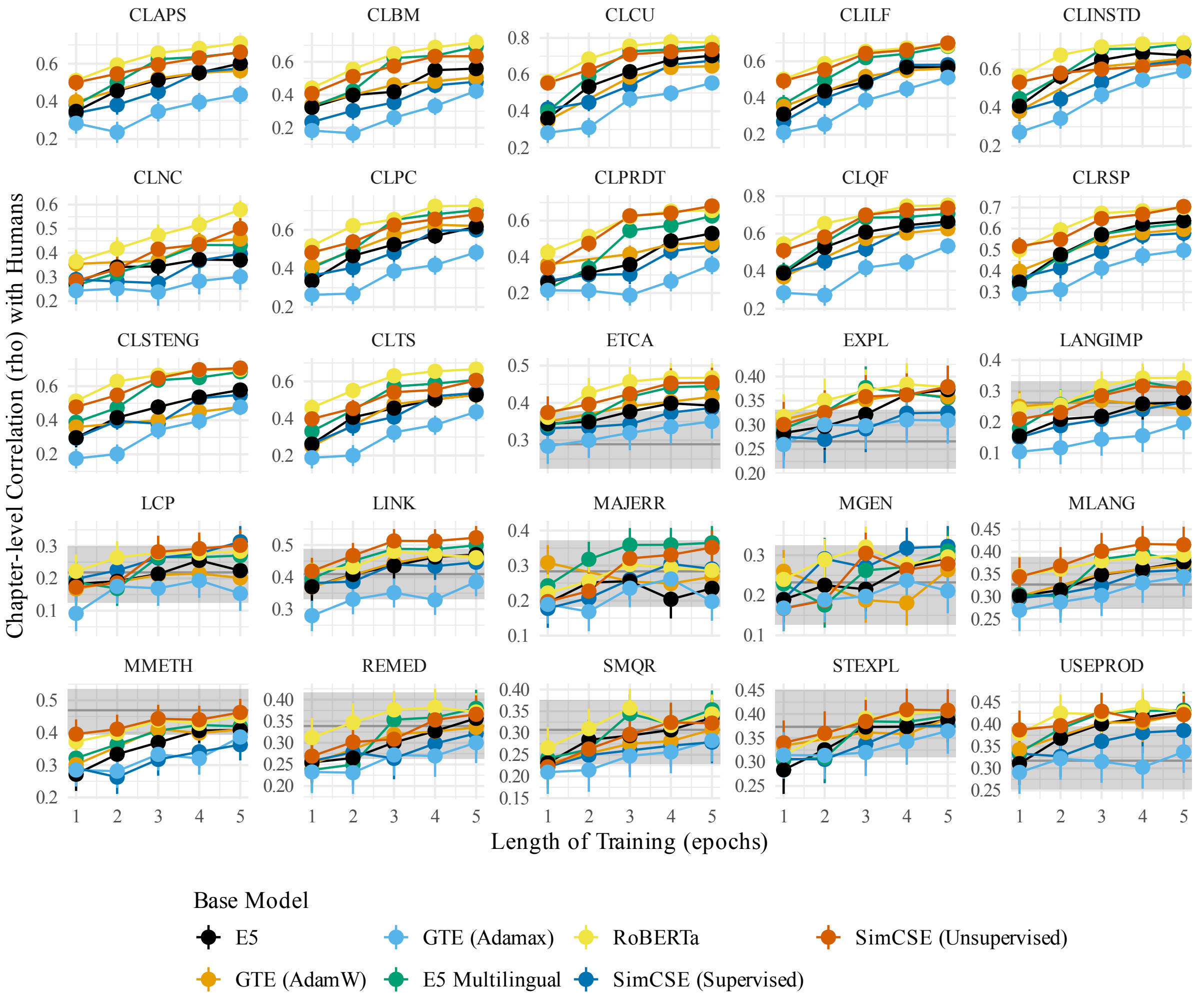


Figure 6: Chapter-level by Item-level Correlation with human experts across training epochs

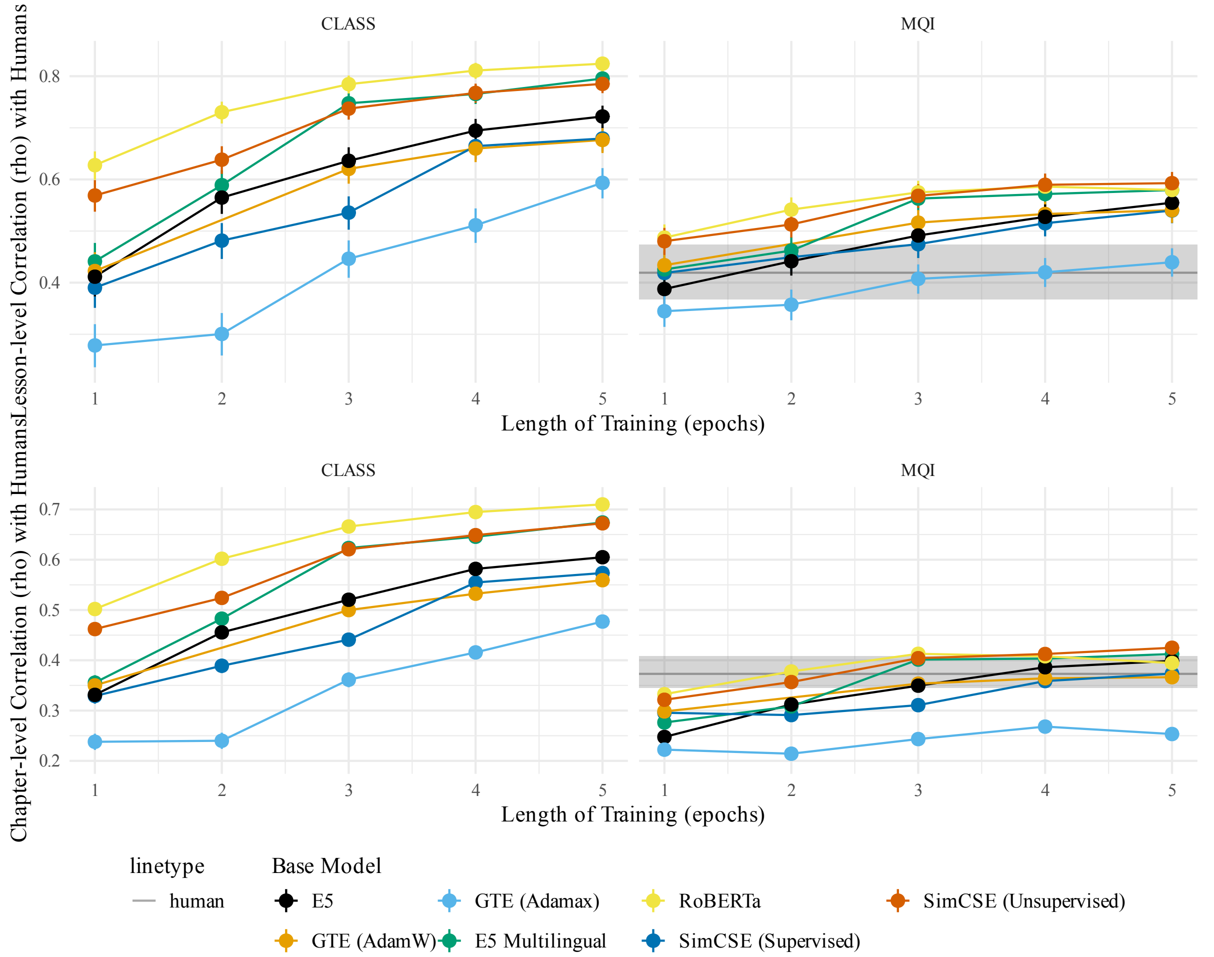


Figure 7: The MQI instrument had at least two human raters per lesson, and the mean and interquartile range of all 63 human MQI raters correlated across the other raters are represented by the gray line and shaded region in the figure. When looking at only the specific chapter from the lesson, however, models only perform in the top half of human raters. This discrepancy may be an artifact of the windowing structure used in the training. The multilevel framework decomposes the correlation into within- and between-cluster components: $\rho_{\text{partial}} = \text{Corr}(\text{rank}(\tilde{R}_{ij}^{(1)}), \text{rank}(\tilde{R}_{ij}^{(2)}) \mid \mathbf{X}_{ij})$ where $\tilde{R}_{ij}^{(k)}$ represents residualized ratings after partialling out covariates $\mathbf{X}_{ij}$ from the multilevel model: $R_{ij}^{(k)} = \boldsymbol{\beta}^{(k)}\mathbf{X}_{ij} + u_j^{(k)} + \epsilon_{ij}^{(k)}$ with random intercepts $u_j^{(k)} \sim N(0, \tau^2)$ for items and residuals $\epsilon_{ij}^{(k)} \sim N(0, \sigma^2)$. The multilevel partial Spearman's correlation accounts for the hierarchical structure while providing a robust, rank-based measure of association that generalizes beyond the specific items sampled, with inference based on the random effects distribution of items.

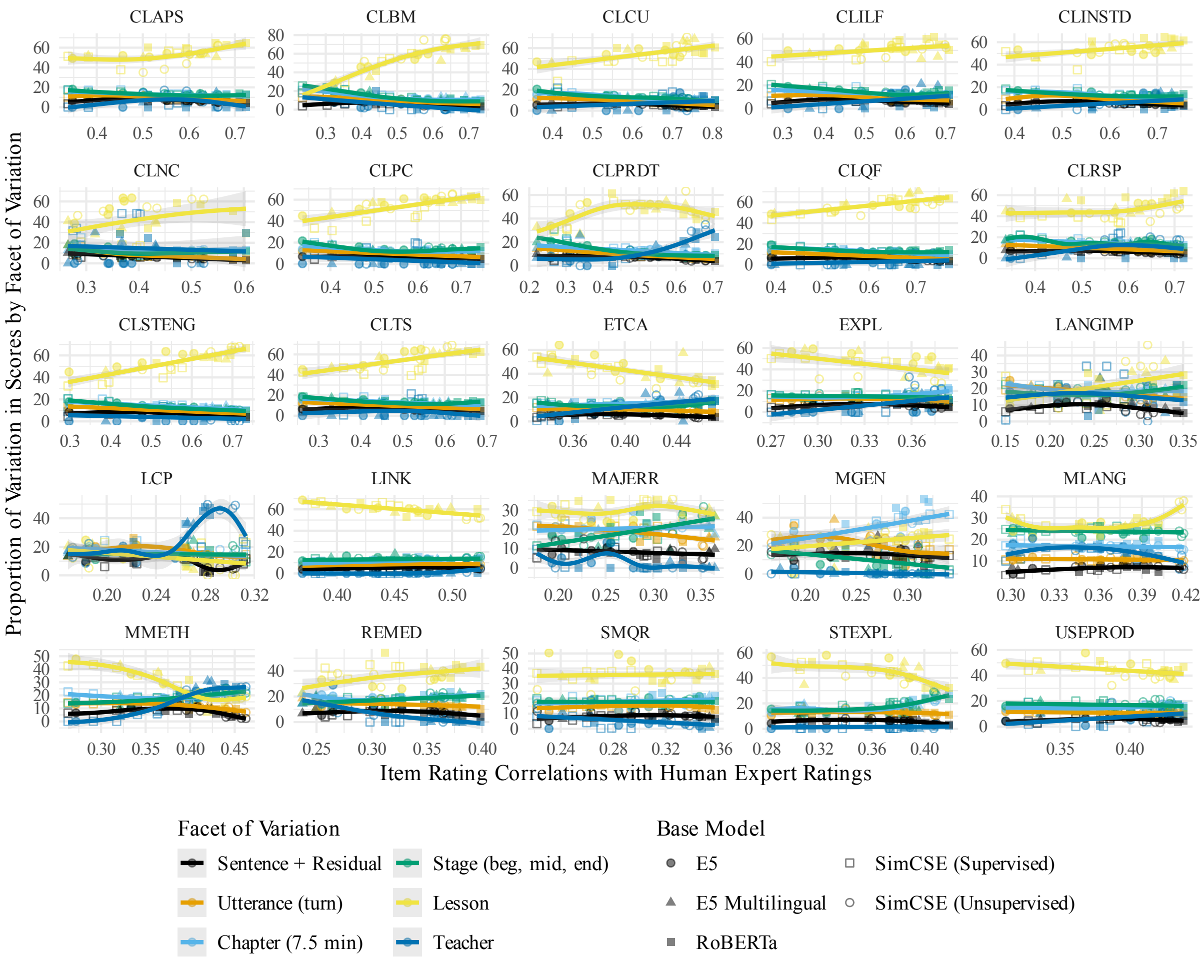


Figure 8: Proportion of variation explained as related to a model's alignment to human expert ratings disaggregated by item using generalizability theory. Random effects were calculated using the `lme4` package in `R`: `SCORE (1|NCTETID) + (1|OBSID) + (1|OBS_CHAPS) + (1|OBS_CHAP) + (1|OBS_CHAP_idx)`

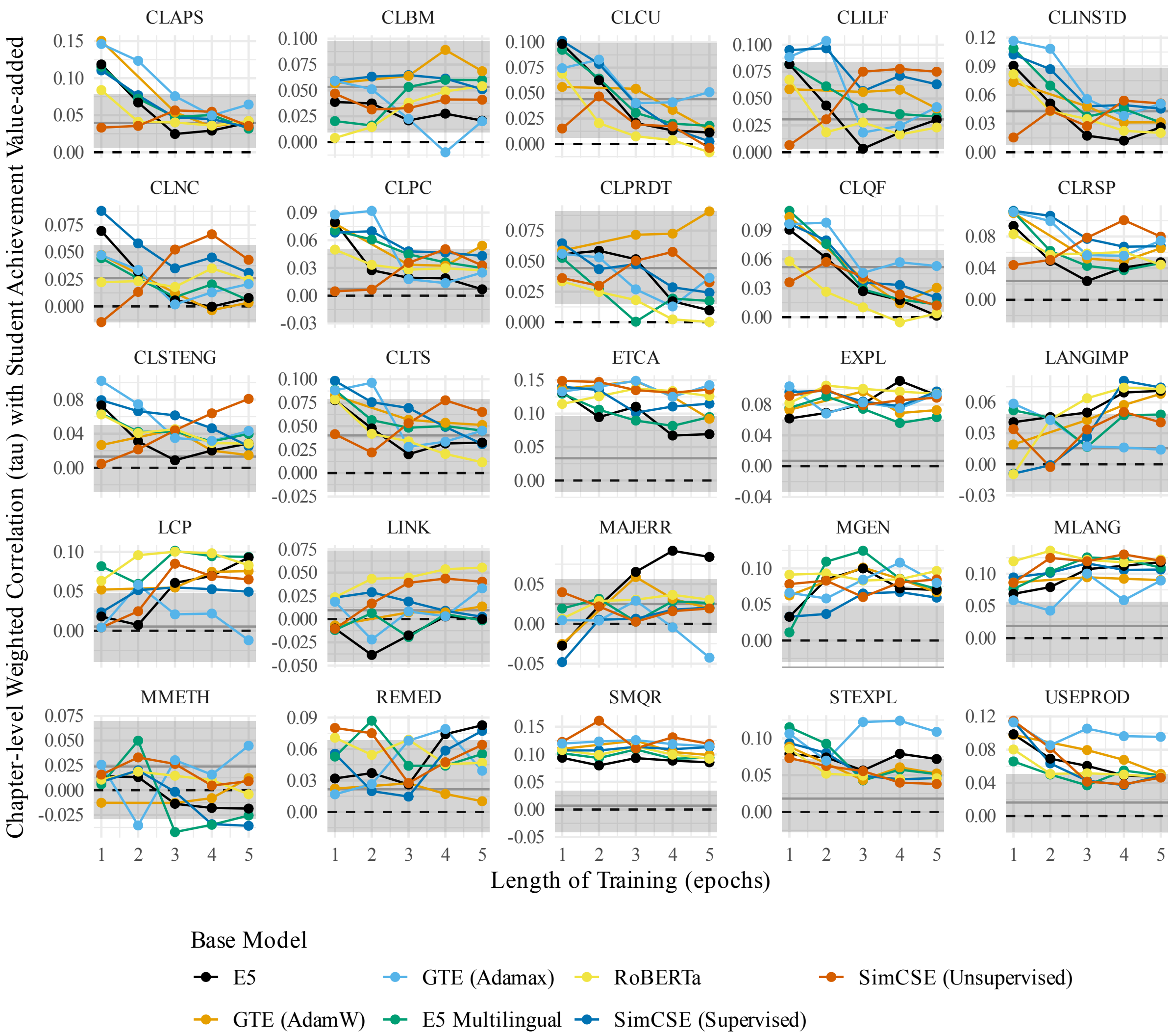


Figure 9: Kendall tau correlations on the stacked VAM outcomes disaggregated at the item level as a function of training epoch. The mean and interquartile range of all human raters evaluated in the same manner are represented by the gray line and shaded region in the figure. Weights are used to account for number of observations per unit.

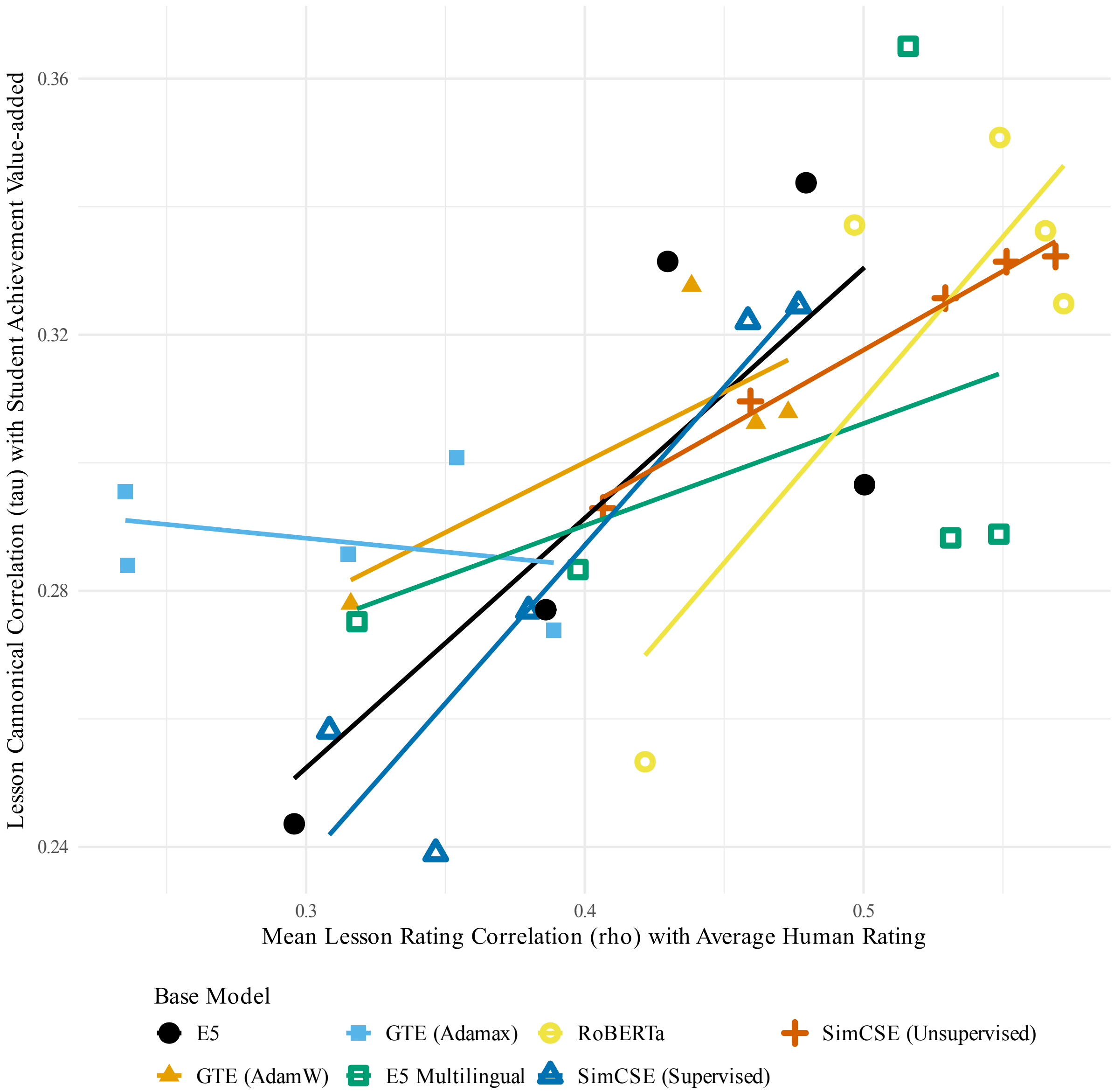


Figure 10: Lesson-level canonical correlations as a function of correlation with human ratings.

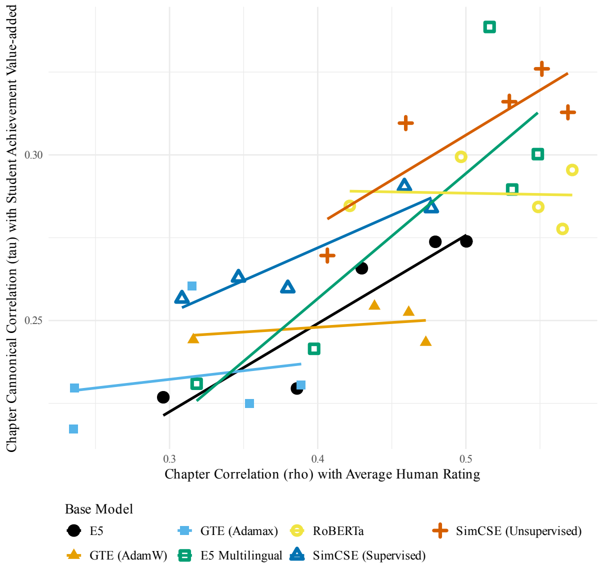


Figure 11: Chapter-level canonical correlations using methods from Yoon et al. as a function of correlation with human ratings.

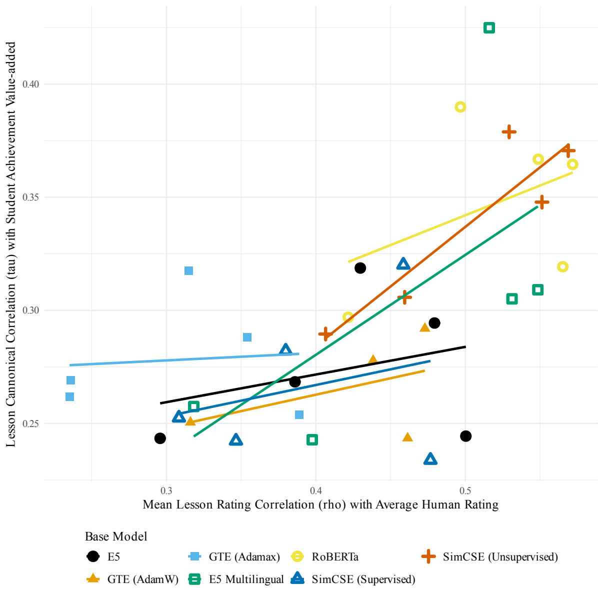


Figure 12: Lesson-level canonical correlations using methods from Yoon et al. as a function of correlation with human ratings.

Abbreviation	Item	Item Description
MQI Instrument		
ETCA	*Enacted Task Cognitive Activation*	Task cognitive demand, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns.
EXPL	*Teacher Explanations*	Teacher explanations that give meaning to ideas, procedures, steps, or solution methods.
LANGIMP†	*Imprecision in Language or Notation*	Imprecision in language or notation, with regard to mathematical symbols and technical or general mathematical language.
LCP†	*Lack of Clarity in Presentation of Mathematical Content*	Lack of clarity in teachers' launching of tasks or presentation of the content.
LINK	*Linking and Connections*	Linking and connections of mathematical representations, ideas, and procedures.
MAJERR†	*Major Mathematical Errors*	Major mathematical errors, such as solving problems incorrectly, defining terms incorrectly, forgetting a key condition in a definition, equating two non-identical mathematical terms.
MGEN	*Developing Mathematical Generalizations*	Developing generalizations based on multiple examples.
MLANG	*Mathematical Language*	Mathematical language is dense and precise and is used fluently and consistently.
MMETH	*Multiple Procedures or Solution Methods*	Multiple procedures or solution methods for a single problem.
REMED	*Remediation of Student Errors and Difficulties*	Remediation of student errors and difficulties addressed in a substantive manner.
SMQR	*Student Mathematical Questioning and Reasoning*	Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims.
STEXPL	*Students Provide Explanations*	Student explanations that give meaning to ideas, procedures, steps, or solution methods.
USEPROD	*Responding to Student Mathematical Productions*	Responding to student mathematical productions in instruction, such as appropriately identifying mathematical insight in specific student questions, comments, or work; building instruction on student ideas or methods.
CLASS Instrument		
CLPC	*Classroom Positive Climate*	
CLNC†	*Classroom Negative Climate*	
CLTS	*Teacher Sensitivity*	
CLRSP	*Regard for Student Perspective*	
CLBM	*Behavior Management*	
CLPRDT	*Productivity*	
CLILF	*Instructional Learning Formats*	
CLCU	*Content Understanding*	
CLAPS	*Applied Problem Solving*	
CLQF	*Quality of Feedback*	
CLINSTD	*Instructional Dialogue*	
CLSTENG	*Student Engagement*	

Table 1: CLASS and MQI item descriptions and corresponding abbreviations. †denotes items that are reverse coded due to being negatively worded with respect to the construct of teacher ability. Bolded items are those evaluated by the **GPT** family of raters and reported by Wang and Demszky. Each member of the Human and Encoder families of raters evaluated all 25 items.

Simulating Rating Scale Responses with LLMs for Early-Stage Item Evaluation

Onur Demirkaya, Hsin-Ro Wei and **Evelyn Johnson**
Riverside Insights
{onur.demirkaya, hsin-ro.wei, ejohnson}@riversideinsights.com

Abstract

This study explores the use of large language models to simulate human responses to Likert-scale items. A DeBERTa-base model fine-tuned with item text and examinee ability emulates a graded response model (GRM). High alignment with GRM probabilities and reasonable threshold recovery support LLMs as scalable tools for early-stage item evaluation.

1 Introduction

Field-testing is essential for developing any assessments as it serves to evaluate the statistical quality of newly developed items before the operational use. However, it remains one of the most resource-intensive and time-consuming stage in developing a test. Traditional approaches require human examinees to try out items, posing challenges related to sample availability, test security, item exposure, and scheduling . (AlKhuzaey et al., 2023; Hsu et al., 2018; Morizot et al., 2007). These challenges are growing as item banks must scale rapidly to support contemporary tests such as adaptive testing, multilingual formats, and artificial intelligent (AI)-generated content.

In response, early attempts predicted item difficulty using text-based features like syntax, semantics, word counts, embeddings, and readability indices (AlKhuzaey et al., 2023; Benedetto et al., 2023). Others used natural language processing (NLP) techniques to estimate item difficulty or discrimination (Benedetto et al., 2021; Zhou & Tao, 2020), but their accuracy remains limited. Importantly, these models often overlook distractors and fail to capture the full complexity of the human test-taking process (Benedetto et al., 2021).

More recent work has been exploring whether large language models (LLMs) can partially or fully simulate examinee responses to streamline item evaluation without sacrificing psychometric validity. For example, Lu and Wang's (2024) "generative students" prompt GPT-4 to mimic 45 learner profiles with different knowledge states, achieving a moderate correlation with undergraduate item scores ($r \approx .72$) but relying on expert-defined misconceptions and a tiny, single-topic item set. Liu, Bhandari, and Pardos (2024) go broader by blending six distinct LLMs into a 50-member ensemble, reproducing human Rasch difficulties on a small college-algebra pool ($r = .93$) yet still showing a compressed ability spread. Collectively, these studies confirm the promise of LLM-based field-testing while exposing the need for scalable methods that reduce expert overhead, widen domain coverage, and capture the full spectrum of item functioning.

Maeda (2025) moves furthest toward full AI substitution by fine-tuning 61 DeBERTa-v3 models, each tied to a specific latent ability, and embedding a two parameter logistic (2-PL) loss to predict option-level probabilities. Across 466 English-grammar items, the system matched human proportion-correct with $r = .82$ and zero mean bias, delivering plausible discrimination and distractor statistics and suggesting substantial cost and security gains. Yet achieving this required training 61 large models exposing heavy computational demand, and several extreme items still failed to calibrate accurately.

Building on Maeda's (2025) foundation, our approach leverages a single LLM that takes both item features such item and domain texts and a student's latent ability (θ) as input to predict selection probabilities of item's options, effectively emulating the graded response model

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 385–392
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

(GRM; Samejima, 1969). Rather than training separate models for each ability level, we condition predictions on continuous θ values and allow the model to learn item parameters implicitly from item features. This study aims to investigate if the proposed architecture enables realistic response simulation for field test Likert-scale items, supports scalable data generation, and reduces computational overhead while preserving psychometric structure, positioning it as a cost-efficient alternative for pretesting in large-scale assessments.

2 Background

2.1 Transformer Language Models

Transformer-based language models such as BERT (Devlin et al., 2018) are pre-trained on large text corpora and can be fine-tuned for various NLP tasks, including classification, summarization, and question answering. These models tokenize input text, convert it into embeddings, and process the sequence through multiple encoder layers to capture rich contextual information.

In this study, we used the DeBERTa-base model (He et al., 2021), an advanced variant of BERT and RoBERTa. DeBERTa improves representation learning by separately encoding the content and relative position of each token and computing distinct attention weights for both. This structure enhances the model's ability to capture nuanced word relationships, making it well-suited for complex language understanding tasks.

2.2 Graded Response Model

The Graded Response Model (GRM; Samejima, 1969) is a widely used item response theory (IRT) model for analyzing ordinal polytomous item responses, such as rating scales or multi-point rubrics in educational assessments. GRM models the probability that an examinee's latent ability θ_j exceeds a series of ordered thresholds for item *i*. Each item has a discrimination parameter a_i and a set of threshold (difficulty) parameters b_{ik}, one for each score category boundary. The probability of responding in category k is defined by the difference between cumulative logistic functions across adjacent thresholds:

$$P(X_{ij} = k \,|\theta_j) = P^*_{ik}(\theta_j) - P^*_{i(k+1)}(\theta_j) \qquad (1)$$

where

$$P^*_{ik}(\theta_j) = \frac{1}{1 + \exp\,[-a_i(\theta_j - b_{ik})]} \qquad (2)$$

The GRM assumes monotonicity, unidimensionality, and local independence, and it enables estimation of both person ability and item parameters on a common scale. It is especially well-suited for assessments where responses reflect degrees of correctness or agreement rather than binary outcomes. For detailed discussion, see Samejima (1969) or Baker & Kim (2004).

2.3 AI-Based Field-Test Data Generation Pipeline

This approach builds on Maeda's (2025) architecture but ease its computational cost by training a single generalized model instead of separate models per ability level and eliminating the sampling process, supporting scalable and flexible field test data generation for Likert-scale items. The model is trained on operational items with known GRM-based probabilities, using both item features and latent ability ($\boldsymbol{\theta}$) as inputs. Once trained, the model generalizes to predict probabilities for new items by conditioning on the item's text representation and examinee ability. Simulated responses are generated by sampling from the predicted probabilities, enabling psychometric analyses such as pre-calibration and item screening without requiring real student administrations. The AI-based field test data simulation pipeline is provided in Figure A1.

Process Item Features Data

We used items of the Devereux Student Strengths Assessment (DESSA). It is a standardized, strength-based behavioral rating scale developed to assess social-emotional competencies in children and adolescents (LeBuffe, et. al., 2009). The DESSA has eight empirically derived domains: self-awareness, social awareness, , self-management, goal-directed behavior, relationship skills, personal responsibility, decision making, and optimistic thinking (Shapiro & LeBuffe, 2004). Items are rated on a 5-point Likert scale (from "never" to "almost always") and yield standard scores with T-score interpretation.

Taking advantage of Likert-scale with same options across items, we only included item stem and its domain information as a text input enabling the model items as unified constructs and deeper theta-text interaction. Item stems were paired with the item's domain label (e.g., "Domain: Self-Awareness") to provide semantic context. The resulting domain-qualified text was used as input features in training the model to predict graded response probability distribution overall possible options (see Figure A2).

Calculate Conditional IRT Probabilities

To generate model training targets, we used the GRM (Samejima, 1969) to compute the conditional probability of each ordinal response category. A total of 1,000 examinee ability levels ($\boldsymbol{\theta}$) were sampled from a standard normal distribution $N(\mu = 0, \sigma^2 = 1)$, which closely approximates the ability distribution of the target population, $\mu = -0.002, \sigma^2 = 0.98$.

For each item-person pair, we used the item's GRM parameters, discrimination parameter a_i and threshold parameters b_{ik}, to compute the probability of a response in category k as given in equation 1. This yields a vector of conditional probabilities across all response categories for each item-θ pair. These probability vectors were used as target labels in training the model to emulate the GRM response function.

Fine-tune transformers with item features and theta

The fine-tuning pipeline employs DeBERTa-base as the text encoder, leveraging its disentangled-attention backbone to yield a 768-dim CLS embedding that captures both content and relative-position information efficiently (He et al., 2021). To make the single latent-ability estimate θ competitive in that high-dimensional space, the model feeds θ through a dedicated "ThetaEncoder" sub-network before concatenation. This process let the network learn either a simple or a richly nonlinear transformation as needed. It is first passed through three hidden layers (sizes 64 → 128→ 256 with GELU activations, LayerNorm, and dropout followed by tanh), producing a θ-embedding that shares scale and distributional properties with the transformer hidden states. This vector is concatenated with the original CLS embedding, giving a 1536-dim joint feature on which a dropout-regularized linear head (1536 → 5) predicts raw logits that are converted to predictive probabilities via soft-max before any loss is computed. Optimization uses cross-entropy with soft targets**:** for every training example the target distribution is the five-category probability vector produced by Samejima's graded-response IRT model, and the loss

$$CEL = -\sum_k P_k \log\hat{P}_k \tag{3}$$

encourages the network to reproduce the entire curve rather than just the arg-max label. Because θ now enters through hundreds of weights instead of one and the loss supplies dense probabilistic feedback, the model learns item-specific category curves that vary smoothly with ability.

Generate Item Responses

Once the fine-tuned LLM model has produced a five-element probability vector $\hat{P}_{ijk} = (\hat{P}_{ij0}, \dots, \hat{P}_{ij4})$ for examinee j on item i, to mimic human-like stochasticity probability based sampling is used to generate a concrete response rather than deterministic arg-max which can distort the latent-response surface and inflate slope estimates later in calibration. A complete response matrix is produced in this way for both training and field-test items for further psychometric analysis.

3 Methods

This study uses DESSA items to simulate a scenario where some set of previously calibrated items are available for training, while a smaller set of new items, represented only by their text, requires field-testing. Item parameters derived from prior field-testing are treated as true item parameters for both training and evaluation.

The dataset included 50 DESSA items, a standardized, strength-based behavioral rating scale developed to assess social-emotional competencies in children and adolescents. The instrument encompasses eight empirically derived domains: self-awareness, social awareness, self-management, goal-directed behavior, relationship skills, personal responsibility, decision making, and optimistic thinking. Each item is rated on a 5-point Likert scale (from "never" to "almost always"). All items had been previously calibrated using the GRM based on responses from a nationally representative sample of 1,350 middle school students. Among the 1,350 respondents, 2.5 %, 8.4 %, 26.3 %, 35.0 %, and 27.7 % endorsed categories 0, 1, 2, 3, and 4, respectively. Overall, nearly two-thirds of the calibration sample endorse the item at a high level, while only about one in ten fall at the negative end of the scale.

To simulate the field-testing context, the items were randomly divided into 85% training items ($n = 38$) and 15% field-test items ($n = 12$), with the constraint that the field-test subset included at least one item from each SEL domain. The training items served as calibrated, operational items, used to fine-tune the language model and anchor the score scale during calibration. The field-test items, excluded from model training, represented new, uncalibrated items used to evaluate model generalization and calibration accuracy.

The DeBERTa-base model (He et al., 2021) was fine-tuned using the item features (domain label and item text) from the 38 training items, along with 1,000 latent ability values (θ) sampled from a standard normal distribution, $N(0,1)$, which reflects the target population's ability distribution. The AdamW optimizer (Loshchilov & Hutter, 2017) was used to minimize the CEL between the GRM-derived target probabilities and the model's predictions (James et al., 2023). Fine-tuning was conducted using the PyTorch library (Paszke et al., 2019) on a Google Colab Pro instance equipped with a NVIDIA A100 Tensor Core 40GB GPU. The model was trained for 15 epochs with a batch size of 16 per device, a learning rate of 2×10^{-5}, and a weight decay of 0.01. Item response data were generated based on $\hat{P}_{ijk}$ for all training and field-test items.

To assess the psychometric quality of the generated data, field-test items were calibrated using the GRM. (Samejima, 1969). The mean and variance of the latent ability (θ) were freely estimated, while the parameters of the training items were anchored by fixing them to their known discrimination (a_i) and threshold (b_{ik}) values, ensuring that field-test items were placed on the same scale. Calibration was conducted using the mirt package in R (Chalmers, 2012). Item parameters previously obtained from field-testing with real human examinees were treated as true values. Estimates derived from the model were evaluated against these true values using mean signed bias, mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficients (r) for each parameter.

4 Results

Table A1 shows that the average item parameters are generally comparable between the training (38 items) and testing sets (12 items). Standard deviations are also consistent across sets, with slightly more variation in the testing items' thresholds, likely due to fewer items. Overall, these similarities suggest that the item parameter distributions are almost balanced across the training and testing subsets.

Figure 1 demonstrates that the model's predicted probabilities track the true category probabilities almost perfectly in testing set (r = 0.97). The better trend observed on the training set (r = 0.99), too. This tight alignment indicates that the model captured the underlying response tendencies with high fidelity which is an essential prerequisite for downstream psychometric uses such as stochastic response simulation and item-parameter recovery.

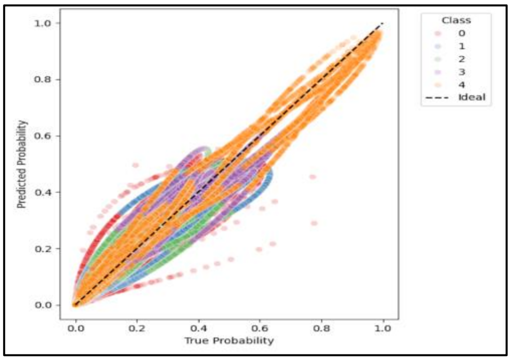


Figure 1: Predicted versus True Probabilities across Response Categories for Testing Items

However, when we translated these well-calibrated probability vectors into single categorical responses (via one draw per item to mimic human variability rather than using deterministic arg-max sampling), discrimination among adjacent score levels became more challenging, especially for the rarer categories (0-1). Table A2 details this pattern, reporting precision, recall, and F1 for every category, together with the overall Cohen's κ that reflects agreement beyond chance. In brief, the model delivers highly calibrated probabilities and moderate-to-strong categorical accuracy where it matters most (levels 2–4), while the expected drop in metrics for the low-frequency categories reflects both class imbalance and the deliberate injection of sampling noise.

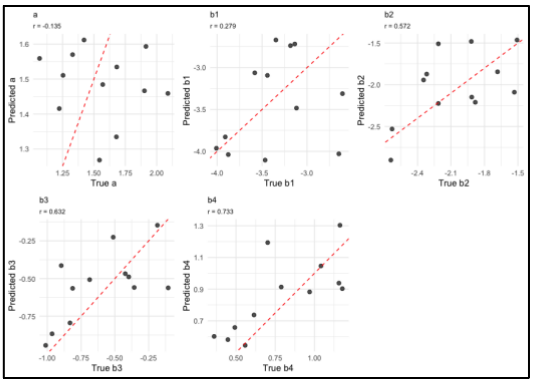


Figure 2: Scatterplots of Predicted versus True Item Parameters for Testing Items

Figure 2 showing field test items' predicted and true IRT parameters and Table 1 indicating overall numerical fit indices of those parameters together provide a consistent picture. For discrimination parameter (a), estimates were weakly and negatively correlated with the true values (r = –0.13), showing both noticeable scatter in Figure 2 and moderate error (RMSE ≈ 0.33). The slight negative bias (-0.07) and compressed range suggested the model flattens steep items and inflates shallow ones. For difficulty thresholds (b's), recovery improved monotonically from b_1 to b_4. The first threshold was the noisiest ($RMSE \approx 0.59, r = 0.28$), but accuracy doubled for the upper thresholds (b_3, b_4) where RMSE fell below 0.25 and correlations climbed above 0.60. The bias pattern was small and positive for $b_2 - b_4$, implying a slight right-shift of predicted step locations. Overall, slopes were poorly recovered, whereas later thresholds were estimated with moderate precision; early thresholds remained a concern.

Parameter	Bias	MAE	RMSE	r
a	-0.07	0.30	0.33	-0.13
b_1	-0.06	0.48	0.59	0.28
b_2	0.05	0.31	0.37	0.57
b_3	0.05	0.19	0.24	0.63
b_4	0.07	0.17	0.21	0.73

Table 1. Parameter Recovery Metrics for Testing Items

Figure 3 overlays the true (solid) and predicted (dashed) category response curves for a sample of items. For most items the ordering of curves was preserved and each predicted peak occurred near the true modal θ, confirming that the threshold structure was broadly captured. Consistent with the numeric bias, predicted curves often shift rightward, especially for the b_1 and b_2 steps, causing lower categories to dominate a wider θ-range than intended. Flattened peaks and broader overlaps reflected the underestimated discriminations, explaining why slope recovery was weak yet the model still yielded plausible probabilities.

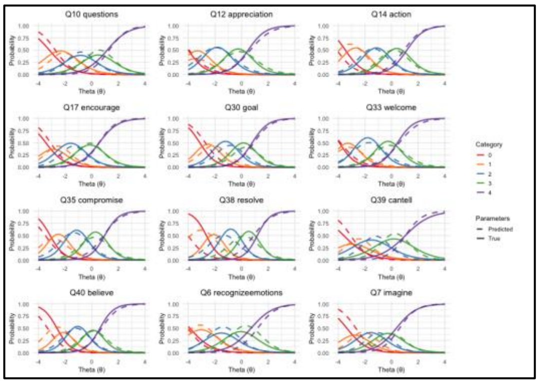


Figure 3: Predicted versus True Category Curves by Testing Items

5 Conclusions

AI-based field-testing approach in this study aims to improve the efficiency of traditional pretesting by simulating human examinee responses using AI, thereby reducing or if possible, eliminating the need for large-scale human data collection. Specifically, we investigated if the proposed approach could emulate graded response model by using a single DeBERTa-base model with item text and examinee θ to generate realistic responses to Likert-type rating scale items. The current study demonstrated that IRT statistics derived from AI-generated responses show moderate alignment with those obtained from human examinees. This suggests that the proposed architecture can approximate key features of human response behavior in rating-scale assessments and serve as a scalable tool for early-stage item evaluation.

Item-parameter recovery paints a nuanced picture: the model captures later thresholds ($b_3 - b_4$) with reasonable precision ($RMSE \leq 0.25, r \geq 0.63$) and preserves the qualitative ordering of category response curves, yet it underestimates discrimination (a) and the earliest threshold (b_1). These findings suggest that the architecture faithfully encodes item difficulty structure but still compresses slope information, a pattern consistent with the "flattened ICC" or items with negative discriminations documented in transformer-generated response data (Byrd & Srivastava, 2022; Maeda, 2025).

This study, while promising, has several limitations that warrant consideration. First, the item pool was restricted to a small set of Likert-type social-emotional learning items, limiting the generalizability of findings to other domains. Second, although the use of stochastic sampling from predicted probabilities offers a realistic alternative to deterministic predictions, it also introduces additional variance that can inflate classification error and reduce parameter recovery precision. Future implementations should incorporate multiple draws from the predicted probability distributions to reduce Monte Carlo variance by using Rubin's Rule (1987). Third, item parameter estimation was conducted on a relatively small number of training and testing items, which may limit the robustness of recovery analyses, particularly for slope parameters. Benedetto (2023) showed that the predictive power of transformers increased with increasing training sample size; therefore, the results of the current study may increase with larger number of training items. Finally, the study relied on a single pretrained DeBERTa model; further work is needed to explore how different model architectures, sizes, and fine-tuning strategies influence response quality and psychometric fidelity.

By modeling probabilistic item responses through a single transformer-based model and evaluating their psychometric viability, this study offers a scalable pathway toward AI-enhanced pretesting workflows. While improvements are needed, particularly in recovering item discriminations, the strong probability calibration and promising threshold estimates position this approach as a compelling tool to reduce workload and improve the speed and consistency in the traditional field-testing pipelines, especially in low-resource or early development contexts.

References

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2023). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education,* 1–53. https://doi.org/10.1007/s40593-023-00362-1

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Benedetto, L. (2023). A quantitative study of NLP approaches to question difficulty estimation (arXiv preprint arXiv:2305.10236). https://arxiv.org/abs/2305.10236

Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., & Turrin, R. (2021). On the application of transformers for estimating the difficulty of multiple-choice questions from text. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pila´n, H. Yannakoudakis & T. Zesch (Eds.), *Proceedings of the 16th workshop on innovative use of NLP for building educational applications* (pp. 147–157). Association for Computational Linguistics.

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys, 55*(9), 1–37.

Byrd, M., & Srivastava, S. (2022). Predicting difficulty and discrimination of natural language questions. In S. Muresan, P. Nakov & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics: Short papers 602 (Vol. 2,* pp. 119–130). Association for Computational Linguistics.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss. v048.i06

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* (arxiv preprint arxiv:1810.04805). https://arxiv.org/abs/1810.04805

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention* (arxiv preprint arxiv:2006.03654). https://arxiv.org/abs/2006.03654

Hsu, F. Y., Lee, H. M., Chang, T. H., & Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management, 54*(6), 969–984.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning*: With applications in python. Springer.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *The Devereux Student Strengths Assessment (DESSA)*. Kaplan Press.

Liu, Y., Bhandari, S., & Pardos, Z.A. (2024). Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis. *ArXiv, abs/2407.10899*.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization* (arxiv preprint arxiv:1711.05101). https://arxiv.org/abs/1711.05101

Lu, X., & Wang, X. (2024). Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*.

Maeda, H. (2025). Field-testing multiple choice questions with AI examinees: English Grammer Items. *Educational and Psychological Measurement, 85*(2), 221-244. https://doi.org/10.1177/00131644241281053

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). Guilford Press.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*. https://arxiv.org/abs/1912.01703

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, *17*(4), 2. doi:10.1002/j.2333-8504.1968.tb00153.x.

Shapiro, V. B., & LeBuffe, P. A. (2004). Strength-based assessment in children: The Devereux Early Childhood Assessment and the Devereux Student Strengths Assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (2nd ed., pp. 215–236). Guilford Press.

Zhou, Y., & Tao, C. (2020). Multi-task BERT for problem difficulty prediction. *In 2020 international conference on communications, information system and computer engineering (CISCE)* (pp. 213–216). Institute of Electrical and Electronics Engineers.

A Appendices

	Parameters	Training Items	Testing Items	All Items
Mean	a	1.49	1.56	1.51
	b_1	-3.32	-3.36	-3.33
	b_2	-1.94	-2.07	-1.97
	b_3	-0.47	-0.60	-0.50
	b_4	0.96	0.79	0.92
SD	a	0.28	0.31	0.28
	b_1	0.70	0.45	0.65
	b_2	0.52	0.39	0.49
	b_3	0.40	0.31	0.38
	b_4	0.40	0.30	0.38

Table A1: Descriptive Statistics of Calibrated Item Parameters by Dataset

Data	Category	Precision	Recall	F1-Score	Kappa	*r*
Training	0	0.13	0.12	0.13	0.16	0.99
	1	0.22	0.23	0.22		
	2	0.36	0.35	0.36		
	3	0.39	0.39	0.39		
	4	0.48	0.49	0.49		
Testing	0	0.08	0.07	0.08	0.16	0.97
	1	0.19	0.21	0.20		
	2	0.35	0.36	0.35		
	3	0.38	0.39	0.38		
	4	0.52	0.50	0.51		

Table A2: Per-category Classification Metrics with Overall Cohen's κ and Probability Correlation

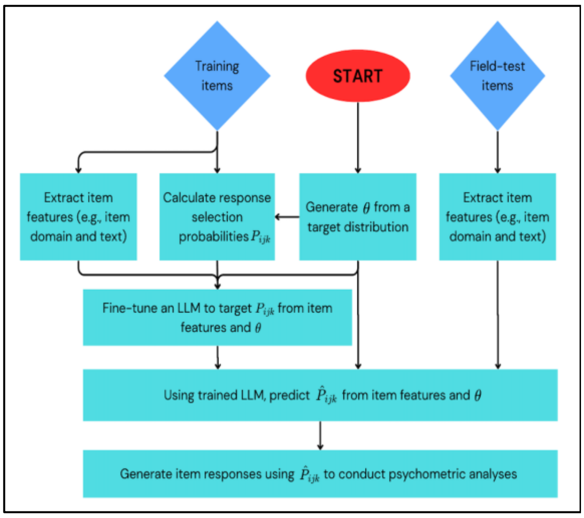


Figure A1: AI-Based Field-Test Data Generation Pipeline

An example of DESSA items:

Domain: Self-Awareness

I can recognize my strengths.

0. Never
1. Rarely
2. Sometimes
3. Often
4. Almost Always

Text consumed by the LLM model after processing item features data:

Domain: Self-Awareness Item: I can recognize my strengths.

Figure A2: DESSA Item Example and Corresponding Preprocessed LLM Input Text

Bias and Reliability in AI Safety Assessment: Multi-Facet Rasch Analysis of Human Moderators

Chunling Niu[1], Kelly Bradley[2], Biao Ma[1], Brian Waltman[1], Loren Cossette[1], & Rui Jin[3]

[1]University of the Incarnate Word
[2]University of Kentucky
[3]Shenzhen University, P. R. China

Abstract

Using Multi-Facet Rasch Modeling on 36,400 safety ratings of AI-generated conversations, we reveal significant racial disparities (Asian: 39.1%, White: 28.7% detection rates) and content-specific bias patterns. Simulations show that diverse teams of 8-10 members achieve over 70% reliability versus 62% for smaller homogeneous teams, providing preliminary evidence-based guidelines for AI-generated content moderation.

1 Background

As conversational AI systems proliferate, ensuring reliable human evaluation of AI-generated content safety becomes critical. Modern generative AI systems like LaMDA rely heavily on human judgment to assess response safety, particularly for nuanced content requiring contextual understanding. However, the demographic composition of evaluation teams and its impact on AI safety assessment remains understudied.

Recent work documents bias in content moderation (Aroyo et al., 2023; Goyal et al., 2022), but few studies examine how rater demographics affect evaluation of AI-generated conversational content specifically. This distinction matters because AI-generated conversations present unique challenges: subtle harmful content, contextual nuances, and adversarial prompting designed to elicit unsafe responses.

The Diversity in Conversational AI Evaluation for Safety (DICES) dataset (Aroyo et al., 2023) established foundations for understanding demographic effects in AI safety evaluation but lacks detailed bias analysis or reliability optimization guidelines. Prior research shows significant demographic disparities in toxicity ratings, particularly affecting African American and LGBTQ populations (Goyal et al., 2022), yet the interaction between rater demographics and AI conversation characteristics remains unexplored.

2 Aims

This study addresses three critical research questions:

1. **Quantify human rater disparities**: Do significant demographic differences exist in safety detection rates for AI-generated conversations, and what is their magnitude?
2. **Identify content-specific patterns**: How do demographic bias patterns vary across different AI conversation topics (health, political, legal, racial content, etc.)?
3. **Optimize rater team composition**: What rater team configurations achieve optimal reliability while maintaining demographic diversity for AI safety evaluation?

We employ Multi-Facet Rasch Modeling (MFRM) to simultaneously model rater, conversation types, and demographic effects while providing actionable guidelines for assembling effective AI-generated content safety evaluation teams.

3 Sample(s)

We analyze the DICES-350 dataset (Aroyo et al., 2023) containing safety evaluations of 350

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 393–397

adversarial human-AI conversations generated using Google's LaMDA. It is a relatively new dataset from NeurIPS that is gaining traction in the field. The dataset includes:

3.1 Conversations

Three hundred and fifty (n=350) multi-turn interactions were created as the corpus data by human agents instructed to generate adversarial prompts designed to elicit unsafe responses. Conversations span health (8%), political (18%), racial (25%), gender/sexual (14%), legal (3%), violence (1%), and miscellaneous (30%) topics. Expert annotations indicate 40% benign, 20% debatable, 20% moderate, and 20% extreme harm levels.

3.2 Raters

One hundred and four (n=104) demographically diverse raters provided 36,400 total safety judgments. Demographics were consolidated for statistical power based on initial exploratory analysis results:

- **Race**: Asian (n=21), White (n=25), Other races (n=58, including Black/African American, Latin X, Latino, Hispanic or Spanish Origin, and Multiracial)
- **Age**: GenZ 18-24 (n=49), Millennial 25-34 (n=28), GenX+ 35+ (n=27)
- **Gender**: Male (n=47), Female (n=57)

3.3 Ratings

Granular safety assessments were collected from the raters across conversation legibility, harmful content (8 sub-questions), unfair bias (4 sub-questions), misinformation, political affiliation, and policy violations using three-point scales (*No/Unsure/Yes*).

4 Methods

4.1 Multi-Facet Rasch Analysis

We implemented MFRM using generalized linear mixed models with logistic regression:

$$logit\left(P(unsafe_{rating} = 1)\right) = \beta^0 + \beta^1(race) + \beta^2(content) + \beta^3(gender) + \beta^4(race \times content) + \beta^5(gender \times content) + (1|rater) + (1|conversation) \quad (1)$$

This simultaneously estimates conversation difficulty (random effect), rater severity (random effect), racial bias (fixed effects), content-specific bias (interaction terms), and gender effects. Model fitted using maximum likelihood estimation in R (lme4 package).

4.2 Empirical Reliability Simulation

We developed bootstrap simulation using real rating patterns:

1. *Sample teams from actual demographic distributions (3-10 members)*
2. *Calculate pairwise reliability for multiply rated conversations*
3. *Estimate consensus via majority vote aggregation*
4. *Bootstrap replicates across 500 iterations for stable estimates*

We tested 12 team configurations across four content types, simulating realistic AI safety evaluation scenarios.

5 Results

5.1 Demographic Effects

Our MFRM analysis reveals differential patterns across demographic groups in safety detection of AI-generated conversational content. The model achieved excellent fit (AIC: 9,716.9, BIC: 9,933.2) with successful convergence across all parameters.

Primary Demographic Effects: While we analyzed race, age, and gender effects simultaneously, racial differences emerged as the most substantial and consistent predictor of safety detection patterns. Age effects were modest (GenZ vs Millennial: $\beta = -0.12$, $p = 0.67$; GenX+ vs Millennial: $\beta = +0.08$, $p = 0.78$), and gender effects were non-significant (Male vs Female: $\beta = -0.03$, $p = 0.85$). Based on these preliminary findings and space constraints, we focus our detailed analysis on racial bias patterns, which showed the strongest effects and clearest interaction patterns with content types.

Figure 1 shows the box plots of rater severity (in logits) across 28 demographic subgroups defined by combinations of race/ethnicity, gender, and age, revealing systematic differences in how different demo groups evaluate AI-generated content.

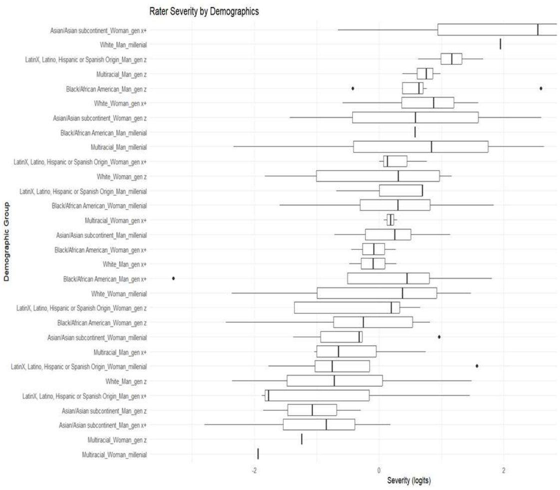


Figure 1: Rater severity by rater demographics

Safety Detection Rate Disparities: Asian raters demonstrated the highest safety detection rates at 39.1%, followed by Other races at 33.9%, and White raters at 28.7%. These differences represent substantial effect sizes, with Asian raters being 36% more likely to identify safety concerns in AI-generated conversations compared to White raters, and 15% more likely than Other race raters.

Figure 2 below shows harm detection rates across various content categories by racial group, revealing substantial variation in detection patterns both within and across demographic groups, with notably higher detection rates for certain harm types like legal issues and violent content.

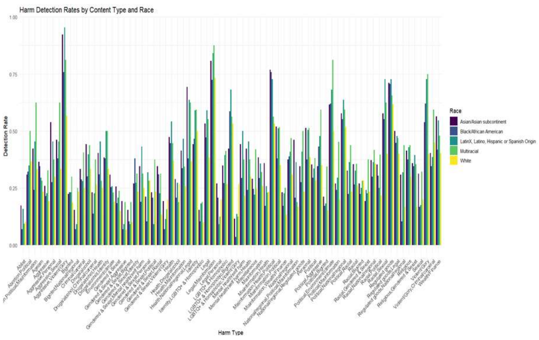


Figure 2: Harm detection rate by content type and race.

Statistical Significance: The racial effects were statistically significant in the expected direction:

- ***Other Race vs Asian***: β = -0.73, SE = 0.38, p = 0.059†
- ***White vs Asian***: β = -1.05, SE = 0.47, p = 0.025*

These findings suggest that raters' race significantly influences the perceived safety of AI-generated conversational content, with important implications for AI safety evaluation team composition.

5.2 Content-Specific Bias

A critical finding is that racial bias in AI safety assessment varies significantly across conversation topics, challenging assumptions of uniform demographic effects across all AI-generated content.

Significant Race × Content Interactions:

- ***Other Race* × *Miscellaneous content***: β = +0.57, SE = 0.23, p = 0.013* (Non-Asian/White raters more likely to detect safety issues in general AI conversations)
- ***White* × *Health content***: β = +0.68, SE = 0.39, p = 0.076† (White raters trend toward higher detection in health-related AI conversations)
- ***Other Race* × *Political content***: β = +0.47, SE = 0.25, p = 0.062† (Non-Asian/White raters are stricter rating political AI content)

Figure 3 below illustrates the significant race × content interactions, where Other Race raters show heightened detection for miscellaneous (β = +0.57) and political content (β = +0.47), while White raters demonstrate increased sensitivity to health-related content (β = +0.68), revealing content-specific deviations from the overall pattern of Asian raters having highest detection rates.

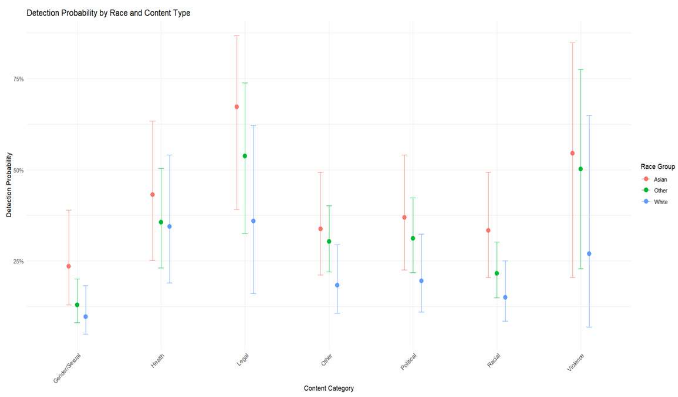


Figure 3: Harm detection probability by race and content type

AI Content Difficulty Hierarchy: Also shown in Figure 3 above, among AI-generated conversations, legal content showed highest baseline safety detection rates (β = +1.91), followed by violence-related (β = +1.52) and health content (β = +0.90), with gender/sexual AI conversations being most difficult to assess for safety violations (baseline category).

These interaction effects suggest that bias in AI safety evaluation is not uniform but depends critically on the topic and content type of AI-generated conversations, requiring content-specific approaches to bias mitigation in AI evaluation workflows.

5.3 Simulation-Based Evidence for Rater Team Configuration and Reliability

Our empirical reliability simulation demonstrates that rater team composition significantly impacts the reliability of AI safety assessments, with clear patterns visible across multiple dimensions.

Team Size Effects for AI Safety Evaluation: The visualization (top left panel) in Figure 4 below reveals a consistent upward trend in reliability as team size increases across all content types, with diminishing returns at larger sizes:

- **Teams of 10**: 70.3% mean reliability (convergence point for all content types)
- **Teams of 9**: 69.5% mean reliability
- **Teams of 8**: 69.0% mean reliability
- **Teams of 6**: 65.8% mean reliability
- **Teams of 3**: 62.2% mean reliability

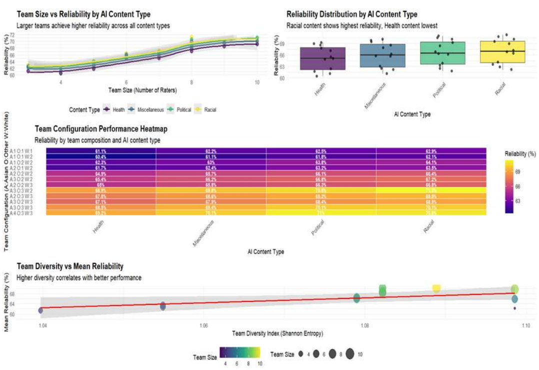


Figure 4: Empirical Reliability Simulation Analysis for Optimal Rater Team Design for AI-Generated Content Moderation.

Optimal Configurations for AI Safety Teams: The heatmap analysis (Figure 4, middle panel) clearly identifies two configurations that achieve the critical ≥70% reliability threshold:

1. **Asian:4 Other:3 White:3** (10 members): 70.3% mean reliability - the top performer across all content types
2. **Asian:3 Other:3 White:2** (8 members): 70.1% mean reliability - demonstrating cost-effective excellence

Content-Specific Reliability Patterns: Box plot analysis (Figure 4, top right panel) reveals systematic differences in evaluation difficulty across AI conversation types:

- **Racial AI content**: 67.0% mean reliability (highest, tightest distribution)
- **Political AI content**: 66.6% mean reliability (moderate variability)
- **Miscellaneous AI content**: 66.0% mean reliability (moderate variability)
- **Health AI content**: 65.2% mean reliability (lowest, highest variability)

The heatmap confirms these patterns, with racial content consistently showing the highest reliability values (yellow/orange cells) across all team configurations, while health content shows the lowest values (purple/blue cells).

Reliability Thresholds for AI Evaluation: Only 17% of tested rater team configurations (2 out of 12) achieved ≥70% reliability, with performance ranging from 60.4% to 71.3%. The top 6 configurations all required balanced demographic representation and achieved 66.4%-70.3% reliability, indicating that 70% represents a practical upper bound for AI safety evaluation.

Diversity-Reliability Relationship: The scatter plot analysis (Figure 4, bottom panel) demonstrates a clear positive relationship between team demographic diversity (Shannon entropy) and mean reliability ($r = 0.579$, $p < 0.01$). Larger, more diverse teams (represented by bigger circles in the upper right) consistently outperform smaller, less diverse configurations, providing quantitative evidence that demographic diversity enhances rather than hinders AI safety evaluation performance.

6 Conclusions

This preliminary study provides the first comprehensive analysis of human rater demographic bias and team reliability in AI-generated conversational content safety assessment using Multi-Facet Rasch Modeling. Our key findings, supported by detailed analysis and visualizations showing clear trends and patterns, demonstrate:

1. **Significant racial disparities** exist in AI safety assessment, with Asian raters 36% more likely to detect safety concerns in AI-generated content than White raters
2. **Content-specific bias patterns** in AI-generated content evaluation require targeted mitigation strategies, with racial content consistently achieving highest reliability and health content presenting greatest challenges
3. **Optimal AI safety rater team composition** involves minimum 8-10 diverse raters to achieve ≥70% reliability, with empirical simulation evidence showing convergence across content types at this threshold
4. **Diversity enhances reliability** in AI safety evaluation, with a strong positive correlation (r = 0.579) between team diversity and performance

Empirical reliability simulation analysis results provide practitioners with actionable guidance for team assembly, clearly demonstrating the reliability benefits of larger, diverse teams and content-specific performance patterns that can inform specialized evaluation strategies.

These findings provide evidence-based guidelines for assembling fair and reliable AI safety evaluation teams. As conversational AI systems scale and become more sophisticated, understanding and optimizing the human evaluation component becomes increasingly critical for maintaining both consistency and equity in AI safety assessment.

Our research also establishes a methodological framework for bias analysis in AI safety evaluation and demonstrates the practical value of psychometric approaches for understanding complex judgment tasks in AI development. Future work should examine intervention strategies for bias reduction and extend this analysis to additional AI systems and conversation domains.

The implications extend beyond academic research to practical AI development: our findings suggest that investing in diverse, appropriately sized evaluation teams is not just an ethical imperative but a technical requirement for reliable AI safety assessment. As the field moves toward more sophisticated conversational AI systems, these insights will become increasingly valuable for ensuring safe and equitable AI deployment.

Acknowledgments

We thank the creators of the Diversity in Conversational AI Evaluation for Safety (DICES) dataset for making this research possible and the anonymous reviewers for their valuable feedback.

References

Aroyo, L., Taylor, A. S., Díaz, M., Homan, C. M., Parrish, A., Serapio-García, G., Prabhakaran, V., & Wang, D. (2023). DICES dataset: Diversity in conversational AI evaluation for safety. *Advances in Neural Information Processing Systems*, *36*, 53330-53342.

Chancellor, S., Blackwell, L., De Choudhury, M., & Davison, L. (2022). Understanding demographic bias in content moderation decisions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1-15.

Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y. L., & Rieser, V. (2022). SafetyKit: First aid for measuring safety in open-domain conversational systems. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 284-299.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Goyal, N., Kivlichan, I. D., Rosen, R., & Vasserman, L. (2022). Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1-28.

Sebok-Syer, S. S., Chahine, S., Watling, C. J., Goldszmidt, M., Cristancho, S., & Lingard, L. (2018). Considering the interdependence of clinical performance: implications for assessment and entrustment. *Medical Education*, *52*(9), 970-980.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., … & Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Dynamic Bayesian Item Response Model with Decomposition (D-BIRD): Modeling Cohort and Individual Learning Over Time

Hansol Lee[1] hansol@stanford.edu
Jason B. Cho[2] bc454@cornell.edu
David S. Matteson[2] dm484@cornell.edu
Benjamin W. Domingue[1] bdomingu@stanford.edu

[1]**Stanford University**
[2]Cornell University

Abstract

We present D-BIRD, a Bayesian dynamic item response model for estimating student ability from sparse, longitudinal assessments. By decomposing ability into a cohort trend and individual trajectory, D-BIRD supports interpretable modeling of learning over time. We evaluate parameter recovery in simulation and demonstrate the model using real-world personalized learning data.

1 Introduction

As personalized learning platforms become more widespread, students increasingly encounter assessments that are short, embedded, and distributed over time. These settings produce sparse but longitudinal data, creating new opportunities—and challenges—for educational measurement. The emerging goal is no longer just to estimate ability at isolated time points, but to track how ability evolves over time, both individually and relative to peers.

Item response theory (IRT) provides a principled framework for estimating latent traits such as ability, but traditional IRT assumes ability is fixed within and across assessments. Dynamic extensions relax this assumption by modeling ability as a time-varying stochastic process (e.g., Martin and Quinn, 2002; Wang et al., 2013; Kim et al., 2023; Tripathi and Domingue, 2019; Imai et al., 2016; Sun et al., 2025). However, most existing models treat students independently or borrow strength only through global priors, limiting their ability to capture cohort-level trends.

We introduce **D-BIRD** (Dynamic Bayesian Item Response model with Decomposition), a fully Bayesian dynamic IRT model that decomposes each student's ability into two components: a cohort-level trend capturing shared change over time, and a student-specific deviation capturing personalized growth. This structure enables the model to borrow information across students while preserving heterogeneity in learning patterns. We perform posterior inference via Pólya-Gamma augmentation (Polson et al., 2013), which enables efficient sampling and calibrated uncertainty quantification for logistic models.

D-BIRD addresses a growing measurement need in personalized education: estimating learning trajectories in a statistically coherent, interpretable way—even under sparsity. By explicitly modeling both shared and individual dynamics, it provides a foundation for learner feedback, program evaluation, and cohort monitoring.

We validate D-BIRD through simulation and empirical analysis. First, we assess parameter recovery and test its key components via ablation. Then, we apply the model to K–12 reading data from a digital learning platform, demonstrating its ability to recover cohort trends and individual trajectories under real-world constraints.

2 Model Specification

We present D-BIRD, a dynamic IRT model that decomposes latent ability into a shared cohort trend and student-specific deviations evolving over time. Let $Y_{i,t,j}$ denote the binary response (correct/incorrect) of student $i \in \{1, \dots, N\}$ at time $t \in \{1, \dots, T\}$ on item $j \in \{1, \dots, J\}$. The goal is to estimate each student's latent proficiency $\theta_{i,t}$ at each time point. The model is defined as:

$$Y_{i,t,j} \sim \text{Bernoulli}(\pi_{i,t,j}), \tag{1a}$$
$$\pi_{i,t,j} = \text{logit}^{-1}(\theta_{i,t} - d_j), \tag{1b}$$
$$\theta_{i,t} = \mu_t + \beta_{i,t}, \tag{1c}$$
$$\Delta\mu_t \sim \mathcal{N}(0, \sigma^2_{\Delta\mu}), \tag{1d}$$
$$\Delta\beta_{i,t} \sim \mathcal{N}(0, \sigma^2_{\Delta\beta_i}), \tag{1e}$$

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 398–405
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

where $\Delta\mu_t := \mu_t - \mu_{t-1}$, $\Delta\beta_{i,t} := \beta_{i,t} - \beta_{i,t-1}$, and d_j is the difficulty of item j.[1]

Equations (1a)–(1b) define a Rasch model (Rasch, 1980), where the probability of a correct response depends on the difference between ability and item difficulty. Like other dynamic extensions of IRT, D-BIRD embeds this structure within a temporal state-space framework by modeling ability as a time-indexed latent process. In doing so, it fits within a broader class of dynamic linear models (West et al., 1985; West and Harrison, 2006), where the key modeling choice lies in the prior placed on the latent trajectory.

In discrete-time settings, common priors over ability include AR(1) processes, as in Wang et al. (2013); Sun et al. (2025), and Gaussian random walks, as in Martin and Quinn (2002) and Kim et al. (2023), where each student's ability is modeled as a single latent process with a shared innovation variance. Other work such as Tripathi and Domingue (2019) has explored continuous-time priors such as Gaussian processes, which are particularly relevant when modeling irregularly spaced assessments. While these approaches support temporal smoothing, they typically assume a uniform degree of smoothness across individuals and do not separate shared trends from individual deviations—limiting interpretability when comparing student growth to broader cohort patterns.

D-BIRD also adopts a random walk over ability but structures it differently from prior models. Its key innovation is an additive decomposition of ability into two components (Equation 1c): (1) a cohort trend μ_t, shared across all students and capturing group-level change, and (2) a student-specific deviation $\beta_{i,t}$, representing individual progress relative to that trend. Both components evolve over discrete time as Gaussian random walks with distinct innovation variances: μ_t with a shared variance $\sigma^2_{\Delta\mu}$ (Equation 1d), and $\beta_{i,t}$ with student-specific variances $\sigma^2_{\Delta\beta_i}$ (Equation 1e). This structure allows for heterogeneous smoothness across individuals while situating trajectories within a common temporal reference.

This decomposition allows D-BIRD to be both flexible and interpretable. It accommodates heterogeneity in student-level learning while supporting cohort-based comparisons and population-level monitoring. In doing so, D-BIRD offers a principled framework for measuring learning progress over time—balancing individualized adaptation with shared structure across the student population.

[1] We assume item difficulties d_j are known *a priori*, reflecting common practice in operational assessments where items are pre-calibrated and drawn from a stable pool. While D-BIRD can be extended to estimate item parameters jointly, we focus here on ability estimation under known difficulties.

3 Inference

We perform fully Bayesian inference for the model specified in Equation (1). Let the observed responses be denoted by $\boldsymbol{y} := \{y_{i,t,j}\}_{i=1,\ldots,N;\, t=1,\ldots,T;\, j=1,\ldots,J}$. The primary latent variables include the cohort-level trajectory $\boldsymbol{\mu} := \{\mu_t\}_{t=1}^T$ and the student-specific deviations $\boldsymbol{\beta} := \{\beta_{i,t}\}_{i=1,\ldots,N;\, t=1,\ldots,T}$.

Prior specification. Initial values follow Gaussian priors: $\mu_1 \sim \mathcal{N}(0, \sigma^2_\mu)$ and $\beta_{i,1} \sim \mathcal{N}(0, \sigma^2_{\beta_i})$. Subsequent values evolve via Gaussian random walks:

$$\mu_t \sim \mathcal{N}(\mu_{t-1}, \sigma^2_{\Delta\mu}), \quad \beta_{i,t} \sim \mathcal{N}(\beta_{i,t-1}, \sigma^2_{\Delta\beta_i}).$$

Variance components include:

- σ^2_μ: initial variance of the cohort trend,
- $\boldsymbol{\sigma}^2_\beta := \{\sigma^2_{\beta_i}\}$: initial variances for student-specific offsets,
- $\sigma^2_{\Delta\mu}$: innovation variance for the cohort trend,
- $\boldsymbol{\sigma}^2_{\Delta\beta} := \{\sigma^2_{\Delta\beta_i}\}$: innovation variances for individual trajectories.

We place improper scale-invariant priors $p(\sigma^2) \propto 1/\sigma^2$ on the innovation variance terms, following the Jeffreys prior (Jeffreys, 1946). This prior is widely used in hierarchical Bayesian models for its invariance under scale transformations and its flexibility in allowing the smoothness of latent trajectories to be learned from the data. It also enables efficient Gibbs sampling via conjugate inverse-gamma updates. While improper and non-regularizing, this prior performs well when sufficient longitudinal data are available per individual (Gelman, 2006), as is typically the case in our setting. By contrast, we place half-Cauchy priors with scale 1, $C^+(0,1)$, on the initial variance parameters σ^2_μ and σ^2_β, to provide regularization and support stable estimation at the first time point.

The full posterior is:

$$p(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2_\mu, \sigma^2_{\Delta\mu}, \boldsymbol{\sigma}^2_\beta, \boldsymbol{\sigma}^2_{\Delta\beta} \mid \boldsymbol{y}).$$

Pólya-Gamma data augmentation. To address the non-conjugacy of the Bernoulli-logistic likelihood, we adopt the Pólya-Gamma (PG) data augmentation framework of Polson et al. (2013). Each observation likelihood can be re-expressed as:

$$
\begin{aligned}
&f(y_{i,t,j}|\mu_t, \beta_{i,t}, d_j) \\
&= \frac{\exp\{(\mu_t + \beta_{i,t}) - d_j\}^{y_{i,t,j}}}{1 + \exp\{(\mu_t + \beta_{i,t}) - d_j\}}. \\
&\propto \int_0^\infty \exp\left\{\kappa_{i,t,j}((\mu_t + \beta_{i,t}) - d_j)\right\} \\
&\qquad \exp\left\{ - \frac{\omega((\mu_t + \beta_{i,t}) - d_j)^2}{2}\right\} p(\omega) d\omega, \\
&\propto \int_0^\infty \mathcal{N}(\kappa_{i,t,j}|\omega(\theta_{i,t} + \beta_{i,t} - d_j), \omega) p(\omega) d\omega,
\end{aligned}
$$

where $\kappa_{i,t,j} = y_{i,t,j} - \frac{1}{2}$ and $\omega \sim PG(1, 0)$. The Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathcal{R}$, is denoted as PG(b,c), is defined as

$$
X \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/4\pi^2},
$$

where the $g_k \sim \text{Gamma}(b, 1)$ and $\overset{D}{=}$ denotes equality in distribution.

We exploit the banded structure of the random walk priors to perform efficient Gibbs sampling using the sparse Cholesky algorithm of Rue (2001). Each iteration scales linearly in the number of students N and time steps T. This structure, combined with the conjugacy induced by PG augmentation, enables exact posterior inference even in high-dimensional settings. PG-based samplers are also geometrically ergodic (Wang and Roy, 2018), providing theoretical guarantees for convergence.

3.1 Comparison with alternative methods

Fully Bayesian inference offers calibrated uncertainty estimates, which are particularly valuable in sparse data settings. However, exact inference in logistic IRT models is challenging due to the non-conjugacy of the likelihood and the high dimensionality introduced by dynamic latent structures.

General-purpose samplers such as the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), implemented in Stan (Carpenter et al., 2017), are widely used for models with complex posteriors due to their automatic tuning and robust convergence properties (Livingstone et al., 2019; Neal et al., 2011). Yet these methods are often computationally infeasible for high-dimensional, structured time-series models like dynamic IRT due to poor scaling and slow mixing (Thomas and Tu, 2021; Sacher et al., 2021).

To improve scalability, many existing dynamic IRT models adopt approximate inference: Wang et al. (2013) approximate the likelihood using a mixture-of-normals; Imai et al. (2016) and Kim et al. (2023) use variational inference. While efficient, these methods may introduce bias and understate posterior uncertainty.

In contrast, Pólya-Gamma augmentation enables exact posterior inference by transforming the logistic likelihood into a conditionally Gaussian form. This allows conjugate updates for latent trajectories and variance components, making it well-suited to dynamic IRT models like D-BIRD. Although less flexible than black-box or amortized inference approaches, PG-based Gibbs sampling provides a tractable, theoretically grounded alternative that supports full Bayesian inference at scale.

4 Simulation Study

4.1 Design

We conduct a simulation study to assess the parameter recovery performance of D-BIRD in comparison with two baselines:

- **Global-RW:** No cohort trend; all students share the same innovation variance (analogous to the model specification used in Kim et al. (2023); Martin and Quinn (2002)):

$$
\theta_{i,t} = \beta_{i,t}, \quad \Delta\beta_{i,t} \sim \mathcal{N}(0, \sigma^2_{\Delta\beta}).
$$

- **Hetero-RW:** No cohort trend; each student has their own innovation variance:

$$
\theta_{i,t} = \beta_{i,t}, \quad \Delta\beta_{i,t} \sim \mathcal{N}(0, \sigma^2_{\Delta\beta_i}).
$$

This design allows us to assess how each feature improves recovery of latent ability trajectories and model parameters under controlled conditions.

We simulate response data for $N = 150$ students over $T = 100$ sessions, with 10 items per session. Ability is generated according to the D-BIRD specification (Equation 1), which includes both a global cohort trend μ_t and individualized deviations $\beta_{i,t}$. The cohort trend is simulated as a smooth Gaussian random walk:

$$
\mu_1 \sim \mathcal{N}(0, 0.1), \quad \Delta\mu_t \sim \mathcal{N}(0, 0.05).
$$

This latent trend is shared across all students and governs the population-wide evolution of ability.

To introduce heterogeneity in latent trajectories, we generate student-specific deviations $\beta_{i,t}$ with varying levels of smoothness. Students are split into two groups: the first 75 have low-variance random walks (more stable learning), while the remaining 75 have higher-variance trajectories:

$$\beta_{i,t} = \hat{\beta}_{i,1} - \frac{1}{150}\sum_{i=1}^{150}\hat{\beta}_{i,t}$$

$$\hat{\beta}_{i,1} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\beta_i}),$$

$$\sigma^2_{\beta_i} \sim \text{Gamma}(5, 10),$$

$$\Delta\hat{\beta}_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\Delta\beta_i}),$$

$$\sigma^2_{\Delta\beta_i} \sim \begin{cases} \text{Gamma}(5, 500), & \text{if } i \leq 75 \text{ (Group A)} \\ \text{Gamma}(5, 10), & \text{if } i > 75 \text{ (Group B)}. \end{cases}$$

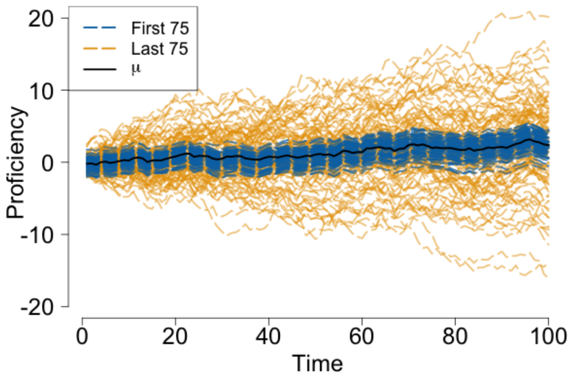


Figure 1: Simulated latent abilities $\theta_{i,t}$ for 150 students over 100 sessions. Orange lines represent the trajectories of the first 75 students; blue lines correspond to the remaining 75. The global trend shared across all students is shown in black.

Figure 1 shows the simulated ability trajectories, where group differences in smoothness and the shared cohort pattern are visible. Item difficulties are drawn from $d_{i,j,t} \sim N(\theta_{i,t}, 0.5)$.

This design creates a data-generating process with two key properties: (1) a smooth global trajectory shared across all students, and (2) heterogeneous individual learning dynamics. D-BIRD is designed to exploit both sources of structure, while the baseline models can only recover one or the other. Each simulation is replicated 250 times, and recovery is evaluated using mean squared error (MSE), empirical coverage (EC), and mean credible interval width (MCIW).

4.2 Results

Table 1 summarizes model performance across 250 replications. D-BIRD consistently achieves the lowest mean squared error (MSE), indicating superior accuracy in recovering latent ability trajectories. This reflects its ability to capture both the global trend and student-specific deviations—structure explicitly encoded in the data-generating process.

By contrast, the Global-RW model performs worst. Because it assumes a single shared innovation variance and lacks a cohort trend, it cannot accommodate the observed heterogeneity in trajectory smoothness across students. This mismatch leads to oversmoothing and inflated error, particularly for students with rapidly changing trajectories.

The Hetero-RW model improves on Global-RW by allowing individualized evolution variances. However, it treats each student's trajectory as independent, ignoring the shared global trend present in the data. As a result, it fails to borrow strength across students and exhibits higher estimation error than D-BIRD. In contrast, D-BIRD strikes a balance: it captures population-level structure via the cohort trend μ_t, while flexibly adapting to individual variation through student-specific deviations $\beta_{i,t}$. This enables more stable and accurate recovery, especially in the presence of sparse data.

D-BIRD also outperforms both baselines in terms of uncertainty quantification. It achieves near-nominal empirical coverage (~96%) with the narrowest credible intervals, as shown by the lowest MCIW. Hetero-RW exhibits undercoverage despite wide intervals, suggesting unstable variance estimation. Global-RW maintains nominal coverage but at the cost of overly wide intervals, due to its inability to represent individual variation. Overall, D-BIRD provides not only more accurate point estimates, but also sharper and more reliable posterior uncertainty.

5 Empirical Application

We apply D-BIRD to longitudinal assessment data from a widely used digital K–12 learning platform to illustrate its practical utility. The goal is to show how the model recovers interpretable learning trajectories at both the cohort and individual levels over time. We also compare D-BIRD to static IRT estimates of ability, highlighting the added insight gained from dynamic modeling of student ability.

Model	MSE	EC	MCIW
D-BIRD	**0.216 (0.008)**	**0.960 (0.004)**	**1.791 (0.03)**
Global-RW	0.270 (0.011)	0.944 (0.005)	1.993 (0.054)
Hetero-RW	0.260 (0.013)	0.901 (0.038)	1.801 (0.134)

Table 1: Posterior recovery metrics for student trajectories $\theta_{i,t}$, comparing our proposed model, D-BIRD, against two baselines, Global-RW and Hetero-RW. Metrics include mean squared error (MSE), empirical coverage (EC), and empirical credible interval width (ECIW), with standard deviations shown in parentheses.

5.1 Data and Setup

Students on the platform begin with a full-length assessment comprising approximately 25 items drawn from a pre-calibrated Rasch item pool. Based on these initial estimates of ability, students are assigned a personalized instructional sequence, with each module followed by a brief 5-item quiz. Full-length assessments are re-administered periodically, providing updated proficiency estimates from static IRT and allowing for instructional adaptation. All item difficulties are known and expressed in Rasch logits.

In our analysis, we focus on two cohorts—Kindergarten (Grade 0) and Grade 5—to capture developmental contrasts in growth patterns. For both cohorts, we restrict the sample to students who completed at least four full-length assessments and truncate time series at 40 weeks. The final analytic sample includes 101 Kindergarten students and 311 Grade 5 students. For Kindergarten, the median observation span was 37 weeks, with a median of 19 active weeks and 10 responses per active week. For Grade 5, the median span was 39 weeks, with 20 active weeks and 14 responses per active week.

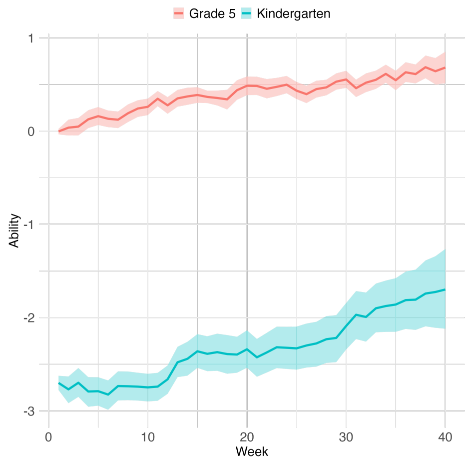


Figure 2: Estimated cohort-level ability trends μ_t for Kindergarten and Grade 5. Bands show 95% credible intervals.

5.2 Methods

To establish a static IRT baseline, we estimate each student's ability at the time of each full-length assessment using a Rasch model with the pre-calibrated item difficulties. Specifically, we compute the maximum a posteriori (MAP) estimate of ability under a logistic item response function and a Gaussian prior $\theta \sim \mathcal{N}(0, 5^2)$. These estimates serve as snapshot summaries of student proficiency at irregular time points and are used for visual comparison with dynamic trajectories estimated by D-BIRD.

We then fit D-BIRD separately for each cohort, using the Bayesian inference procedure described in Section 3. The model is estimated using 10,000 burn-in iterations followed by 10,000 posterior samples. We use the pre-calibrated item difficulties provided by the platform. D-BIRD yields posterior distributions for both the cohort-level trend μ_t and the individual-specific deviations $\beta_{i,t}$ at weekly resolution.

5.3 Results

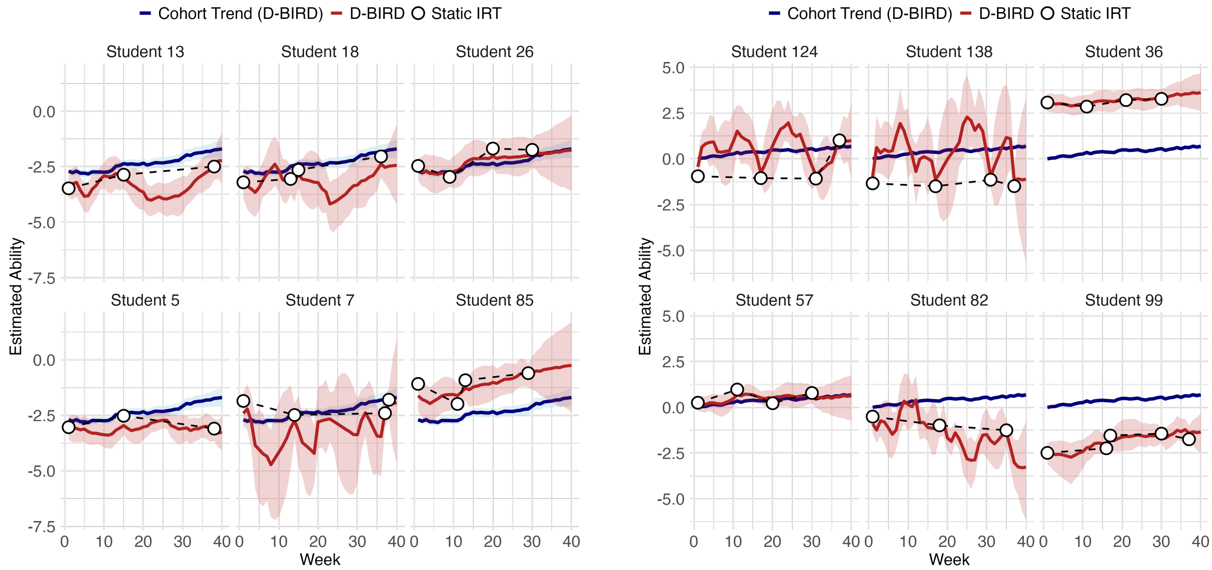
5.3.1 Cohort-Level Trends

Figure 2 shows the estimated cohort-level trends and their 95% credible intervals over the 40-week period for Kindergarten and Grade 5 cohorts. As expected, Kindergarten students exhibit lower baseline ability ($\hat{\mu}_1^{\mathrm{G0}} = -2.34$; 95% CI: [-2.55, -2.12]) than Grade 5 students ($\hat{\mu}_1^{\mathrm{G5}} = 0.39$; 95% CI: [0.30, 0.49]). Kindergarten students exhibited steady growth (mean slope = 0.026 logits/week), while Grade 5 trends were flatter (mean = 0.018 logits/week), suggesting slower average gains.

5.3.2 Individual Ability Trajectories

Figures 3a and 3b present D-BIRD ability trajectories for selected students in Kindergarten and Grade 5, respectively.

Kindergarten cohort. Students 26 and 85 both show upward trends in their static IRT scores, but



(a) Kindergarten (b) Grade 5

Figure 3: Estimated ability trajectories for selected students in Kindergarten (left) and Grade 5 (right). Red lines represent D-BIRD posterior means of ability with 95% credible intervals; blue lines show estimated cohort trend with 95% credible intervals. White circles indicate static IRT estimates from full-length assessments, connected by dashed lines for visual continuity (not model-derived).

D-BIRD reveals important distinctions. While Student 26 tracks closely with the cohort trend, Student 85 consistently outperforms it—something obscured without the group-level benchmark. In contrast, Student 5 appears to decline over time, falling further below the cohort average.

Static scores for Students 13, 18, and 26 appear similar at first glance, but D-BIRD uncovers meaningful differences in learning dynamics and uncertainty. Student 13 and 18 both show a mid-year dip, suggesting potential struggle despite an upward endpoint. Student 18's wide posterior band reflects high uncertainty due to sparse data. Student 26 maintains steady growth in line with the cohort, highlighting the value of interpreting performance in temporal and contextual terms.

Grade 5 cohort. Students 36, 57, and 99 follow visually similar static score trajectories, yet D-BIRD differentiates them sharply when viewed against the cohort trend. Student 36 consistently outperforms the cohort while showing stable progress; Student 57 remains aligned with the cohort; and Student 99 lags well behind. These distinctions demonstrate how D-BIRD contextualizes student ability trajectories to the cohort trend.

Student 124 illustrates a different case. Their static scores remain low until a notable jump on the last full-length test. However, D-BIRD estimates their ability to have already increased in the weeks prior, indicating that quiz-level responses captured learning gains before they appeared in test scores. This fluctuating trajectory contrasts with the smoother paths of Students 36, 57, and 99, highlighting D-BIRD's sensitivity to between-test dynamics.

Finally, Students 82 and 138 both underperform on full-length tests, but their trajectories diverge. D-BIRD estimates a relatively stable, slightly declining path for Student 82, with a brief upward bump around week 10. Student 138, in contrast, shows more variability and potential mid-year recovery. These differences underscore D-BIRD's ability to distinguish between superficially similar learners by leveraging the full sequence of assessment interactions.

6 Discussion

This paper introduces D-BIRD, a Bayesian dynamic IRT model that decomposes student ability into a shared cohort trend and an individual-specific trajectory. This structure is designed to support an important goal of educational measurement in personalized learning environments: tracking individual growth over time while situating it within broader group-level patterns. By explicitly modeling both individual and cohort dynamics, D-BIRD enables interpretable inferences even under sparse, irregular assessment conditions—a common feature of modern digital learning systems.

D-BIRD combines two key ideas: structured borrowing across students and flexible modeling of individual change. The cohort trajectory provides a stable, data-driven reference against which individual deviations can be interpreted. Student-specific innovation variances allow each learner's ability to evolve with a level of smoothness appropriate to their observed responses. Exact Bayesian inference via Pólya-Gamma augmentation ensures well-calibrated posterior estimates, avoiding common approximations such as variational inference.

Several modeling choices limit the generalizability of D-BIRD and point to directions for future work. First, we assume item difficulties are known, consistent with operational settings that use pre-calibrated item pools. Future work could relax this assumption to jointly estimate item and ability parameters, exploring identifiability under sparsity. Second, D-BIRD is formulated in discrete time, where each time index may correspond to a learning opportunity (Koedinger et al., 2023), a day of instruction (Wang et al., 2013), or—as in our empirical application—a week. Extensions to continuous time, such as placing Gaussian process priors over latent ability (Tripathi and Domingue, 2019), could support finer-grained modeling of learning dynamics, particularly in irregular data streams. Third, D-BIRD currently models dichotomous responses using the Rasch model. A natural extension is to adapt the framework for polytomous item models (Ostini and Nering, 2006), enabling broader applicability to complex assessment formats.

More broadly, D-BIRD contributes to a growing body of work at the intersection of psychometrics and AI-driven learning systems. As adaptive platforms increasingly rely on real-time data to personalize instruction, there is a pressing need for interpretable models that capture both individual learning progress and broader cohort-level trends. D-BIRD helps meet this need by offering a principled approach to longitudinal ability estimation—balancing flexibility with structure, and individual adaptation with population-level insight. In doing so, it advances longstanding goals in educational measurement while aligning with the practical demands of emerging digital learning environments.

References

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of statistical software*, 76:1–32.

Andrew Gelman. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).

Matthew D. Hoffman and Andrew Gelman. 2014. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Kosuke Imai, James Lo, and Jonathan Olmsted. 2016. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656.

Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Yunsung Kim, Sreechan Sankaranarayanan, Chris Piech, and Candace Thille. 2023. Variational temporal irt: Fast, accurate, and explainable inference of dynamic learner proficiency. *arXiv preprint arXiv:2311.08594*.

Kenneth R Koedinger, Paulo F Carvalho, Ran Liu, and Elizabeth A McLaughlin. 2023. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, 120(13):e2221311120.

Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. 2019. On the geometric ergodicity of hamiltonian monte carlo. *Bernoulli*, 25(4A):3109–3138.

Andrew D Martin and Kevin M Quinn. 2002. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political analysis*, 10(2):134–153.

Radford M Neal and 1 others. 2011. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Remo Ostini and Michael L Nering. 2006. *Polytomous item response theory models*. 144. Sage.

Nicholas G Polson, James G Scott, and Jesse Windle. 2013. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

G. Rasch. 1980. *Probabilistic models for some intelligence and attainment tests*, expanded ed. edition. University of Chicago Press, Chicago.

Havard Rue. 2001. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338.

Szymon Sacher, Laura Battaglia, and Stephen Hansen. 2021. Hamiltonian Monte Carlo for Regression with High-Dimensional Categorical Data. Papers 2107.08112, arXiv.org.

Jingyu Sun, Yang Liu, Xiaojing Wang, and Ming-Hui Chen. 2025. Bayesian variable selection in dynamic item response theory models. *Journal of Educational and Behavioral Statistics*, page 10769986251314527.

Samuel Thomas and Wanzhu Tu. 2021. Learning hamiltonian monte carlo in r. *The American Statistician*, 75(4):403–413. PMID: 37465458.

Ajay Tripathi and Benjamin Domingue. 2019. Curve fitting from probabilistic emissions and applications to dynamic item response theory. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1336–1341.

Xiaojing Wang, James O. Berger, and Donald S. Burdick. 2013. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1):126–153.

Xin Wang and Vivekananda Roy. 2018. Geometric ergodicity of polya-gamma gibbs sampler for bayesian logistic regression with a flat prior. *Electronic Journal of Statistics*, 12.

Mike West and Jeff Harrison. 2006. *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Mike West, P Jeff Harrison, and Helio S Migon. 1985. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

Enhancing Essay Scoring with GPT-2 Using Back Translation Techniques

Aysegul Gunduz
University of Alberta
gunduz@ualberta.ca

Mark J. Gierl
University of Alberta
mark.gierl@ualberta.ca

Okan Bulut
University of Alberta
okan.bulut@ualberta.ca

Abstract

Advancements in artificial intelligence and transformer-based language models have significantly influenced educational assessment, particularly in the development of Automated Essay Scoring (AES) systems. This study examines the effectiveness of the GPT-2 small model in evaluating student essays from the Automated Student Assessment Prize (ASAP) dataset[1].It also explores the effect of a back-translation data augmentation technique(translating essays into Turkish and then back into English) On model performance. Evaluation metrics include Cohen's kappa and Quadratic Weighted Kappa (QWK). The model achieved QWK scores ranging from 0.60 to 0.80 across essay sets, with a peak of 0.77 on Essay Set 5. Notably, back translation led to substantial improvements, particularly in Essay Set 8, where QWK increased by 33%. These findings highlight the potential of data augmentation to mitigate class imbalance and improve scoring robustness. However, the limited semantic depth of the GPT-2 small model points to the need for more advanced, rubric-aware architectures. The study underscores the importance of balanced data distributions in enhancing the validity and fairness of AES systems.

Keywords: *artificial intelligence, language modeling, automated essay scoring (AES), GPT-based models, GPT-2*

1 Introduction

Recent advances in large language models (LLMs), particularly those developed under the Generative Pretrained Transformer (GPT) architecture, have significantly influenced Automated Essay Scoring (AES). Early AES systems relied on surface-level linguistic features and traditional machine learning algorithms (Kumar and Boulanger, 2020; Klebanov and Madnani, 2022), while more recent approaches have incorporated transformer-based models capable of capturing deeper semantic and syntactic patterns (Taghipour and Ng, 2016). Among these, encoder-only architectures such as BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) have been widely applied to AES tasks, achieving strong performance and serving as reliable baselines (Firoozi et al., 2023; Wang et al., 2022).

In contrast, decoder-based generative models, particularly GPT variants, have received comparatively limited attention in AES despite their proven success in other natural language processing applications. Recent studies have demonstrated that advanced generative models such as GPT-3.5 and GPT-4 can achieve near human-level performance in essay scoring benchmarks (Mizumoto and Eguchi, 2023; Xiao et al., 2025; Yamashita, 2025). However, these models are proprietary and resource-intensive, limiting their accessibility for educational researchers and practitioners. Smaller, open-source alternatives like GPT-2 remain underexplored in AES, even though evidence from related classification tasks indicates that fine-tuned GPT-2 can rival or surpass BERT-based hybrids in performance (Bouchiha et al., 2025). This observation supports the need for systematic evaluation of GPT-2 in AES, both as a practical and a methodological contribution.

Another persistent challenge in AES is the limited size and imbalance of available datasets, which can compromise model generalization and fairness (Jong et al., 2022; Guo et al., 2024). Data augmentation techniques such as back-translation have been proposed as a potential solution, offering greater linguistic diversity and reducing the effects of class imbalance (Lun et al., 2020). Yet, it remains unclear whether performance improvements attributed to back-translation derive from genuine linguistic variation or simply from the increased size of training data. Addressing this ambiguity requires careful experimental design that controls

[1] https://www.kaggle.com/c/asap-aes

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 406–416
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

for training size and duplication.

This study makes two key contributions. First, it provides a controlled evaluation of GPT-2 for AES on the ASAP dataset, positioning it against established encoder-based baselines and recent parameter-efficient fine-tuning approaches such as LoRA (Liu et al., 2024). Second, it investigates the effect of back-translation as a data augmentation strategy under controlled conditions, clarifying whether observed improvements stem from data diversity rather than dataset expansion alone.

The study is guided by two research questions:

1. How reliably can a fine-tuned GPT-2 model score essays from the ASAP dataset compared to established baselines?
2. To what extent does back-translation improve GPT-2's AES performance beyond the effect of increasing training set size?

2 Related Work

2.1 Overview of Automated Essay Scoring

Automated Essay Scoring (AES) refers to the use of computational methods to evaluate and score student essays (Shermis, 2014). While manual scoring is often time-consuming and prone to rater inconsistency, AES offers efficiency, objectivity, and scalability, making it an increasingly valuable tool in educational contexts (Yan et al., 2020).

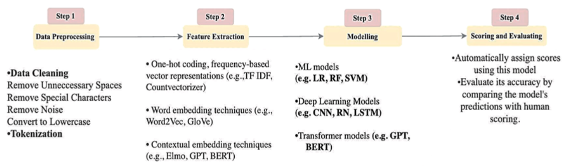


Figure 1: The AES Process described in Four Steps (Gierl et al., 2014).

As can be seen in Figure 1, the AES process consists of four steps: text preprocessing, feature extraction, model training, and performance evaluation (Gierl et al., 2014). To detail, it involves the preprocessing (Step 1) and conversion of essays written in a training environment into numerical vectors using text representation techniques (Step 2), combining these vectors with machine learning algorithms or deep learning networks to create a scoring model (Step 3), and automatically assigning scores using this model and evaluating the scoring model to see if it can predict human scoring (Step 4). Advances in machine learning and natural language processing have significantly improved the first three stages, particularly through enhanced text representation and modeling techniques (**?**).

Early feature extraction methods employed frequency-based techniques, such as term frequency (TF) and TF-IDF (Salton et al., 1975), but these approaches were unable to capture semantic meaning. Later, word embedding models like Word2Vec and GloVe (Mikolov et al., 2013) improved semantic representation but were still context-independent. Contextual embedding models such as ELMo, BERT, and GPT addressed this limitation by incorporating surrounding context into each word's representation (Peters et al., 2018; Radford et al., 2018; Liu et al., 2020). These advances improved the quality of input features used in scoring models.

Earlier AES studies employed deep neural networks (DNNs) and recurrent neural networks (RNNs) to model sequential patterns in text (Alikaniotis et al., 2016; Tay et al., 2018). RNNs often struggle with capturing long-range dependencies and information across time steps; however, they are designed to suit the sequence-to-sequence design effectively (Nugaliyadde et al., 2019). This limitation has motivated the use of transformer-based architectures, which replace recurrence with attention mechanisms. The self-attention mechanism introduced by Vaswani et al. (2017) enables the model to learn dependencies across all positions in a sequence simultaneously, resulting in a richer representation of global structure and semantic relationships. As a result, transformer models such as GPT have become increasingly prevalent in recent AES research.

2.2 Transformer-Based Architectures in AES

In 2017, the paper 'Attention Is All You Need' revolutionized the field of natural language processing (NLP) by introducing the transformer architecture (see Figure 2) (Vaswani et al., 2017). This model leveraged self-attention mechanisms to capture long-range semantic and syntactic dependencies in text. In AES tasks, encoder-only transformers such as BERT (Devlin et al., 2019) and RoBERTa have also demonstrated state-of-the-art performance in both predictive and analytic scoring (Firoozi et al., 2023; Klebanov and Madnani, 2022).

These models provide robust contextual embeddings and strong baselines for AES. These models are typically fine-tuned on prompt-specific essay datasets, where only the top classification layer is updated while the encoder layers provide contextualized embeddings. This parameter-efficient strategy has proven effective in score prediction, especially under constrained computational resources. However, their reliance on bidirectional masked language modeling may limit their utility in generative tasks and document-level coherence modeling. Although rubric-integrated encoder architectures have improved interpretability and alignment with human scoring rubrics (Liu et al., 2020), their generalization across unseen prompts and diverse discourse structures remains limited.

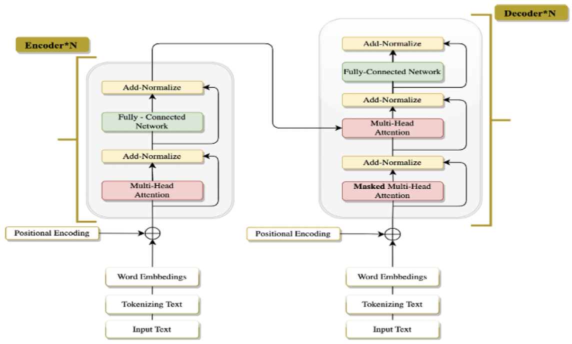


Figure 2: Transformer Architecture (Vaswani et al., 2017).

These limitations have motivated research on decoder-based models, which are inherently more suited to sequence-level generation and whole-document representation. Importantly, GPT-2 has not only been effective in generative tasks but has also demonstrated competitive performance in classification settings. For instance, GPT-2 has matched or even surpassed BERT-based hybrids in hierarchical text classification (Bouchiha et al., 2025), performed strongly in text classification and natural language inference benchmarks (Montesinos, 2020), and shown competitive results against BERT in low-resource classification tasks (Wang et al., 2024). Such findings indicate that GPT-2 is a viable model for AES, where both classification accuracy and generative capabilities are critical.

2.3 Decoder-Only Transformers: The GPT Family

Decoder-only models, particularly the GPT series introduced by OpenAI, are trained with autoregressive objectives and unidirectional attention, which makes them inherently generative (Radford et al., 2018, 2019). GPT-2 expanded this architecture to 1.5 billion parameters and demonstrated strong transfer performance. Subsequent models such as GPT-3, GPT-3.5, and GPT-4 further scaled capacity and achieved near human-level accuracy in AES benchmarks under zero-shot and few-shot prompting conditions (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Gunduz and Gierl, 2024). GPT-4 also introduced multimodal input processing, although its architecture and training data remain undisclosed

While these larger models have shown impressive results, their proprietary nature restricts reproducibility and accessibility. In contrast, GPT-2 remains fully open-source and scalable across different sizes, making it a practical option for academic and educational research. Importantly, GPT-2 has proven effective beyond generative applications. Prior studies have shown that fine-tuned GPT-2 can outperform BERT-based hybrids in hierarchical text classification (Bouchiha et al., 2025), perform strongly in text classification and natural language inference benchmarks (Montesinos, 2020), and achieve competitive results against BERT in low-resource classification tasks (Wang et al., 2024).

These findings indicate that GPT-2 is not only cost-efficient and accessible but also capable of delivering robust performance in classification-oriented tasks. Nevertheless, systematic evaluations of GPT-2 on AES benchmark datasets such as ASAP remain limited. This study addresses this gap by providing a controlled and reproducible assessment of GPT-2 for essay scoring, with particular attention to the role of data augmentation.

2.4 Recent Applications of GPT-Based AES

In recent years, GPT models have been increasingly applied to educational assessment tasks, including both short-answer and essay scoring. One line of research has focused on data augmentation to mitigate class imbalance.Fang et al. (2023) employed GPT-4 to generate synthetic responses for minority scoring classes, which improved the performance of a DistilBERT scoring model. Similarly, Gaddipati et al. (2020) compared transfer learning models such as ELMo, BERT, GPT, and GPT-2 for short-answer grading, showing that while ELMo provided strong baselines, transformer-based models offered greater scalability for downstream use. Several studies have investigated larger GPT models for direct scoring. Mizumoto and Eguchi (2023) evaluated GPT-3 on TOEFL essays and reported that combining linguistic features with

Model	Architecture	Parameters	Training Data	Release Date
GPT-1	12H Decoder	117M	BookCorpus	2018
GPT-2	12–48H Decoder	1.5B	WebText	2019
GPT-3	Modified GPT-2	175B	CC + WebText	2020
GPT-3.5	Undisclosed	175B	–	Mar 2022
GPT-4	Undisclosed	$\sim$1.7T	–	Mar 2023

Table 1: Overview of OpenAI's GPT-n series.

Model	Params	Layers	Hidden	Input
Small	117M	12	768	768
Medium	345M	24	1024	1024
Large	774M	36	1280	1280
XL	1558M	48	1600	1600

Table 2: GPT-2 Model Configurations across Four Sizes.

model outputs improved agreement with human raters.Yancey et al. (2023) assessed GPT-3.5 and GPT-4 on essays from English language learners, finding that GPT-4 achieved performance comparable to state-of-the-art Automated Writing Evaluation (AWE) systems, though alignment varied by learners' first language. Henkel et al. (2023) used GPT-4 for scoring short-answer reading comprehension tasks in low- and middle-income countries, demonstrating its potential in resource-limited educational contexts.Obata et al. (2023) tested ChatGPT for essay scoring in English and Japanese and showed that validity improved when combined with linguistic features.Xiao et al. (2025) further argued that GPT-3.5 and GPT-4 are most effective when augmenting human raters in hybrid scoring systems.

Despite these advances, most work has concentrated on proprietary models such as GPT-3.5 and GPT-4, limiting reproducibility and transparency. Benchmark studies on open-source models remain scarce. Gunduz and Gierl (2024) compared GPT-3.5 and GPT-4 under different prompting conditions on the ASAP dataset, but no systematic evaluation of GPT-2 has yet been conducted. Considering GPT-2's accessibility, scalability, and demonstrated competitiveness in classification tasks (Bouchiha et al., 2025; Wang et al., 2024), further investigation is warranted. This study addresses this gap by fine-tuning GPT-2 on the ASAP dataset and evaluating the effects of back-translation as a data augmentation strategy, offering a reproducible and transparent alternative to proprietary systems.

2.5 Data Augmentation and Back-Translation in AES

Data augmentation is widely used in NLP to improve generalization and mitigate label imbalance through techniques such as synonym replacement, paraphrasing, and translation-based methods (Wei and Zou, 2019). In AES, augmentation helps balance score distributions and enrich training data diversity (Lun et al., 2020; Jong et al., 2022).

Back-translation, which generates paraphrases by translating text into a target language and back, has been shown to increase linguistic variety and robustness in low-resource tasks (Sennrich et al., 2016; Edunov et al., 2018). In AES, augmentation methods have been applied to enrich training data (e.g., (Firoozi, 2023; Guo et al., 2024)), yet the specific impact of back-translation on score distributions, particularly under imbalanced data conditions, remains underexplored. This study addresses this gap by applying back-translation to the ASAP dataset under controlled conditions, clarifying its contribution beyond simple dataset expansion.

3 Method

3.1 Dataset

This study utilizes the Automated Student Assessment Prize (ASAP) dataset, developed under the sponsorship of the Hewlett Foundation in 2012, to encourage scalable and reliable approaches to AES (Shermis, 2014). The dataset comprises eight distinct essay sets written by students in Grades 7 through 10, encompassing various genres, including narrative, persuasive, and expository writing. Each essay set varies in terms of grade level, rubric type, essay length, and scoring range (see Table 3). Essays were scored by two or three expert raters using holistic, trait-based, or composite rubrics. The score ranges and aggregation methods for domain scores differ across sets. Table 3 summarizes the specific score ranges for Rater 1, Rater 2, and the derived domain score used for model training. Among the essay sets, Set 4 stands out for its relatively balanced score distribution across all score categories. As shown in Table 3, both individual rater scores and the domain score span the full range from 0 to 3, with sufficient representation in each category. This balanced distribution is particularly beneficial for training reliable AES models, as it reduces the risk of class imbalance and supports more effective learning dynamics.

3.2 Data Preprocessing

Text Preprocessing. To prepare the essays for model input, standard text preprocessing steps were applied. All texts were lowercased and lemmatized using the NLTK library (Bird et al., 2009). The cleaned essays were then tokenized using the GPT-2 tokenizer from the Hugging Face Transformers library (Wolf et al., 2019). Since transformer models require fixed-length input, padding and truncation were used to standardize sequence lengths.

Score Preprocessing. Each essay was scored by two or three raters, and domain scores were computed according to the scoring rules in Table 3. However, some sets (Essay Sets 1, 7, and 8) had wide or unbalanced score ranges. To improve model performance, these scores were rescaled into fewer ordinal categories. For instance, Set 1 domain scores (2–12) were converted to a 1–6 ordinal scale. Similarly, Set 7 scores (0–24) were mapped to a 0–3 scale, and Set 8 scores (0–60) were compressed into six ordinal categories based on trait aggregation logic (see Table 3).

3.3 Model Development

GPT-2 Architecture. The model developed in this study builds upon the GPT-2 architecture (see Figure 3), a decoder-only transformer pretrained on over 8 million web pages (Radford et al., 2019). GPT-2 generates contextualized word embeddings using masked self-attention and is optimized for predicting the next token. Among its four variants, the smallest version—GPT-2 Small (124M parameters, 12 decoder layers, 768 hidden units)—was selected due to computational efficiency. All training was conducted using Google Colab Pro (Tesla V100 GPU, 32GB RAM). Each decoder block in

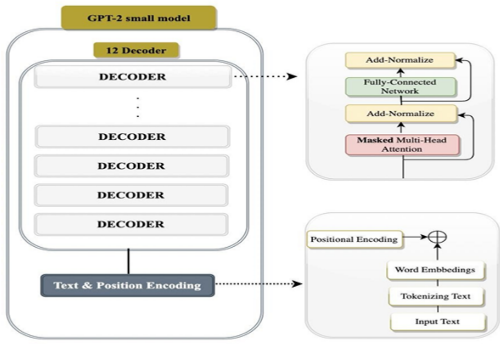


Figure 3: GPT-2 Small Architecture.

GPT-2 includes masked multi-head self-attention, a feedforward network, residual connections, and layer normalization (Vaswani et al., 2017). The input sequence is processed from left to right, making the model suitable for both generative and classification tasks.

Classification Head. To adapt GPT-2 for AES, we added a task-specific classification head on top of the pretrained transformer (see Figure 4). This consisted of a dropout layer (with rates of 0.1, 0.2, and 0.5 tested) followed by a fully connected linear layer that maps the last hidden state of the model to a fixed number of score classes per essay set (e.g., 4 classes in Essay Set 4). The `num_labels` parameter was dynamically set based on the scoring range of each set. All training was performed on Google Colab Pro using the Hugging Face Transformers library (Wolf et al., 2019). The small-scale GPT-2 variant enabled faster iteration while maintaining competitive performance for AES tasks.

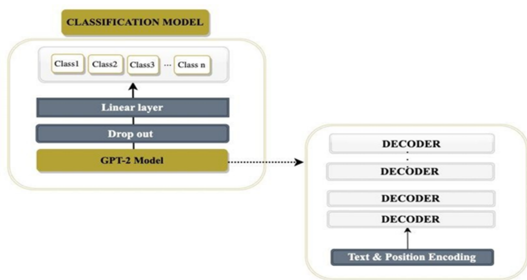


Figure 4: Classification Model Architecture.

3.4 Experimental Setup and Hyperparameter Tuning

All essays were tokenized using the GPT-2 tokenizer from the HuggingFace Transformers library.

Set	Grade	Essay Type	Train Size	Avg. Len.	Rubric Type	Raters	Score Range	Domain Score Explanation
1	8	Persuasive	1783	350	Holistic	2	2–12	Sum of R1 and R2 (2–12)
2a	10	Persuasive	1800	50	Trait	2	1–6	Equals R1's score (1–6)
2b	10	Persuasive	1800	50	Trait	2	1–4	Equals R1's score (1–4)
3	10	Source-Dep.	1726	50	Holistic	2	0–3	Max(R1, R2) (0–3)
4	10	Source-Dep.	1772	50	Holistic	2	0–3	Near max(R1, R2) (0–3)
5	8	Source-Dep.	1805	50	Holistic	2	0–4	Near max(R1, R2) (0–4)
6	10	Source-Dep.	1800	50	Holistic	2	0–4	Near max(R1, R2) (0–4)
7	7	Expository	1569	50	Composite	2	0–12	Sum of R1 and R2 (0–24)
8	10	Expository	723	50	Composite	3	0–30	R1+R2 or R3 used (0–60)

Table 3: Descriptive Statistics and Scoring Guidelines for the Eight ASAP Essay Sets.

Essays exceeding the maximum sequence length of 1,024 tokens (as imposed by the GPT-2 Small architecture) were truncated.

The pre-trained GPT-2 Small model was initialized with default configurations: 12 decoder layers, 768-dimensional hidden states and embeddings, 12 self-attention heads, GELU activation, and dropout probability of 0.1 across embedding, attention, and fully connected layers. Layer normalization used an epsilon value of 1e-5. In total, the model contains approximately 117M parameters. To adapt GPT-2 for essay scoring, a linear classification head with dropout was appended. The number of output classes was defined per essay set.

To adapt GPT-2 for essay scoring, a linear classification head with dropout was appended. The number of output classes was defined per essay set using the `num_labels` parameter. The final hidden state of the first token was passed to the classification layer. Model training was optimized using the AdamW optimizer (Kingma, 2014) with a fixed learning rate of 1e-4 and categorical cross-entropy loss. The loss function for k classes is defined in Equation 1

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \tag{1}$$

where y is the one-hot true label and $\hat{y}$ is the predicted class distribution.

To ensure robust evaluation, each essay set was randomly partitioned into training (60%), validation (20%), and test (20%) subsets following standard practice.

3.5 Data Augmentation Strategy

The distribution of essays across score levels is utilized by the GPT-2 Small architecture, which features the performance and generalizability of AES models. To address this, we employed data augmentation to enhance the training set, particularly for underrepresented classes.

Text Augmentation. Text data augmentation involves generating additional samples by modifying existing texts, without requiring new data collection. Common methods include synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD), which introduce lexical variability while preserving sentence structure (Firoozi, 2023).

Back-Translation. Among these methods, back-translation has emerged as a particularly effective strategy for producing fluent and semantically consistent variations. This technique translates a sentence into an intermediate language and back to the original, creating paraphrased versions that enrich the training data. In our study, the source language was English, and the target language was Turkish. We translated English essays into Turkish and then back into English using the Google Translate API. Turkish was intentionally chosen as the pivot language due to its agglutinative morphology and syntactic divergence from English, contributing to greater linguistic variety in the augmented texts.

This method was selectively applied: in balanced sets (e.g., Set 4), each score class was augmented by 20% following the strategy proposed in Firoozi's Doctoral Thesis (Firoozi, 2023), while in imbalanced sets, score levels with fewer than 50 samples were doubled. This targeted approach aimed to reduce class imbalance, minimize model bias, and improve performance across the entire score spectrum. The augmentation process is illustrated in Figure 5.

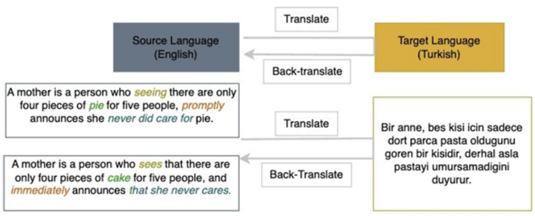


Figure 5: Back-Translation Data Augmentation Pipeline.

3.6 Performance Metrics

To evaluate the effectiveness of the AES model, we employed multiple metrics capturing both agreement with human raters and classification performance.

Cohen's Kappa. Cohen's Kappa (κ) measures inter-rater agreement corrected for chance, and is commonly used to assess the consistency between model predictions and human scores. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

where P_o is the observed agreement and P_e is the expected agreement by chance. Agreement levels are interpreted based on the guidelines by Landis and Koch (1977).

Quadratic Weighted Kappa (QWK). QWK extends Cohen's Kappa by penalizing disagreements based on the distance between score levels, making it especially suitable for ordinal tasks, such asutilized the GPT-2 Small architecture, featurings:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{3}$$

The QWK score is then defined by:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{4}$$

Accuracy. Accuracy reflects the proportion of essays for which the predicted score exactly matches the human-assigned score. Although it does not account for ordinal distance between misclassified levels, it remains a useful baseline metric for evaluating overall classification correctness.

4 Results

4.1 Hyperparameter Settings

To optimize performance, we fine-tuned key hyperparameters for each essay set, as detailed in Table 4. All models used the GPT-2 Small architecture with 768-dimensional token embeddings, 1024-dimensional positional encodings, and 12 decoder layers.

A classification head consisting of a dropout and a linear layer was appended to map outputs to a variable number of score classes (`num_labels`) per essay set. Dropout rates were adjusted individually; Sets 1 and 2a performed best with 0.5, while 2b, 3, 5, 7, and 8 performed best with 0.1.

A fixed learning rate of 1e-4 was used across sets, optimized via the `AdamW` optimizer. Epochs and batch size varied by set, reflecting differences in convergence behavior and dataset size. For example, Set 6 performed best with 30 epochs, dropout 0.3, and batch size 2.

4.2 RQ1: AES Model Performance

The fine-tuned GPT-2 model demonstrated moderate scoring reliability across the eight essay sets. On average, it achieved a Quadratic Weighted Kappa (QWK) of 0.68, Cohen's Kappa of 0.43, and classification accuracy of 61%. According to the interpretability thresholds proposed by Williamson et al. (2012), the model explained a substantial portion of human scoring variance. These results suggest that even the most minor GPT-2 variant can offer competitive performance in AES tasks under computational constraints.

4.3 RQ2: Effect of Data Augmentation

Back-translation-based data augmentation led to notable performance improvements. The average QWK score increased from 0.68 to 0.74 (+0.06), while Cohen's Kappa rose from 0.43 to 0.48 (+0.05). This gain was most evident in essay sets with initially imbalanced score distributions, confirming the effectiveness of targeted augmentation in enhancing agreement between machine predictions and human raters.

5 Discussion

This study examined the performance of the open-source GPT-2 model for AES on the ASAP dataset, with a focus on fine-tuning and back-translation-based data augmentation. Results show that even the most minor GPT-2 variant, when fine-tuned with optimized hyperparameters, achieved a competitive average QWK score of 0.68 (close to human-level performance at 0.74) and outperformed GPT-3.5 in certain sets.The model performed best in balanced essay sets with sufficient

Parameter	**1**	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**
Embedding Dim.	768	768	768	768	768	768	768	768	768
Positional Encoding	1024	1024	1024	1024	1024	1024	1024	1024	1024
Decoder Layers	12	12	12	12	12	12	12	12	12
Num Labels	6	6	4	4	4	5	5	4	6
Dropout Rate	0.5	0.1	0.5	0.2	0.2	0.1	0.1	0.3	0.1
Learning Rate	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Epochs	20	25	35	20	30	20	30	20	30
Batch Size	2	2	2	4	2	2	2	1	2

Table 4: Final selection of Hyperparameters used for Fine-tuning GPT-2 across all Essay Sets.

Model	**1**	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
GPT-2	0.75	0.64	0.66	0.71	0.74	0.77	0.73	0.77	0.43	0.68
Human Raters	0.71	0.78	0.72	0.81	0.86	0.74	0.77	0.68	0.63	0.74
Discrepancy	0.04	0.14	0.06	0.10	0.12	0.03	-0.04	0.09	0.18	0.06

Table 5: Comparison of GPT-2 and Human Raters using QWK across all Essay Sets.

Essay Set	**1**	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**
Before BT	1783	1800	1800	1726	1772	1805	1800	1569	723
After BT	1875	1831	1829	1765	2124	1829	1844	1676	1220
Discrepancy	+92	+31	+29	+39	+352	+24	+44	+107	+497

Table 6: Comparison of Training Dataset Size Before and After Back-Translation for each Essay Set.

Model	**Performance**	**1**	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
GPT-2	Cohen's Kappa	0.52	0.46	0.45	0.41	0.43	0.45	0.40	0.41	0.31	0.43
	QWK	0.75	0.64	0.66	0.71	0.74	0.77	0.73	0.71	0.45	0.68
	Accuracy	0.67	0.61	0.63	0.61	0.60	0.66	0.58	0.60	0.53	0.61
GPT-2 + BT	Cohen's Kappa	0.54	0.50	0.53	0.47	0.48	0.47	0.43	0.49	0.38	0.48
	QWK	0.79	0.65	0.68	0.77	0.81	0.79	0.79	0.74	0.60	0.74
	Accuracy	0.72	0.62	0.74	0.62	0.68	0.64	0.67	0.67	0.59	0.66
Discrepancy	Cohen's Kappa	+0.02	+0.04	+0.08	+0.06	+0.05	+0.03	+0.01	+0.08	+0.11	+0.05
	QWK	+0.04	+0.01	+0.02	+0.06	+0.06	+0.02	+0.06	+0.03	+0.15	+0.06
	Accuracy	+0.05	+0.01	+0.11	+0.01	+0.09	+0.02	+0.06	+0.07	+0.06	+0.05

Table 7: Comparison of GPT-2 Model Performance Before and After Back-Translation (BT) across all Essay Sets.

training data, while lower reliability was observed in sparse or imbalanced sets such as Set 8. Data augmentation proved particularly effective for underrepresented score classes, improving both QWK and Cohen's Kappa scores and reducing class imbalance. These findings affirm the value of tuning smaller, accessible models for educational NLP tasks, highlighting the trade-off between model complexity and interpretability in low-resource contexts.

In conclusion, GPT-2, despite its smaller architecture, offers substantial potential for AES when carefully fine-tuned and supported by data augmentation. Its open-source nature and customizable hyperparameters make it a practical choice for scalable, interpretable assessment systems. Back-translation significantly improved performance in low-resource score categories, demonstrating its value in addressing data sparsity. These results reinforce that high-quality AES systems can be developed without relying solely on larger proprietary models, and suggest future directions in combining linguistic measures and augmentation techniques to enhance model robustness and fairness.

6 Limitations and Future Work

Despite promising results, this study has several limitations. First, it focuses exclusively on classification-based essay scoring and does not incorporate rubric-specific features, which are central to many human scoring protocols. The absence of rubric-aligned modeling limits interpretability and may hinder the effectiveness of feedback-oriented applications. Second, the dataset includes essay sets with imbalanced score distributions and small sample sizes, which may constrain generalizability, particularly in underrepresented categories. Third, experiments were limited to the GPT-2 Small model; while acceptable, tuning significantly improved performance, larger models (e.g., GPT-3, GPT-4) could better capture complex linguistic structures if similarly fine-tuned. Lastly, only one augmentation strategy—back-translation—was explored. Future work should investigate rubric-aware scoring frameworks, incorporate alternative augmentation methods (e.g., synonym substitution, sentence permutation), and evaluate larger-scale models on more balanced datasets to improve the robustness and educational utility of AES systems.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Djelloul Bouchiha, Abdelghani Bouziane, Noureddine Doumi, Benamar Hamzaoui, and Sofiane Boukli-Hacene. 2025. Hierarchical text classification: Fine-tuned gpt-2 vs bert-bilstm. *Applied Computer Systems*, 30(1):40–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *arXiv preprint arXiv:2310.18365*.

Tahereh Firoozi. 2023. Using automated procedures to score written essays in persian: An application of the multilingual bert system.

Tahereh Firoozi, Okan Bulut, and Mark Gierl. 2023. Language models in automated essay scoring: Insights for the turkish language. *International Journal of Assessment Tools in Education*, 10(Special Issue):149–163.

Sasi Kiran Gaddipati, Deebul Nair, and Paul G Plöger. 2020. Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv preprint arXiv:2009.01303*.

Mark J Gierl, Syed Latifi, Hollis Lai, André-Philippe Boulais, and André De Champlain. 2014. Automated essay scoring and the future of educational assessment in medical education. *Medical education*, 48(10):950–962.

Aysegul Gunduz and M Gierl. 2024. Automated essay scoring with chatgpt 3.5 and 4.0. Presentation at the UBlberta Graduate Student Research in Education Conference.

Weiqin Guo, Yong Yang, and Ge Ren. 2024. Research of automatic scoring of essays based on data augmentation. In *Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum*, pages 635–641.

Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. 2023. Can llms grade short-answer reading comprehension questions: An empirical study with a novel dataset. *arXiv preprint arXiv:2310.18373*.

You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. 2022. Improving performance of automated essay scoring by using back-translation essays and adjusted scores. *Mathematical Problems in Engineering*, 2022(1):6906587.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. 2024. ALoRA: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 622–641, Mexico City, Mexico. Association for Computational Linguistics.

Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13389–13396.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Dimas Munoz Montesinos. 2020. Modern methods for text generation. *arXiv preprint arXiv:2009.04968*.

Anupiya Nugaliyadde, Upeka Somaratne, and Kok Wai Wong. 2019. Predicting electricity consumption using deep recurrent neural networks. *arXiv preprint arXiv:1909.08182*.

Ayaka Obata, Takumi Tagawa, and Yuichi Ono. 2023. Assessment of chatgpt's validity in scoring essays by foreign language learners of japanese and english. In *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pages 105–110. IEEE.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Gerard Salton, Chung-Shu Yang, and Clement T Yu. 1975. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS 2019)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Mark D Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Recurrently controlled recurrent networks. *Advances in neural information processing systems*, 31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Yu Wang, Wen Qu, and Xin Ye. 2024. Selecting between bert and gpt for text classification in political science research. *arXiv preprint*, arXiv:2411.05050.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 293–305.

Taichi Yamashita. 2025. Exploring potential biases in gpt-4o's ratings of english language learners' essays. *Language Testing*, 42(3):344–358.

Duanli Yan, André A Rupp, and Peter W Foltz. 2020. *Handbook of automated scoring: Theory into practice*. CRC Press.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Mathematical Computation and Reasoning Errors by Large Language Models

Liang Zhang
Institute for Intelligent Systems
University of Memphis, Memphis, TN, USA
AI4STEM Education Center
University of Georgia, Athens, GA, USA
liangzhang@uga.edu

Edith Aurora Graf *

Lawrenceville, NJ, USA
eag2718@gmail.com

Abstract

Large Language Models (LLMs) are increasingly utilized in AI-driven educational instruction and assessment, particularly within mathematics education. The capability of LLMs to generate accurate answers and detailed solutions for math problem-solving tasks is foundational for ensuring reliable and precise feedback and assessment in math education practices. Our study focuses on evaluating the accuracy of four LLMs (OpenAI GPT-4o and o1, DeepSeek-V3 and DeepSeek-R1) solving three categories of math tasks, including arithmetic, algebra, and number theory, and identifies step-level reasoning errors within their solutions. Instead of relying on standard benchmarks, we intentionally build math tasks (via item models) that are challenging for LLMs and prone to errors. The accuracy of final answers and the presence of errors in individual solution steps were systematically analyzed and coded. Both single-agent and dual-agent configurations were tested. It is observed that the reasoning-enhanced OpenAI o1 model consistently achieved higher or nearly perfect accuracy across all three math task categories. Analysis of errors revealed that procedural slips were the most frequent and significantly impacted overall performance, while conceptual misunderstandings were less frequent. Deploying dual-agent configurations substantially improved overall performance. These findings offer actionable insights into enhancing LLM performance and underscore effective strategies for integrating LLMs into mathematics education, thereby advancing AI-driven instructional practices and assessment precision.

1 Introduction

Large Language Models (LLMs) have significantly impacted mathematical instruction and assessment. Educational platforms are increasingly integrating LLMs to enhance teaching and evaluation methods. For instance, Khan Academy utilized the LLM-powered tool *Khanmigo* for Socratic-style math assistance (Anand, 2023). Coursera uses LLMs to streamline assessment creation, automate grading, and offer personalized feedback (Maggioncalda, 2024). Quizlet's *Q-Chat* integrates LLM-based conversational AI to dynamically adapt question difficulty levels and deliver guided hints (Bayer, 2025). Beyond merely producing final answers, LLMs excel at clearly articulating intermediate computational steps and reasoning processes, significantly enhancing their value in mathematics education contexts (Gupta et al., 2025). These capabilities of LLMs facilitate personalized tutoring, interactive problem-solving, and real-time feedback, significantly reduce grading workloads and ensure consistent evaluations in mathematics education.

The capability of LLMs to produce accurate answers and detailed, step-by-step solutions in math problem-solving is foundational for reliable assessment and precise feedback in mathematics education (Gupta et al., 2025; Jin et al., 2025). Specifically, LLM-based automated assessment involves evaluating granular math skills through step-level grading of student solutions (Jin et al., 2025), performing automatic step-level corrections (Li et al., 2025), and providing targeted instructional hints (Tonga et al., 2025). However, this raises an important question: if LLMs cannot reliably produce correct answers or accurately solve math problems, can their outputs still be considered effective and trustworthy for instructional guidance and learner assessment? This motivated our idea to systematically test the capability of LLMs to accurately solve diverse math tasks, and subsequently extend their applicability toward realistic assessment and instructional scenarios. While state-of-the-art LLMs have demonstrated high accuracy on various math benchmarks, benchmark success alone does not present a comprehensive picture. Performance sig-

*This research was conducted while the second author was at Educational Testing Service.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 417–424
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

nificantly declines on certain math tasks like fundamental numerical understanding and basic computational problems (Yang et al., 2024; Boye and Moell, 2025; Petrov et al., 2025). Current LLM computation and reasoning processes remain prone to calculation mistakes (e.g., arithmetic slips, algebraic simplification errors) and logical reasoning errors (e.g., invalid inference steps, omission of necessary procedural steps, and self-contradictory reasoning) (Li et al., 2024; Roy et al., 2025). These persistent errors significantly limit the reliability and efficacy of LLM outputs for instructional feedback and learner assessment purposes.

In this study, we specifically examine the capabilities and limitations of LLMs in math problem-solving, focusing on assessing the accuracy of generated answers and identifying errors within solution steps. Instead of using benchmarks, we built math problems in arithmetic, algebra, and number theory to evaluate LLMs' proficieny in math computation and reasoning. We explored four distinct LLMs, two base models GPT-4o (Hurst et al., 2024) and DeepSeek-V3 (Liu et al., 2024) and two reasoning-enhanced models OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), across three math problem-solving tasks, including arithmetic, algebra, and number theory. Two interaction paradigms are considered: (i) a single-agent setting in which one model works through each math task step-by-step, and (ii) a dual-agent setting that lets two peer LLMs chat, cross-validate, and refine their reasoning, echoing recent advances in collaborative intelligence (Zhang et al., 2025a; Latif et al., 2024; Zhang et al., 2025b). Every solution was decomposed into granular steps and coded with an expert-verified rubric, which enabled us to quantify step-level accuracy and localize procedural or conceptual errors. Our investigation is guided by two **R**esearch **Q**uestions: **Q1:** How accurately do LLMs generate final answers to math problems? **Q2:** What recurring error patterns (procedural, conceptual, or logical) emerge in their step-level solutions?

This study will provide researchers and practitioners with precise insight into where LLMs excel and where they falter in math computation and reasoning. We also provide a rubric that can be used to evaluate the accuracy of LLM-based solutions, and to identify the nature of errors when they occur. Our study demonstrates that knowledgeable LLMs have potential to reliably support math instruction and assessment, and we offer actionable guidance for their effective use.

2 Related Works

LLMs have made notable strides in solving math problems, yet they frequently struggle with precise computations and multi-step numerical reasoning tasks (Wolfram, 2023; Li et al., 2024). Frieder et al. (Frieder et al., 2023) provided a detailed evaluation demonstrating that GPT-4 (the 2023 version of ChatGPT) effectively handles many undergraduate-level questions but exhibits significant difficulties when confronted with graduate-level math challenges, particularly in proof-based tasks and complex symbolic computations. Yang et al. (Yang et al., 2024) found that LLMs frequently make surprising mistakes in basic numerical understanding and processing tasks. Some studies (Li et al., 2024; Pan et al., 2025) have identified recurring error patterns in math reasoning by LLMs, including calculation errors, counting errors, formula confusion, question misinterpretation, missing solution steps, conceptual confusion, and nonsensical outputs, among others. Numerous additional studies have consistently reported similar challenges, emphasizing the ongoing limitations faced by LLMs in performing math tasks (Arkoudas, 2023; Wolfram, 2023; Wang et al., 2023; Zhang et al., 2024; McLeish et al., 2024).

Mitigating math problem-solving errors and boosting LLM performance is a multifaceted endeavor. Technical advances center on modular reasoning strategies, such as Chain-of-Thought and Program-of-Thought prompting (Wei et al., 2022; Chen et al., 2022), alongside fine-tuning and math-specific training regimens (Zhang et al., 2023; Ahn et al., 2024), novel architectures that integrate external tools or structured reasoning modules, and rigorous evaluations that precisely expose weaknesses. On the usability side, carefully designed prompts, curriculum-aligned task sequencing, interactive dialogue, and built-in self-checking routines can substantially reduce errors in real-world use.

3 Methods

Dataset. For the dataset, we used the problem categories and instances developed for Graf et al. (2025). Three distinct types of math tasks were utilized to evaluate the performance of LLMs: (1) multiplying two 5-digit numbers; (2) solving algebraic word problems involving quadratic equations; and (3) finding solutions to

Diophantine equations. See the math dataset in the GitHub repository: https://github.com/LiangZhang2017/math_number_computing. In our approach, we leveraged item models (Bejar, 2002; LaDuca et al., 1986) and automated item generation (AIG) (Embretson, 1999; Gierl and Haladyna, 2013; Irvine and Kyllonen, 2002). An item model is a set of items that share a common structure, defined through the use of variables and constraints. Mirzadeh et al. (2024) used a model-based approach to assess LLM math reasoning by generating template-based variants of existing tasks. These new instances avoided leakage, and performance often differed from the original tasks—especially when variations involved numeric values or complexity rather than names. Each problem category was represented as an item modeland used to generate 10 instances. The item models were defined as follows: (1) Item model 1 involves finding the product of two 5-digit whole numbers. (2) Item model 2 involves finding two distinct two-digit whole numbers with a given sum and a given product, where the sum is less than or equal to 100, and neither number is divisible by 10. (3) Item model 3 involves finding a pair of positive integers *x, y* that satisfy an equation of the form $p_1 x^a = p_2 y^b$, where p_1 and p_2 are distinct primes such that each is less than or equal to 11. The exponents *a, b* are relatively prime and each is less than or equal to 9.

LLM Models Setup. We defined two scenarios for configuring LLMs as agents to solve math problems: single-agent and dual-agent setups. Both scenarios are designed to elicit comprehensive, step-level solutions as well as accurate final answers. In the single-agent scenario, individual LLMs, including OpenAI GPT-4o, DeepSeek-V3, OpenAI o1, and DeepSeek-R1, are configured as math problem-solving assistant agents that independently perform math problem-solving tasks. In the dual-agent scenario, two base LLM models (OpenAI GPT-4o and DeepSeek-V3) collaborate as peer agents through interactive, chat-based discussions, exchanging ideas and jointly deriving solutions. Each setup was repeated in three independent runs to ensure reliability.

Evaluations. For instances of the first two item models, the answer key is either a single value (Item Model 1) or, assuming the two integers are interchangeable, a single pair of values (Item Model 2). For instances from Item Model 3, however, there are infinitely many pairs x, y that satisfy the given equation. Since only one pair x, y is requested however, evaluating correctness can be accomplished by substituting the provided values for x, y into the given equation–if this yields a true result, the response is correct, otherwise, it is incorrect. Since solutions are always possible, any response that states there are no solutions is incorrect. To evaluate the solutions, we used a structured coding process: (1) each solution was segmented into discrete, logical steps. (2) each step was labeled according to a predefined rubric detailed in Table 1, categorizing step labels as CC (Conditionally Correct), PE (Procedural Error), CE (Conceptual Error), or IE (Impasse Error). The labeling was performed by the o1 LLM model, followed by verification from human experts. We applied a conditional scoring approach to avoid penalizing LLMs for errors made in earlier solution steps. Analysis of labeling patterns will be reported separately.

Table 1: Math Problem Solution Coding Rubric.

Step Code	**Definition**
Conditionally Correct (CC)	A step that demonstrates procedural and conceptual accuracy, controlling for any errors that may have occurred on previous steps.
Procedural Error (PE)	A step that contains one or more transcription errors, arithmetic mistakes, or symbolic manipulation errors, but without underlying conceptual misunderstanding.
Conceptual Error (CE)	A step demonstrating one or more incorrect applications or misunderstandings of relevant math concepts or principles. It may include misunderstanding the problem, representing it incorrectly, or committing reasoning errors between steps.
Impasse Error (IE)	A step where the solver is unable to proceed further logically or mathematically, indicating a critical gap or blockage in problem-solving understanding.

4 Results and Discussion

The preliminary results presented below include systematic testing of LLMs' performance across three distinct types of math problems, labeling outcomes that identify solution errors across these problem types, and an initial exploration of collaborative LLM-based agents for math tasks.

Accuracy of Final Answers from Single-Agent. Figure 1 presents the performance of four LLMs on a math problems involving the multiplication of two 5-digit numbers. GPT-4o exhibited the lowest performance, with only two correct answers overall. DeepSeek-V3 started strong (8/10 correct) and quickly achieved perfect accuracy in the subsequent iterations (28/30 total). The o1 model demonstrated flawless accuracy from the outset, solving all problems correctly across all three iterations. DeepSeek-R1 achieved only four correct answers across all three runs combined. Interestingly, while the reasoning-enhanced o1 model significantly outperformed its base counterpart GPT-4o, this was not the case with DeepSeek-R1 relative to DeepSeek-V3. We found that DeepSeek-R1 struggled substantially on our proposed tasks. Upon examining its detailed solutions, the model appeared to exhibit an "overthinking" phenomenon (akin to "spinning wheels"), characterized by excessive reflection on intermediate reasoning steps, causing it to overlook critical components necessary for accurate solutions. This outcome of DeepSeek-R1 deviates from performance reported in prior benchmark evaluations (Guo et al., 2025). As shown in Figure 2, we evaluated the number of correct answers provided by each model for algebraic word problems involving quadratic equations. GPT-4o demonstrated moderate performance, correctly solving 9 out of 10 problems in Iterations 1 and 2, but experiencing a slight drop to 7 correct solutions in Iteration 3. In contrast, DeepSeek-V3, o1, and DeepSeek-R1 consistently achieved perfect accuracy, correctly solving all 10 problems across each of the three iterations. Figure 3 presents the accuracy results, showing the performance ranking as follows: o1 (25/30) > DeepSeek-V3 (21/30) > DeepSeek-R1 (20/30) > GPT-4o (8/30). The advanced o1 model clearly outperformed its base counterpart GPT-4o; however, this was not the case within the DeepSeek series.

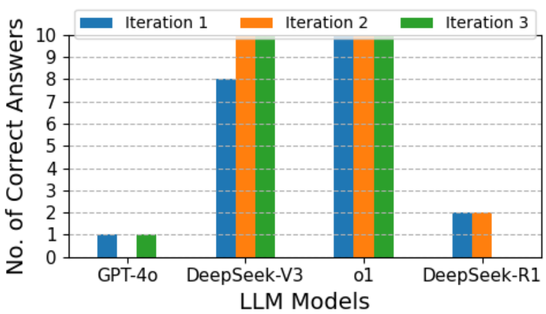


Figure 1: Correctness Across Three Iterations for the Multiplying Two 5-digit Numbers.

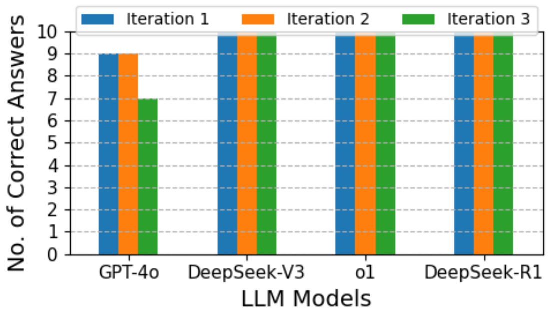


Figure 2: Correctness Across Three Iterations for Solving Algebraic Word Problems Involving Quadratic Equations.

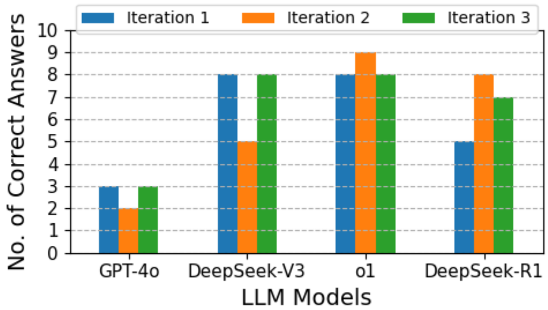


Figure 3: Correctness Across Three Iterations for Solving Diophantine equations.

Evaluations by LLM-based Labeling in the Single-Agent Scenario. The labeled steps in math tasks across the three iterations of the single-agent scenario are shown in Figure 4. We specifically selected these math tasks due to their tendency to highlight significant errors, reflecting notably lower accuracy for some LLM models. Among all tasks, the "CC" label consistently occurs with the highest frequency across models. However, the presence of "CE" labels, notably observed for DeepSeek-R1 (Q3), DeepSeek-V3 (Q3) and GPT-4o (Q3), indicates gaps in understanding fundamental math concepts and principles necessary for accurate solutions, potentially explaining their reduced performance. GPT-4o (Q1) and GPT-4o (Q3) demonstrate the most frequent "PE" occurrences, significantly impacting its overall performance (see Figure 1). Conversely, DeepSeek-V3 (Q1) and GPT-4o (Q2) and o1 (Q3) exhibit no clear conceptual misunderstandings or procedural errors, consequently achieving the highest overall accuracy across the math tasks. In these cases, some incorrect final answers occurred despite no clearly identifiable errors (a phenomenon consistent with our experience using LLMs). A straightforward explanation is that LLMs inherently rely on token prediction rather than explicit numerical computation, rendering them vulnerable to subtle numerical inaccuracies. Further research is necessary to better understand this behavior. Although DeepSeek-

R1 (Q1) mostly exhibits minor procedural errors ("PE") in technically correct steps, its overly complex and inefficient reasoning significantly impedes achieving correct final answers.

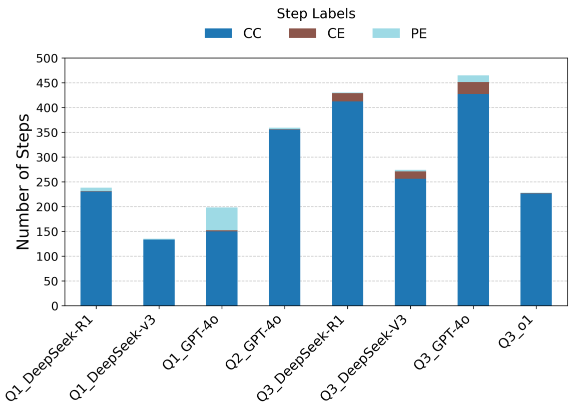


Figure 4: Frequencies of Step Labels in Math Tasks (Where LLMs Stumble). Note: Q1= multiplying two 5-digit numbers; Q2= solving algebraic word problems, Q3= finding solutions to Diophantine equations. We only labeled sets of math problems in which the LLMs produced incorrect final answers, excluding those with 100% correctness.

As a case study, an expert human coder verified the automated labels produced by GPT-4o and o1 on 70 solution steps drawn from first-iteration problems on the multiplication of two five-digit numbers. We used a verification process rather than an independent coding procedure due to time constraints; however the expert evaluated each step using the rubric in Table 1. Cohen's κ shows that GPT-4o achieves only *fair* agreement with the human coder ($\kappa = 0.366$), whereas o1 attains *substantial* agreement ($\kappa = 0.737$), nearly doubling reliability. These results suggest that LLMs with stronger math competence, such as o1, yield more dependable step-level annotations, reinforcing their suitability for automated formative assessment.

Accuracy of Final Answers from Dual-Agent Collaboration. Figure 5 presents performance results for the dual-agent scenario in solving problems involving the multiplication of two 5-digit numbers. The dual-agent configuration with GPT-4o significantly outperformed the single-agent setup, correctly answering 14 out of 30 questions compared to only 2 out of 32 questions in the single-agent scenario. Figure 6 illustrates that both LLM models in the dual-agent scenario achieved perfect accuracy on the quadratic equations questions, surpassing the performance of GPT-4o operating individually as a single agent, which correctly answered only 25 out of 30 problems. Figure 7 demonstrates improved accuracy in dual-agent scenarios compared to single-agent setups on the Diophantine equations questions: GPT-4o improved from 8 out of 30 to 15 out of 30, while DeepSeek-V3 notably increased from 21 out of 30 to a perfect 30 out of 30. These results align with findings from Zhang et al.'s study (Zhang et al., 2025a,b), highlighting that dual-agent collaboration among LLMs can replicate key benefits of human collaboration. Specifically, collaboration in dual-agent scenarios enhances efficiency by enabling two LLM-based agents to share diverse perspectives, cross-validate solutions, and foster emergent reasoning (Chen et al., 2023; Liang et al., 2024). Such collaborative mechanisms hold promise for future improvements in math assessment.

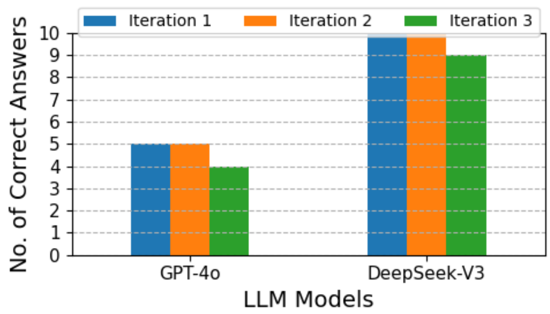


Figure 5: Dual-agent Correctness Across Three Iterations for Multiplying Two 5-digit Numbers.

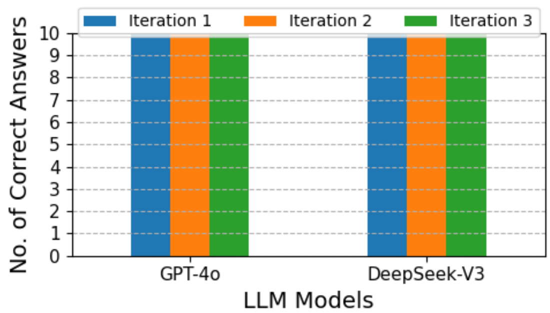


Figure 6: Dual-agent Correctness Across Three Iterations for Solving Algebraic Word Problems Involving Quadratic Equations.

5 Future Work

Future work should include more detailed labeling to better understand solution errors. For instance, in multiplication problems, errors often occurred in the final step, incorrectly summing

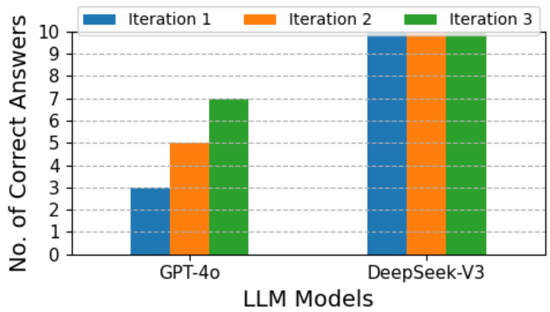


Figure 7: Dual-agent Correctness Across Three Iterations for Solving Diophantine equations.

partial products. Prompt revisions explicitly instructing step-by-step calculations could reduce such errors. Another promising direction is integrating third-party tools including calculators, spreadsheets, or computer algebra systems to handle computations, with LLMs providing reasoning and explanations. This raises the question of how to decide when to delegate tasks to external tools. Developing fine-tuning methods to improve LLM–tool integration could further enhance accuracy. Assuming math problem-solving performance can be improved through prompt revision or integrating LLMs with third-party tools, a critical question remains: Can such systems effectively support instruction and assessment? Beyond correctness, effective instructional use requires pedagogically sound approaches, and effective assessment demands accurate identification of genuine understanding. Addressing these questions is essential for practical classroom implementation and represents an important next research step. Multi-agent approaches to math problem-solving like the one in this study, which leverage collaborative thinking and collective intelligence (Zhang et al., 2025a; Latif et al., 2024), should be further explored. Finally, our findings indicate that stronger performance on final math answers tends to correlate with higher accuracy in step-level assessment. Future studies should investigate the mechanisms underlying this relationship. We also see substantial value in examining how these insights can inform the design of math items, improve formative feedback systems, and enhance the reliability of automated assessment frameworks.

6 Limitations

As it is based on only three item models, the dataset is limited and needs to be scaled up to include both more item models and more instances of each model. We used a rather general rubric in this study; it is possible that a fine-grained rubric with more categories could uncover more insights about the nature of error patterns within solutions. In the interest of saving time, the LLM labeling and the human labeling were not independent; rather, the human verified the LLMs' labels for a portion of the data. Future work would examine agreement between LLM labeling and human labeling as independent processes. Nevertheless, LLM labeling with human verification reached 91.5% exact-match. Due to financial constraints, additional commercial models such as OpenAI o3 or more other LLM models like Anthropic Claude were not tested but could provide valuable insights and further evidence if included in future evaluations.

7 Conclusion

This study systematically evaluated four LLMs, including two base models (OpenAI GPT-4o and DeepSeek-V3) and two advanced reasoning models (OpenAI o1 and DeepSeek-R1), across parallel arithmetic, algebraic, and number-theoretic item models in both single- and dual-agent paradigms. Models with stronger numerical competence, exemplified by o1, achieved step-level annotations that nearly doubled inter-rater agreement with human experts, underscoring their promise for scalable formative assessment. Dual-agent collaboration, mirroring the key benefits of human collaboration, further enhanced math problem-solving performance through cross-validation and emergent reasoning. In this study, the publicly released dataset, coding rubric, and benchmarking protocol equip researchers and practitioners with practical tools for pinpointing procedural versus conceptual breakdowns and for designing AI-enhanced teaching strategies. Future work will expand the problem bank, refine the error taxonomy, and integrate LLMs with external computational engines, bringing us closer to classroom-ready, pedagogically sound AI-based math instruction and assessment.

Acknowledgments

Thanks to Ikkyu Choi at ETS for generating the item model instances used in this research. Portions of the manuscript were refined with the assistance of AI tools solely for grammar checking and sentence refinement, without involvement in other study tasks.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Preeti Anand. 2023. Khan academy creates GPT-4-based helper Khanmigo, marking the formal entry of ai into education.

Konstantine Arkoudas. 2023. Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*.

Lex Bayer. 2025. Studying with Q-Chat (support article). `https://quizlet.com/blog/meet-q-chat`. Accessed: 2025.

Isaac I. Bejar. 2002. Generative testing: From conception to implementation. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item generation for test development*, pages 199–218. Lawrence Erlbaum Associates, Mahwah, NJ.

Johan Boye and Birger Moell. 2025. Large language models and mathematical reasoning failures. *arXiv preprint arXiv:2502.11574*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Susan E Embretson. 1999. Generating items during testing: Psychometric issues and models. *Psychometrika*, 64:407–433.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36:27699–27744.

Mark J Gierl and Thomas M Haladyna. 2013. *Automatic item generation: Theory and practice*. Routledge.

Edith Aurora Graf, Carol Forsyth, Shona Ruiz Diaz, Duanli Yan, and Yang Jiang. 2025. Mathematical explorations in an llm. In *Applications of Generative AI to Mathematics Education: Opportunities and Challenges*, Denver, CO. Paper presented at the annual meeting of the National Council on Measurement in Education.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Adit Gupta, Jennifer Reddig, Tommaso Calo, Daniel Weitekamp, and Christopher J MacLellan. 2025. Beyond final answers: Evaluating large language models for math tutoring. *arXiv preprint arXiv:2503.16460*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Sidney H. Irvine and Patrick C. Kyllonen, editors. 2002. *Item Generation for Test Development*. Psychology Press.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Hyoungwook Jin, Yoonsu Kim, Dongyun Jung, Seungju Kim, Kiyoon Choi, Jinho Son, and Juho Kim. 2025. Investigating large language models in diagnosing students' cognitive skills in math problem-solving. *arXiv preprint arXiv:2504.00843*.

Anthony LaDuca, WI Staples, B Templeton, and GB Holzman. 1986. Item modelling procedure for constructing content-equivalent multiple choice questions. *Medical education*, 20(1):53–56.

Ehsan Latif, Yifan Zhou, Shuchen Guo, Yizhu Gao, Lehong Shi, Matthew Nayaaba, Gyeonggeon Lee, Liang Zhang, Arne Bewersdorff, Luyang Fang, and 1 others. 2024. A systematic assessment of openai o1-preview for higher order thinking in education. *arXiv preprint arXiv:2410.21287*.

Junsong Li, Jie Zhou, Yutao Yang, Bihao Zhan, Qianjun Pan, Yuyang Ding, Qin Chen, Jiang Bo, Xin Lin, and Liang He. 2025. Teaching llms for step-level automatic math correction via reinforcement learning. *arXiv preprint arXiv:2503.18432*.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. *arXiv preprint arXiv:2406.00755*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jeff Maggioncalda. 2024. Coursera launches a new suite of academic-integrity features to help universities verify learning in an age of AI-assisted cheating. Coursera Blog.

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and 1 others. 2024. Transformers can do arithmetic with the right embeddings. *Advances in Neural Information Processing Systems*, 37:108012–108041.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *Preprint*, arXiv:2410.05229.

Zhuoshi Pan, Yu Li, Honglin Lin, Qizhi Pei, Zinan Tang, Wei Wu, Chenlin Ming, H Vicky Zhao, Conghui He, and Lijun Wu. 2025. Lemma: Learning from errors for mathematical advancement in llms. *arXiv preprint arXiv:2503.17439*.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating llms on 2025 usa math olympiad. *arXiv preprint arXiv:2503.21934*.

Tiasa Singha Roy, Aditeya Baral, Ayush Rajesh Jhaveri, and Yusuf Baig. 2025. Can llms understand math? exploring the pitfalls in mathematical reasoning. *arXiv preprint arXiv:2505.15623*.

Junior Cedric Tonga, KV Srivatsa, Kaushal Kumar Maurya, Fajri Koto, and Ekaterina Kochmar. 2025. Simulating llm-to-llm tutoring for multilingual math feedback. *arXiv preprint arXiv:2506.04920*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Stephen Wolfram. 2023. Wolfram|alpha as the way to bring computational knowledge superpowers to chatgpt. stephen wolfram writings.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2024. Number cookbook: Number understanding of language models and how to improve it. *arXiv preprint arXiv:2411.03766*.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, and 1 others. 2024. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836.

Liang Zhang, Jionghao Lin, Xiaoming Zhai, Diego Zapata-Rivera, Carol Forsyth, Yifan Zhou, Bolun Sun, Arthur C Graesser, and Xiangen Hu. 2025a. Exploring communicative strategies for dual llm agents in mathematical problem solving.

Liang Zhang, Xiaoming Zhai, Jionghao Lin, Jennifer Kleiman, Diego Zapata-Rivera, Carol Forsyth, Yang Jiang, Xiangen Hu, and Arthur C Graesser. 2025b. Exploring communication strategies for collaborative llm agents in mathematical problem-solving. In *International Conference on Artificial Intelligence in Education*, pages 258–265. Springer.

Mengxue Zhang, Zichao Wang, Zhichao Yang, Weiqi Feng, and Andrew Lan. 2023. Interpretable math word problem solution generation via step-by-step planning. *arXiv preprint arXiv:2306.00784*.

Author Index