AIME-Con 2025

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

Volume 2: Works in Progress

October 27-29, 2025

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

Platinum

duolingo english test

Pearson

Gold

Curriculum Associates i-Ready

ets research institute

Silver

Cambium Assessment

CollegeBoard

edCount LLC because all students count

Gates Foundation

NBME

Supporters

FOCALPOINT
Teaching & Learning, Measured, Managed

MI
MEASUREMENT
INCORPORATED

University of Pittsburgh
Learning Research and Development Center

WestEd
WestEd.org

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.

The future of *i-Ready Assessment* is invisible.

- Voice technology is coming to *i-Ready Literacy Tasks*
- Built to hear students' voices of all accents and dialects
- Creating the best possible solution by collaboratively learning with teachers in the classroom

AI Labs™
Curriculum Associates®

Learn more about our vision for the future

---

*ets research institute

Shaping the Future of AI in Assessment

ETS advances responsible AI research to promote fairness, trust, and innovation.

As AI transforms education, ETS brings decades of expertise to ensure that new solutions are not only powerful, but also valid, equitable, and transparent. Our work is driving the next-generation of measurement science, standing at the intersection of AI, learning, and assessment.

Highlights from ETS research at NCME AIME 2025:
- Investigating racial and ethnic subgroup representation in automated essay scoring
- Using generative AI teaching simulations to support teacher training
- Designing fairness-promoting, automated fraud detection systems
- Validating AI generated scoring rationales

REVIEW OUR
GUIDELINES FOR
RESPONSIBLE AI→

Copyright © 2025 by ETS. ETS is a registered trademark of ETS.
The Eight-Point logo is a trademark of ETS.

Advancing Assessment with AI

Grounded in science and responsible best practices, we use AI to enhance how we measure what students know and can do.

● 19 states
we serve use hybrid scoring

24M essays & short answers
auto-scored by our AI engines

2M verbal responses
auto-scored by our AI engines

More AI-Powered Features - *Coming Soon!*

- WriteOn with Cambi
- Item Parameter Estimation
- Cheating Analysis
- Teacher Authoring with AI passage generation
- Hotline for student-at-risk work detection

CambiumAssessment

Data reflects the 2024-2025 academic year

CollegeBoard

College Board Is a Proud Sponsor of AIME-Con

Join our engaging sessions to learn how we're advancing innovative and responsible use of AI in educational measurement.

edCount is pleased to sponsor 2025 NCME AIME-Con

Over 20 years of service to students and educators!

Our Belief Statement
Every individual brings unique experiences, skill sets, and perspectives that work to advance our purpose: continuously improving the quality, fairness, and accessibility of education for all students.

Our Services

- Assessment Design, Development, and Evaluation
- Instructional Systems and Capacity Building
- Policy Analysis and Technical Assistance

edCount LLC®
because all students count

www.edCount.com
(202) 895-1502 | info@edCount.com

NBME®

www.NBME.org

ADVANCING ASSESSMENT,
SUPPORTING OPTIMAL CARE

Through research and collaboration, NBME is evolving how we evaluate and support learners, with a focus on applying new technology to develop assessments that measure and build the knowledge and skills needed to provide optimal, effective care to all.

©2025 National Council on Measurement in Education (NCME)

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 979-8-218-84229-1

Preface

AIME-CON
2025
Pittsburgh, PA, USA ncme

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 precision/accuracy, 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 Work-in-Progress Papers

AIME-Con invited submissions of Working Papers, which were submissions of up to $1,000$ words in the form of a structured summary. This format was designed for work-in-progress or pilot studies. Working Papers required a title, short abstract, and followed a structured format with the following headings:

- Background
- Aims
- Sample(s)
- Methods
- Results (or Anticipated Results)
- Conclusions (or Anticipated Contributions)
- References, tables, and figures included as needed.

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

Automated Item Neutralization for Non-Cognitive Scales: A Large Language Model Approach to Reducing Social-Desirability Bias
Sirui Wu and Daijin Yang . . . 1

AI as a Mind Partner: Cognitive Impact in Pakistan's Educational Landscape
Eman Khalid, Hammad Javaid, Yashal Waseem and Natasha Sohail Barlas . . . 14

Detecting Math Misconceptions: An AI Benchmark Dataset
Bethany Rittle-Johnson, Rebecca Adler, Kelley Durkin, L Burleigh, Jules King and Scott Crossley 20

Optimizing Opportunity: An AI-Driven Approach to Redistricting for Fairer School Funding
Jordan Abbott . . . 25

Automatic Grading of Student Work Using Simulated Rubric-Based Data and GenAI Models
Yiyao Yang and Yasemin Gulbahar . . . 34

Cognitive Engagement in GenAI Tutor Conversations: At-scale Measurement and Impact on Learning
Kodi Weatherholtz, Kelli Millwood Hill, Kristen Dicerbo, Walt Wells, Phillip Grimaldi, Maya Miller-Vedam, Charles Hogg and Bogdan Yamkovenko . . . 40

Chain-of-Thought Prompting for Automated Evaluation of Revision Patterns in Young Student Writing
Tianwen Li, Michelle Hong, Lindsay Clare Matsumura, Elaine Lin Wang, Diane Litman, Zhexiong Liu and Richard Correnti . . . 49

Predicting and Evaluating Item Responses Using Machine Learning, Text Embeddings, and LLMs
Evelyn Johnson, Hsin-Ro Wei, Tong Wu and Huan Liu . . . 66

Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity
Yue Huang and Joshua Wilson . . . 71

Comparing AI tools and Human Raters in Predicting Reading Item Difficulty
Hongli Li, Roula Aldib, Chad Marchong and Kevin Fan . . . 84

When Machines Mislead: Human Review of Erroneous AI Cheating Signals
William Belzak, Chenhao Niu and Angel Ortmann Lee . . . 90

Fairness in Formative AI: Cognitive Complexity in Chatbot Questions Across Research Topics
Alexandra Barry Colbert and Karen D Wang . . . 98

Keystroke Analysis in Digital Test Security: AI Approaches for Copy-Typing Detection and Cheating Ring Identification
Chenhao Niu, Yong-Siang Shih, Manqian Liao, Ruidong Liu and Angel Ortmann Lee . . . 107

Talking to Learn: A SoTL Study of Generative AI-Facilitated Feynman Reviews
Madeline Rose Mattox, Natalie Hutchins and Jamie J Jirout . . . 117

AI-Powered Coding of Elementary Students' Small-Group Discussions about Text
Carla Firetto, P. Karen Murphy, Lin Yan and Yue Tang . . . 125

Evaluating the Reliability of Human–AI Collaborative Scoring of Written Arguments Using Rational Force Model
Noriko Takahashi, Abraham Onuorah, Alina Reznitskaya, Evgeny Chukharev, Ariel Sykes, Michele Flammia and Joe Oyler . . . 135

Evaluating Deep Learning and Transformer Models on SME and GenAI Items
Joe Betts and William Muntean 141

Comparison of AI and Human Scoring on A Visual Arts Assessment
Ning Jiang, Yue Huang and Jie Chen 147

Explainable Writing Scores via Fine-grained, LLM-Generated Features
James V Bruno and Lee Becker 155

Validating Generative AI Scoring of Constructed Responses with Cognitive Diagnosis
Hyunjoo Kim 166

Automated Diagnosis of Students' Number Line Strategies for Fractions
Zhizhi Wang, Dake Zhang, Min Li and Yuhan Tao 178

Medical Item Difficulty Prediction Using Machine Learning
Hope Oluwaseun Adegoke, Ying Du and Andrew Dwyer 185

Examining decoding items using engine transcriptions and scoring in early literacy assessment
Zachary Schultz, Mackenzie Young, Debbie Dugdale and Susan Lottridge 191

Addressing Few-Shot LLM Classification Instability Through Explanation-Augmented Distillation
William Muntean and Joe Betts 197

Identifying Biases in Large Language Model Assessment of Linguistically Diverse Texts
Lionel Hsien Meng, Shamya Karumbaiah, Vivek Saravanan and Daniel Bolt 204

Implicit Biases in Large Vision–Language Models in Classroom Contexts
Peter Baldwin 211

Enhancing Item Difficulty Prediction in Large-scale Assessment with Large Language Model
Mubarak Mojoyinola, Olasunkanmi James Kehinde and Judy Tang 218

Leveraging LLMs for Cognitive Skill Mapping in TIMSS Mathematics Assessment
Ruchi J Sachdeva and Jung Yeon Park 223

Automated Item Neutralization for Non-Cognitive Scales: A Large Language Model Approach to Reducing Social-Desirability Bias

Sirui Wu
University of British Columbia
2125 Main Mall,
Vancouver, BC V6T1Z4, Canada
sirui.wu@ubc.ca

Daijin Yang
Northeastern University
360 Huntington Ave,
Boston, MA 02130, USA
yang.dai@northeastern.edu

Abstract

This study evaluates item neutralization assisted by the large language model (LLM) to reduce social desirability bias in personality assessment. GPT-o3 was used to rewrite the International Personality Item Pool Big Five Measure (IPIP-BFM-50), and 203 participants completed either the original or neutralized form along with the Marlowe–Crowne Social Desirability Scale. The results showed a preserved reliability and a five-factor structure, with gains in conscientiousness and declines in Agreeableness and Openness. The correlations with social desirability decreased for several items, but inconsistently. Configural invariance held, though metric and scalar invariance failed. Findings support AI neutralization as a potential but imperfect bias-reduction method.

1 Introduction

Large language models have primarily been applied to generate cognitive test items and have shown strong performance. With proven powerful contextual understanding and generation abilities in multiple domains (Fitria, 2023; Yang et al., 2025; Ullah et al., 2024), systems such as GPT-3 (Floridi and Chiriatti, 2020) have already produced acceptable multiple choice reading passages (Shin et al., 2025), chemistry and physics items (Chan et al., 2025), and tasks that assess fluid reasoning and visual processing (Ryoo et al., 2022). However, using LLMs for non-cognitive assessments (personality, attitudes, social-emotional skills) is still rare. These constructs are often abstract, value-laden, and context-dependent, which makes automatic item-writing challenging.

Nonetheless, early research is beginning to explore this space. Li et al. (2024a) used GPT-4 to create short, scenario-based questions, named situational judgment items. These items ask people how they would respond in everyday situations, as a way of measuring the Big Five personality traits. In another example, Xue et al. (2025b) relied on GPT-3.5 to expand and translate a university-belongingness questionnaire, maintaining good reliability despite some noisy items. These findings suggest that LLMs can assist non-cognitive scale development, but their robustness and effectiveness remain unverified. Studies have shown that LLM outputs for complex social constructs, such as political or moral values, tend to be overly uniform (Park et al., 2024).

Most prior work has focused on generating new items from scratch, but refining existing validated items through targeted edits is an equally promising yet understudied approach. As emphasized by The Standards for Educational and Psychological Testing (Eignor, 2013), adapting item wording — whether for clarity, cultural context, or bias reduction — can enhance accessibility and fairness while preserving construct validity. McCrae et al. (McCrae et al., 2005) demonstrate that systematic item refinement, like simplifying complex terms in the NEO-PI-3, improves readability and reliability without changing the test's core structure. Studies also show that employing various refinement strategies, such as rephrasing and balancing item tone, can enhance validity while preserving construct discrimination (Bäckström et al., 2014).

To research how LLM could serve as a precise editors, we adopted LLM to identify and decrease the social desirability bias. Social-desirability bias is a tendency for a person to respond in a way that seems socially appealing, regardless of his or her true characteristic (Grimm, 2010; Furr, 2021; Braun et al., 2001). It can contaminate true levels of trait and comparison of individuals, especially on traits such as agreeableness, conscientiousness, and emotional stability (Bäckström and Björklund, 2014), so curbing it is critical. It was chosen for our study not only because it is a common threat to non-cognitive tests, but also because traditional approaches to reducing social desirability, including

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 1–13
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

forced choice (Cao and Drasgow, 2019), balanced keying (Gignac, 2013; Li et al., 2024c), and manual "neutralization" of wording (Bäckström and Björklund, 2020; Bäckström and Björklund, 2013), can work but are labor intensive and may create unintended dimensions (Zhang et al., 2025).

Recent studies demonstrate that LLMs not only display social desirability response patterns similar to humans, but can also detect when they are being evaluated and shift their answers toward socially valued traits. This ability to recognize and reproduce bias suggests that LLMs could also be leveraged to diagnose and potentially mitigate social desirability effects in human surveys (Lee et al., 2024; Salecha et al., 2024). For instance, Dukanovic and Krpan (2025) conducted a real-world hiring study. They required candidates completed both a standard multiple-choice personality questionnaire and a short conversation with an AI chatbot. The chatbot analyzed their written answers and generated personality scores, and they found chatbot-based scores were less influenced by social desirability than the traditional questionnaire scores. However, the chatbot scores were also less effective at predicting external outcomes such as education level or job role. Nevertheless, few studies have evaluated whether LLMs can rewrite test items to reduce their social desirability without compromising reliability and validity.

To address this gap, we used prompt engineering to guide GPT-o3 in revising the IPIP-BFM-50 (Tao et al., 2009), maintaining the test's structure while reducing social desirability bias. The prompt integrated established debiasing strategies (Kajonius, 2017; Bäckström et al., 2014) and incorporated role-playing (Kong et al., 2023), chain-of-thought prompting (Wei et al., 2022), and transparency mechanisms (Schneider, 2024). We evaluated the AI-neutralized items with participants against the original form, examining reliability, factor structure, and correlations with the Marlowe–Crowne Social Desirability Scale.

The results show that AI-based neutralization attenuated social desirability bias while preserving the Big Five structure within each form. Reliability was maintained in most domains, improved for Conscientiousness, but decreased for Agreeableness and Openness. Confirmatory factor analyses supported configural invariance, though full metric and scalar invariance across versions was not achieved. Correlations with social desirability weakened for many items, though effects were uneven across traits.

The discussion highlights both the promise and limitations of AI-assisted item editing. AI neutralization provides a viable tool for reducing response bias without altering trait constructs, but its uneven performance and lack of cross-form equivalence indicate the need for domain-specific fine-tuning, iterative refinement, and human-in-the-loop validations. Taken together, this work demonstrates the potential of large language models to contribute to fairer psychological assessment through targeted item rewriting.

2 Methods

2.1 Instruments

2.1.1 The International Personality Item Pool Big Five Personality Scale (IPIP-BFM-50)

We employed the IPIP-BFM-50 as the foundational measure of the Big Five personality traits, including 50 items (Goldberg et al., 2006; Zheng et al., 2008). Each personality was measured by 10 items. This version of the IPIP-BFM-50 has been previously validated and shown to retain acceptable psychometric properties across multiple studies (Tao et al., 2009). Across multiple cultural validations, Cronbach's alphas are generally high (.80 – .90) for Extraversion, Conscientiousness, Emotional Stability, and Openness, though Agreeableness is sometimes lower (.65 – .70) (Goldberg et al., 2006; Ypofanti et al., 2015; Zheng et al., 2008). Studies also observed validity evidence based on internal structure and relations to other scales. Factor analyses consistently replicate the expected five-factor structure, with strong invariance across gender and ethnic groups (Constantinescu and Constantinescu, 2016; Buchanan et al., 2005; Ehrhart et al., 2008). Validity is shown through substantial correlations with other Big Five instruments including the NEO Five-Factor Inventory (Gow et al., 2005), the Ten Item Personality Inventory (Ypofanti et al., 2015), and the Eysenck Personality Questionnaire–Revised (Gow et al., 2005), often above .60.

2.1.2 IPIP-BFM-50 with AI-neutralization (IPIP-BFM-50-AI)

To systematically reduce social desirability bias in personality assessments, we developed a tailored prompt for GPT-o3, producing the IPIP-BFM-50-AI. GPT-o3 was selected for its strong instruction following, long-context reasoning, and coherent, multi-step outputs (Kim et al., 2025; Ballon et al., 2025; OpenAI, 2025). Our design draws on Bäckström et al.'s manual rewriting strategies (Bäckström et al., 2014), emphasizing reduced evaluative language, preserved behavioral meaning, and midpoint-oriented phrasing—methods shown to reduce item popularity while maintaining validity.

Beyond psychometric strategies, the prompt incorporates techniques to boost effectiveness and interpretability. It frames GPT-o3 as an expert psychometrician (Kong et al., 2023), applies chain-of-thought prompting (Wei et al., 2022) to structure reasoning, and enforces transparency through structured outputs with justifications, bias ratings, and fidelity checks (Schneider, 2024). The full prompt and generated items are included in the appendix.

2.1.3 The Marlowe-Crowne Social Desirability Scale (MC-SDS) short form

The SDS is a validated and widely used measure for assessing socially desirable responding. The Marlowe–Crowne Social Desirability Scale (MC-SDS) short forms, particularly the 13-item Reynolds version, exhibit acceptable internal consistency ($\alpha = .76$ (Reynolds, 1982)) and very high correlations (.80–.90 (Ii and Sipps, 1985)) with the full 33-item scale. This evidence supports their reliability and validity.

2.2 Participants and Data Collection

Participants were recruited online through public advertisements and social networks. Eligibility required age 18 or older, and consent to participate. After reading the study information page and providing informed consent, participants were randomly assigned to complete either the original or the AI neutralized version of the IPIP-BFM-50, creating a between-group design with two independent samples. The two forms used identical content domains but different wording where applicable for the AI-neutralized version. To prevent memory and sequence effects, item order was independently randomized within each version, and the version order was counterbalanced across participants. The response format used a 5-point radio-button scale for all items. The Marlowe–Crowne Social Desirability Scale (short form) was administered after one of the two IPIP administrations. Demographic information (age, gender, education, occupation) was collected at the end to minimize priming. All participants were voluntary recruited by an online link, operated and delivered by a free online survey tool Wjx [1]. We collected 203 response, 102 for Original version and 101 for Neutralized version. After excluding all cases with incomplete items, the sample size was finalized to be 200, each version with 100 responses.

2.3 Evaluation Strategy and Hypothesis

2.3.1 Effectiveness of neutralization.

Item and scale level indicators. We will compare item popularity (means, SDs) and scale means between original and AI-neutralized items to check that highly evaluative items show reduced extreme endorsement without loss of variability.

Desirability linkage. Estimate the correlation of each domain with MC-SDS for the original and neutralized versions within persons. Test whether the neutralized version shows a smaller association with MC-SDS.

2.3.2 Validity evidence following the Standards.

Internal structure and reliability. For each version, test unidimensionality within each domain via CFA or IRT dimensionality checks, then test cross-version invariance (configural, metric, scalar) and report reliability (Cronbach alpha).

Relations to other variables. As discriminant evidence, verify that neutralized scales show weaker correlations with social desirability than originals, while preserving expected convergent patterns with established Big Five constructs.

2.3.3 Hypothesis

Results from all analysis above can be used to check the following hypothesis:

1. **H1 - reliability**: Neutralized domains will demonstrate acceptable reliability that is comparable to originals.

2. **H2 - structure**: Each domain will show a single intended factor per version and acceptable cross-version invariance indices.

3. **H3 - relations**: Neutralized domains will maintain expected convergent patterns with Big Five constructs while showing reduced linkage to social desirability.

---

[1] https://www.wjx.cn/

3 Results

Two balanced groups completed the original and AI-neutralized versions (*n* = 100 each). As shown in Table 1, most participants were between 26 and 40, and there were also respondents aged from 41 to 50, as well as a small 60+ group. Gender distributions were comparable across versions, with roughly equal numbers of men and women. The groups appear demographically similar, supporting a fair comparison of psychometric results between original and neutralized items.

[Table 1 about here.]

3.1 H1: Reliability

Reliability was largely preserved after neutralization. As shown in Table 2, extraversion and Neuroticism remained high in both versions. Conscientiousness improved in the neutralized form. Agreeableness and Openness decreased, with Agreeableness dropping to the mid $.50s$ – $.60s$. Overall, alpha and omega were acceptable for most domains, indicating that neutralization did not broadly undermine internal consistency, though Agreeableness warrants caution. These findings support H1 with noted caveats.

[Table 2 about here.]

3.2 H2: The Validity Evidence from Internal Structure

Single-group confirmatory factor analyses (CFAs) supported the intended five-factor structure for each version. As shown in Table 3, model fit was acceptable for the original version (CFI $\approx$ .97, TLI $\approx$ .97, RMSEA $\approx$ .06) and marginally weaker for the neutralized version (CFI $\approx$ .97, TLI $\approx$ .96, RMSEA $\approx$ .08). Both versions retain the five-factor structure, but the higher RMSEA in the neutralized form points to a few items needing targeted wording revision.

Multi-group tests showed that configural form held, but metric and scalar constraints produced significant misfit with elevated RMSEA, indicating a lack of full cross-version equivalence. Thus, Hypothesis 2 is partially supported: the structure replicates within versions, but strict invariance across versions was not achieved. Configural invariance was supported, indicating that the neutralized and original versions share the same five-factor pattern and item-to-factor assignments. This shows that neutralization preserved the construct blueprint. However, subsequent metric and scalar constraints did not hold, which implies differences in loadings and intercepts across forms. Scores can be interpreted within each form using the same domain structure, but cross-form comparisons of factor means should be deferred until partial invariance or alignment is applied.

[Table 3 about here.]

3.3 H3: The Validity Evidence from Relations to Other Variables

As shown in Table 4, the results highlight differences in correlations between individual items across the five dimensions and the SDR score. It is expected to observe a decrease in difference for absolute value of correlation (no matter a positive or negative), indicating a decrease of influence by SDR. However, we can observe correlations are increase for some items. We conduct the Steiger's Z test to check whether the change in correlation significant, and 6 items indicate a significant change. Among them, five are decrease and one increase.

The neutralized items demonstrated reduced associations in several cases, supporting the intended effect. However, the presence of increases underscores uneven performance across content. Overall, H3 is partially supported: the tool attenuates social desirability bias for many items, but not consistently across the full instrument.

Table 5 further showed details about what items was assessed to have significant change in correlations with SDR after neutralization. The changes align with specific linguistic mechanisms. For Extraversion, neutralized phrasings replace overt status claims with modest, observable behaviors or internal states. This lowers self-presentational stakes and reduces the incentive to answer in a socially approved way. For Openness, edits remove prestige cues (for example, "rich vocabulary") and normalize difficulty with abstract content. Endorsing these becomes less face-threatening, so links to desirability weaken. The Agreeableness increase arises from hedged, evaluative wording ("others might find rude" and "sometimes"). This introduces norm salience and plausible deniability, inviting impression management more than the blunt behavior label "insult people." In short, SDR decreases when wording is concrete, behavioral, and low in status or virtue signals; SDR increases when wording invokes social judgment, hedges frequency, or allows reframing of intent.

To sum up, AI neutralization works, but not uniformly. It maintains reliability in most domains, preserves the factor structure within forms, and reduces desirability in several areas. The costs are local and fixable: a handful of items drive non-invariance and dips in Agreeableness and Openness. Treat scores as within-form for now, apply partial invariance or alignment for cross-form comparisons, and revise the flagged items to restore behavioral precision while keeping neutral tone.

[Table 4 about here.]

[Table 5 about here.]

4 Discussion

The findings indicate that AI-based neutralization can reduce socially desirable responding while preserving the intended construct structure of a Big Five inventory. Single-group CFAs recovered the five-domain pattern in both versions, which suggests that the core representation of the traits remained intact after neutralization. Multi-group analyses supported configural invariance but not metric or scalar invariance, which implies that some item–factor relations and intercepts changed across versions. Reliability remained acceptable for most domains, improved for Conscientiousness, and declined for Agreeableness and Openness. Associations with a social desirability criterion decreased for several items, with notable exceptions in Agreeableness. Together, the results support AI neutralization as a viable wording intervention that targets response bias without altering trait identity.

4.1 Construct representation and measurement comparability

The preserved five-factor structure indicates that neutralization did not shift the meaning of the constructs, which aligns with evidence that the Big Five structure is robust across formats and raters (McCrae and Costa, 1987). The lack of metric and scalar invariance signals that item functioning changed across versions, so cross-form comparisons of means should not be made without partial invariance or alignment solutions (Byrne et al., 1989; Putnick and Bornstein, 2016). Within each form, factors can be interpreted in the usual way. Across forms, unit and intercept differences should be addressed before comparing group or condition means.

4.2 Domain-specific reliability shifts

Conscientiousness reliability increased in the neutralized form, which is consistent with the idea that removing evaluative phrasing can sharpen behavioral focus and raise inter-item coherence. Declines in Agreeableness and Openness suggest that some edits broadened meanings or removed construct-diagnostic cues that previously fostered homogeneity. This pattern is compatible with prior work showing that evaluative wording can inflate internal consistency by cueing a general “goodness” factor, and that neutralizing language can reduce that inflation while leaving substantive variance intact (Bäckström et al., 2014; Bäckström and Björklund, 2013).

4.3 Why correlation with SDR changed

Reductions in correlation with social desirability appear, when wording shifts from status or virtue claims to concrete behaviors or internal states. This likely weakens impression management, which is one facet of socially desirable responding (Paulhus and Reid, 1991). Increases were observed when neutralized items introduced hedges or explicit social judgment cues, which can heighten norm salience and invite self-presentation. These mechanisms align with research on common method bias and evaluative content as drivers of spurious covariance and inflated correlations (Podsakoff et al., 2003; Bäckström and Björklund, 2013).

4.4 Implications for AI-assisted item editing

The results indicate the potential of AI-assisted item editing. Recent research has shown that LLMs themselves exhibit human-like social desirability biases when responding to personality questionnaires, which implies that they are sensitive to the evaluative cues embedded in item wording and may therefore be leveraged to identify and mitigate such bias (Chan et al., 2025). This capacity provides a foundation for the observed reduction in socially desirable responding when items are neutralized with AI support.

However, as the results suggest, a one-time output from a single prompt may not achieve the ideal output. Studies across multiple domains have found that single-shot generation often produces variable quality and is less reliable for tasks requiring precision, nuance, or consistency (Patel et al., 2023; Sahoo et al., 2024). The variability is partly due to the probabilistic nature of LLMs and the difficulty of capturing subtle linguistic properties in a single attempt. Research on prompting and iterative generation shows that multiple candidates and refinement loops generally outperform one-shot outputs, which supports the interpretation that item editing requires more than a single pass (Cheng et al., 2024; Xue et al., 2025a).

Besides using single prompts, other techniques for enhancing large language model behavior are suggested. For the model itself, domain-specific fine-tuning has been shown to substantially improve performance even when only a small amount of high-quality training data is available (Jeong, 2024; Satterfield et al., 2024). In this context, including pairs of successfully human-edited and neutralized items could increase the model's ability to generate valid revisions. However, such data are difficult to obtain, and constructing this type of dataset is therefore an important future direction.

To add control to the system, multiple agents can be combined to provide feedback and review of generated items. One approach is to use another large language model as a reviewer, which can rate and critique generated items. Generate–feedback loops of this kind have proven effective in other domains, such as reasoning and dialogue, by reinforcing higher quality outputs through self-critique and refinement (Li et al., 2024b; Madaan et al., 2023). Beyond automated feedback, incorporating humans in the loop transforms item generation into an iterative process. In such a cycle, participants test the items, results are analyzed, and the items are further refined based on psychometric evidence. This practice reflects established best practices in test development (Eignor, 2013), where iterative pilot testing and expert review are essential to ensure reliability and validity. Yet, the human-LLM collaboration still remains unexplored in the item editing field.

In summary, the results highlight the potential of AI-assisted item editing but also point to current limitations when relying on single-prompt outputs. Future development will benefit from domain-specific fine-tuning, multi-agent or human-in-the-loop feedback mechanisms, and iterative refinement processes that mirror traditional psychometric standards. Together, these strategies can convert AI neutralization into a reproducible pipeline that reduces bias while maintaining the measurement of intended psychological constructs.

4.5 Limitations

The study used a single language, a single instrument, and a between-groups design in a low-stakes context. Social desirability effects can be stronger under incentives to self-present, which limits generalizability to high-stakes settings. All measures were self-report and collected in one session, which raises the possibility of common method variance despite anonymity instructions. The analyses focused on internal structure, reliability, and associations with a bias criterion, so criterion-related validity with external outcomes remains untested for the neutralized form.

Future work should test neutralized items in high-stakes contexts, use within-person designs to estimate per-respondent reductions in bias, and include informant or behavioral criteria to address common method concerns. Partial invariance searches or alignment should be applied to enable cross-form comparisons, and results should document the number and type of freed parameters. The AI pipeline should be benchmarked across models and prompts, with a reusable library of prompt patterns and failure cases by domain. Replication across languages and populations, test–retest studies, and evaluation of predictive validity will clarify whether bias reduction is achieved without loss of criterion-related information.

5 Conclusion

AI-based neutralization reduced social desirability bias while preserving the Big Five construct structure. Reliability shifts varied across domains, improving for Conscientiousness but declining for Agreeableness and Openness, reflecting the influence of evaluative language on internal consistency. Configural invariance was supported, but metric and scalar invariance were not, indicating that cross-form comparisons require partial invariance or alignment methods. The discussion highlights that AI-assisted item editing is promising but uneven, and future development should emphasize domain-specific fine-tuning, iterative refinement, and human-in-the-loop validation to ensure stable and valid measurement.

References

Martin Bäckström and Fredrik Björklund. 2014. Social desirability in personality inventories. *Journal of Individual Differences*, 35(3):144–157.

Martin Bäckström and Fredrik Björklund. 2020. The properties and utility of less evaluative personality scales: Reduction of social desirability; increase of construct and discriminant validity. *Frontiers in psychology*, 11:560271.

Martin Bäckström, Fredrik Björklund, and Magnus R Larsson. 2014. Criterion validity is maintained when items are evaluatively neutralized: Evidence from a full–scale five–factor model inventory. *European Journal of Personality*, 28(6):620–633.

Marthe Ballon, Andres Algaba, and Vincent Ginis. 2025. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*.

Henry I Braun, Douglas N Jackson, and David E Wiley. 2001. Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement*, pages 61–84. Routledge.

Tom Buchanan, John A Johnson, and Lewis R Goldberg. 2005. Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, 21(2):115–127.

Barbara M Byrne, Richard J Shavelson, and Bengt Muthén. 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin*, 105(3):456.

Martin Bäckström and Fredrik Björklund. 2013. Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, 54(2):152–159.

Mengyang Cao and Fritz Drasgow. 2019. Does forcing reduce faking? a meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, 104(11):1347.

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. 2025. Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 8:100344.

Yu Cheng, Jieshan Chen, Qing Huang, Zhenchang Xing, Xiwei Xu, and Qinghua Lu. 2024. Prompt sapper: A LLM-empowered production tool for building AI chains. *ACM Trans. Softw. Eng. Methodol.*, 33(5).

PM Constantinescu and I Constantinescu. 2016. The adaptation of the big-five IPIP-50 questionnaire in romania revisited. *Bulletin of the Transilvania University of Braşov. Series VII: Social Sciences• Law*, pages 129–138.

Danilo Dukanovic and Dario Krpan. 2025. Comparing chatbots to psychometric tests in hiring: reduced social desirability bias, but lower predictive validity. *Frontiers in Psychology*, 16:1564979.

Karen Holcombe Ehrhart, Scott C Roesch, Mark G Ehrhart, and Britta Kilian. 2008. A test of the factor structure equivalence of the 50-item IPIP five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment*, 90(5):507–516.

Daniel R Eignor. 2013. The standards for educational and psychological testing.

Tira Nur Fitria. 2023. Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing english essay. In *ELT Forum: Journal of English Language Teaching*, volume 12, pages 44–58.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

R Michael Furr. 2021. *Psychometrics: an introduction*. SAGE publications.

Gilles E Gignac. 2013. Modeling the balanced inventory of desirable responding: Evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment*, 95(6):645–656.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Alan J Gow, Martha C Whiteman, Alison Pattie, and Ian J Deary. 2005. Goldberg's 'IPIP'big-five factor markers: Internal consistency and concurrent validation in scotland. *Personality and Individual Differences*, 39(2):317–329.

Pamela Grimm. 2010. *Social Desirability Bias*. John Wiley & Sons, Ltd.

Avery Zook Ii and Gary J Sipps. 1985. Cross-validation of a short form of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 41(2):236–238.

Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific LLMs. *arXiv preprint arXiv:2401.02981*.

Petri J Kajonius. 2017. Cross-cultural personality differences between east Asia and northern Europe in IPIP-NEO. *International Journal of Personality Psychology*, 3(1):1–7.

Su Hwan Kim, Severin Schramm, Lena Schmitzer, Kerem Serguen, Sebastian Ziegelmayer, Felix Busch, Alexander Komenda, Marcus Makowski, Lisa C Adams, Keno K Bressem, and 1 others. 2025. Evaluating large language model-generated brain MRI protocols: Performance of GPT-4o, o3-mini, DeepSeek-R1 and Qwen2. 5-72B. *medRxiv*, pages 2025–04.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Sanguk Lee, Kai-Qi Yang, Tai-Quan Peng, Ruth Heo, and Hui Liu. 2024. Exploring social desirability response bias in large language models: Evidence from GPT-4 simulations. *arXiv preprint arXiv:2410.15442*.

Chang-Jin Li, Jiyuan Zhang, Yun Tang, and Jian Li. 2024a. Automatic item generation for personality situational judgment tests with large language models. *arXiv preprint arXiv:2412.12144*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. LLMs-as-judges: a comprehensive survey on LLMs-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Mengtong Li, Bo Zhang, Lingyue Li, Tianjun Sun, and Anna Brown. 2024c. Mixed-keying or desirability-matching in the construction of forced-choice measures? an empirical investigation and practical recommendations. *Organizational Research Methods*, page 10944281241229784.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Robert R McCrae, Paul T Costa, Jr, and Thomas A Martin. 2005. The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of personality assessment*, 84(3):261–270.

OpenAI. 2025. OpenAI o3 and o4-mini system card. System card, OpenAI, San Francisco, CA. Version 2 of the Preparedness Framework.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

Dhavalkumar Patel, Ganesh Raut, Eyal Zimlichman, Satya Narayan Cheetirala, Girish Nadkarni, Benjamin S Glicksberg, Robert Freeman, Prem Timsina, and Eyal Klang. 2023. The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. *MedRxiv*, pages 2023–08.

Delroy L Paulhus and Douglas B Reid. 1991. Enhancement and denial in socially desirable responding. *Journal of personality and social psychology*, 60(2):307.

Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.

Diane L Putnick and Marc H Bornstein. 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41:71–90.

William M Reynolds. 1982. Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of clinical psychology*, 38(1):119–125.

Ji Hoon Ryoo, Sunhee Park, Hongwook Suh, Jaehwa Choi, and Jongkyum Kwon. 2022. Development of a new measure of cognitive ability using automatic item generation and its psychometric properties. *Sage Open*, 12(2):21582440221095016.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Nolan Satterfield, Parker Holbrooka, and Thomas Wilcoxa. 2024. Fine-tuning llama with case law data to improve legal domain performance. *OSF Preprints*.

Johannes Schneider. 2024. Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289.

Dongkwang Shin, Jang Ho Lee, and Kyungmin Kim. 2025. An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal*, page 00336882251326284.

Peng Tao, Dong Guoying, and Stuart Brody. 2009. Preliminary study of a Chinese language short form of the marlowe–crowne social desirability scale. *Psychological reports*, 105(3_suppl):1039–1046.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

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.

Eric Xue, Zeyi Huang, Yuyang Ji, and Haohan Wang. 2025a. Improve: Iterative model pipeline refinement and optimization leveraging llm agents. *arXiv preprint arXiv:2502.18530*.

Mingfeng Xue, Yunting Liu, and HuaXia Xiong. 2025b. Enhancing non-cognitive assessments with GPT: Innovations in item generation and translation for the university belonging questionnaire. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 157–172. PMLR.

Daijin Yang, Erica Kleinman, and Casper Harteveld. 2025. GPT for games: An updated scoping review (2020-2024). *IEEE Transactions on Games*, pages 1–16.

Maria Ypofanti, Vasiliki Zisi, Nikolaos Zourbanos, Barbara Mouchtouri, Pothiti Tzanne, Yannis Theodorakis, and Georgios Lyrakos. 2015. Psychometric properties of the international personality item pool big-five personality questionnaire for the greek population. *Health psychology research*, 3(2):2206.

Xijuan Zhang, Muhua Huang, Jessie Sun, and Victoria Savalei. 2025. Improving the measurement of the big five via alternative formats for the bfi-2. *Journal of Personality Assessment*, pages 1–22.

Lijun Zheng, Lewis R Goldberg, Yong Zheng, Yufang Zhao, Yonglong Tang, and Li Liu. 2008. Reliability and concurrent validation of the IPIP big-five factor markers in China: Consistencies in factor structure between internet-obtained heterosexual and homosexual samples. *Personality and individual differences*, 45(7):649–654.

A Appendix

The Prompt for Neutralizing Self-Report Items

You are an expert psychometrician. Your goal is to reword self-report survey items so they measure the intended vocational interest while minimizing social desirability bias.

Social desirability bias is a type of response bias in research where participants tend to answer questions in a way that they believe will be viewed favorably by others, rather than providing completely honest or truthful responses.

Follow these rules:

- Evaluate each item's social desirability bias. Give each item a score within -5 to 5 where 0 represents the lowest social desirability bias, 5 represents positive social desirability bias (people want to choose the item because they think the item is favorable), and -5 represents negative social desirability bias (people do not want to choose the item because they think the item is unfavorable). Keep the item unchanged if its social desirability bias score is in the zone from -1 to 1.
- Think step-by-step — identify value-laden terms, propose alternatives, and self-check that the new wording still reflects the original behaviour, and that the new wording reduces the social desirability bias — but **do not reveal your reasoning**.
- Remove or soften status-, value-, or social desirability-laden words.
- Construct an item that you would find less desirable yourself.
- If the adjective is evaluatively positive, use a less evaluative one, or rephrase in a way that makes the adjective less evaluative.
- Do not change an item from positive to negative (direction).
- Think of whether the item is reversed or not.
- Preserve each item's core behavioural meaning.
- Pay attention to the dimension of each statement. Do **NOT** change the dimension of each statement.
- Explain your change in natural language for each statement, and give your change a score to indicate its new social desirability bias.

Output format:
Please output the results in a 5-column table titled **Neutralized Items**, with the following headers:

Original | SD Score | Neutralized | Reason | SD Score

Each statement is tagged with a dimension based on the Big Five personality traits. Use the following codes:

- A: Agreeableness
- C: Conscientiousness
- N: Neuroticism

- O: Openness to Experience
- E: Extraversion

The sign "+" or "–" indicates whether the item is positively or negatively phrased within that dimension.

List of Tables

1 Socio-demographics Variable 12
2 Reliability for All Subscales 12
3 Confirmatory Factor Analysis Model Fit on the Big Five Personality Model 12
4 Difference in the Correlations with the SDR 13
5 Original and neutralized items with SDR correlation changes 13

Table 1: Socio-demographics Variable

Age	Original count	Neutralized count
18–25	9	13
26–30	17	24
31–35	59	52
36–40	8	6
41–50	6	3
Over 60	1	2
Gender		
Male	49	46
Female	51	54

Table 2: Reliability for All Subscales

	Original		**Neutralized**	
Subscale	**Alpha**	**Omega**	**Alpha**	**Omega**
Extraversion	0.90	0.91	0.87	0.89
Agreeableness	0.67	0.71	0.59	0.63
Conscientiousness	0.73	0.77	0.79	0.81
Neuroticism	0.91	0.91	0.94	0.94
Openness	0.78	0.78	0.66	0.71

Table 3: Confirmatory Factor Analysis Model Fit on the Big Five Personality Model

Single-group CFA fit

Group	χ^2 (scaled)	df	p (scaled)	CFI	TLI	RMSEA [90% CI]	SRMR
Original	1284.957	1165	<0.001	0.972	0.971	0.060 [0.033, 0.079]	0.093
Neutralized	1336.244	1165	<0.001	0.965	0.963	0.078 [0.055, 0.097]	0.101

Multi-group invariance

Model	Df	AIC	BIC	χ^2	Δdf	$\Delta\chi^2$	p
Configural	2330	25254	26309	4576.3	—	—	—
Metric	2375	25361	26268	4773.8	45	197.51	<0.001
Scalar	2420	25624	26383	5127.0	45	353.16	<0.001

Table 4: Difference in the Correlations with the SDR

	Extraversion	**Agreeableness**	**Conscientiousness**	**Neuroticism**	**Openness**
delta	-0.14	0.06	-0.11	0.12	0.00
p	0.26	0.66	0.43	0.34	0.98
delta	-0.09	-0.03	-0.01	0.11	-0.29
p	0.47	0.84	0.96	0.41	0.03*
delta	0.16	0.28	0.07	0.04	-0.29
p	0.19	0.03*	0.58	0.77	0.04*
delta	-0.45	0.10	0.03	0.16	-0.03
p	<0.001	0.49	0.83	0.20	0.81
delta	-0.12	-0.12	0.10	-0.13	0.21
p	0.37	0.36	0.41	0.29	0.14
delta	-0.13	0.00	-0.05	0.14	-0.02
p	0.29	0.98	0.71	0.24	0.86
delta	-0.14	-0.11	-0.05	0.06	0.06
p	0.28	0.37	0.72	0.60	0.62
delta	-0.24	-0.09	0.09	0.18	-0.07
p	0.05*	0.53	0.45	0.16	0.64
delta	-0.07	0.08	0.18	0.05	-0.26
p	0.58	0.58	0.20	0.68	0.06
delta	0.10	0.01	0.00	0.04	0.14
p	0.44	0.94	0.98	0.73	0.28

Table 5: Original and neutralized items with SDR correlation changes

Version	**Items**	**Personality**	**Direction**	**Correlation with SDR after neutralization**
Original	Don't mind being the center of attention.	Extraversion	Positive	
Neutralized	Feel fine when attention is on me.	Extraversion	Positive	Decrease
Original	Am the life of the party.	Extraversion	Positive	
Neutralized	Often take an active role in group conversations.	Extraversion	Positive	Decrease
Original	Insult people.	Agreeableness	Negative	
Neutralized	Sometimes say things that others might find rude.	Agreeableness	Negative	Increase
Original	Have a rich vocabulary.	Openness	Positive	
Neutralized	Know and use a variety of words.	Openness	Positive	Decrease
Original	Have difficulty understanding abstract ideas.	Openness	Negative	
Neutralized	Find abstract ideas challenging.	Openness	Negative	Decrease

AI as a Mind Partner: Cognitive Impact in Pakistan's Educational Landscape

Eman Khalid [1][0009-0004-7616-4447] Hammad Javaid [2][0009-0006-1432-5924] Yashal Waseem [3][0009-0004-6043-8167] and Natasha Sohail Barlas [4][0000-0002-8094-7274]
1-4 Lahore University of Management Sciences, Lahore PB 54792, Pakistan

Abstract

This study explores how high school and university students in Pakistan perceive and use generative AI as a cognitive extension. Drawing on the Extended Mind Theory, it examines impacts on critical thinking, motivation, and ethics. Findings reveal over-reliance, mixed emotional responses, and institutional uncertainty about AI's role in learning.

1 Background

The rapid adoption of generative artificial intelligence (AI) tools in education has sparked growing academic interest in how such technologies affect learning. AI-based tools now assist students with summarizing, content generation, and problem-solving. Current research largely focuses on university students in the Global North, with limited attention to high school learners in the Global South.

This study draws conceptually from the Extended Mind Theory, understanding how cognitive processes extend beyond the brain into external objects, tools, and technologies [8]. When adolescents use generative AI to offload or scaffold thinking, these tools act as external cognitive resources, per-ordering thinking or bypassing critical cognitive steps. Usage for lower-order cognitive tasks, such as information retrieval and summarization undermines original thinking and self-regulation in students [2,3]. This also has significant motivational and emotional effects on adolescents, both positive and negative [4].

In parallel, global research revealed contrasting perceptions: some students find AI empowering and helpful for independent learning, while others express concerns around ethical use, authenticity, and academic integrity [5]. Over-reliance on AI for quick solutions over conceptual understanding shows shallow cognitive engagement [6]. Work on aligning AI use with Bloom's Taxonomy, also highlights the tension between ease of use and the depth of cognitive effort [7].

Existing studies often focus on usage patterns, surface level attitudes, but lack cognitive integration in perception studies. This study addresses these gaps by centering adolescent perspectives across high school and university contexts, in the Pakistani context.

2 Aims

Using the Extended Mind theory we explore three axes of investigation:
2.1 To examine how students' perception of AI as an extension of their cognitive processes shape their attitudes, emotions, and ethical concerns in educational settings.
2.2 To explore the impact of AI reliance on students' critical thinking and sense of cognitive ownership.
2.3 To analyze how educational policies and institutional frameworks enable or restrict the recognition of AI as an extension of students' cognition.

3 Related Works

Research on generative AI (GenAI) in education consistently shows students framing tools like ChatGPT as pragmatic assistants for brainstorming, summarizing, and drafting while calibrating trust and preferring human input when stakes are high [8,9,10]. In South Asian contexts, surveys likewise report convenient, frequent use

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 14–19
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

coupled with worries about originality and integrity [2,5,6]. Conceptually, this ambivalence aligns with the Extended Mind perspective, students oscillate between treating AI as a tool and as a cognitive partner embedded in their study routines [1].

Cognitively, outcomes hinge on how learners engage rather than mere access. Self-regulated learning accounts emphasize planning, monitoring, and reflection as determinants of benefit, with structured prompting and goal setting linked to better metacognition and reduced dependency [12]. Systematic reviews converge on a design-contingent pattern: perceived efficiency gains are common, but effects on higher-order thinking are variable without scaffolds that force critique and explanation[13,14]. Cognitive offloading research explains the mechanism: outsourcing memory or reasoning can save effort yet depress effort thresholds, weaken retention, and blur ownership of ideas if unregulated [15,16]. Students' reported mixtures of feeling simultaneously empowered and "less original" map onto this tension between support and erosion of cognitive ownership [12,13].

Relational and affective dynamics, especially among adolescents, further shape attitudes toward AI. Youth readily anthropomorphize conversational agents, forming companion-like ties that raise engagement but can induce over-trust without clear boundaries and provenance cues [17,18]. Recent evidence shows that perceived "mind" in chatbots modulates acceptance of support and can normalize reliance in everyday study decisions [19,20]. These dynamics help explain why some learners describe AI as a "friend" or neutral sounding board, blurring tool–partner distinctions within academic contexts.

Institutional policies and classroom norms strongly influence whether students conceptualize AI as part of their cognitive process or as a prohibited shortcut. Studies document uneven or ambiguous guidance that pushes usage underground and heightens anxiety about ethics [11,21,22]. Conversely, clearer acceptable-use matrices, disclosure/citation norms, and AI literacy interventions are associated with more reflective, bounded reliance and a healthier sense of authorship [23]. Instructors' modeling and feedback practices also matter: teacher-in-the-loop analytics and feedback pipelines can channel AI toward reflective uptake rather than answer-consumption [24,25,26].

4 Methods

The sample size for this study consisted of 137 high school and university students, aged 16-22 from Lahore, Pakistan. Students were from diverse academic backgrounds, with gender representation was ensured.

Surveys and semi-structured interviews were conducted. Senior management helped distribute surveys to students in grades 11–12 and first and second-year university students. Research protocols were shared to ensure parental consent, especially for minors. The survey examined students' AI perceptions and usage across three areas: (a) frequency and purpose, (b) self-perception, and (c) institutional perception. Students could volunteer for follow-up interviews by leaving contact details at the end of the survey. We conducted 19 interviews—10 with high school students (7 females, 3 males) and 9 with university students (4 females, 5 males)—across disciplines such as sciences, humanities, and social sciences. The interviews explored the same three themes as the survey, with a deeper focus on (a) contextualizing AI use, (b) comparing self- vs. peer-perceptions, (c) understanding regulation, and (d) analyzing AI's cognitive impact.

5 Findings and Discussion

Four key themes emerged from our study: (a) Perception of AI, (b) Usage Patterns, (c) Cognitive and Emotional Impact, and (d) Institutional Relationship with AI. These findings offer a foundation for understanding AI's role in Pakistan's private education sector.

5.1 Perception of AI

Across interviews, students widely acknowledged AI's utility and convenience, especially in academic contexts. Many likened ChatGPT to a *"replacement for Google"* and even a *"lifesaver"* when under time pressure. Several rated its helpfulness as high as 4 or 5, with one participant comparing the rise of AI to the industrial revolution, underscoring how momentous they believe its impact could be. This framing resonates with global research documenting students' pragmatic use of AI for brainstorming, summarizing, and drafting, while still calibrating trust when stakes are high [9,10].

Yet, a strong undercurrent of apprehension and caution ran through the responses. Students expressed concerns about growing dependency, using terms such as *"compromised self-ability," "loss of critical thinking,"* and *"hindering learning."* One student admitted: *"Sometimes I think I should use my own brain,"* highlighting internal conflict. Such ambivalence mirrors findings from systematic reviews showing that efficiency gains often come at the expense of originality and deep engagement [2]. Cognitive offloading theory helps explain this tension: reliance on external tools can reduce effort thresholds, blur authorship, and weaken retention if left unscaffolded [15].

A subset of students went further, describing AI as a "double-edged sword." They recognized productivity gains but worried it *"makes everyone the same,"* eliminating individuality and hindering innovation. This aligns with concerns raised by Gonsalves [3], who argues that overuse of generative AI risks flattening cognitive diversity and undermining Bloom's higher-order processes. Others emphasized that AI's usefulness was constrained by user skill, noting: *"I have to feed it the solution to get a good explanation."* This perspective reflects the growing recognition that prompting skill is itself a form of digital literacy shaping outcomes [8].

Despite varied views, a common sentiment was that AI is best seen as a supportive tool, not a replacement for thinking. As one participant phrased it: *"Only savour AI, grow your own legs to stand on."* Overall, students perceived AI as simultaneously empowering and risky, powerful for productivity but potentially detrimental to long-term cognitive development. This perception maps onto the broader literature that frames generative AI as both an extension of cognition and a possible inhibitor of originality depending on context, scaffolding, and regulation [1,2]. To further emphasize this "double-edged" narrative, critiques of digital native myths similarly caution that technological convenience does not guarantee deeper learning [27].

5.2 Usage Patterns

AI tools have become a regular part of students' academic workflows, ranging from highly structured routines to spontaneous, on-the-go help. ChatGPT, in particular, emerged as a go-to assistant for brainstorming, summarizing, translating, and even generating entire drafts. For instance, one student explained how they used it to structure a moot court proposal without prior experience, saying it *"was very helpful and the proposal got accepted."* Another highlighted its utility in theory-heavy subjects for *"creating structured approaches,"* while acknowledging its inaccuracy in math-based courses. Such pragmatic integration reflects broader findings that students adopt AI primarily for efficiency and scaffolding, while remaining cautious of its limits [9].

Students also reported interactive uses, routinely uploading images of class slides, book chapters, or questions and requesting explanations or practice questions. This echoes international studies showing adolescents view conversational AI as a responsive partner for inquiry and revision [8]. At the same time, several mentioned using auxiliary tools like Quillbot and Scribbr in tandem with GPT, Quillbot to "humanize" AI text, Scribbr to bypass AI detection. This mirrors concerns raised in higher education about AI misuse and plagiarism anxiety [11].

In terms of timing and workflow, many students noted they rely on AI during "peak study times" to convert informal ideas into formal emails, summarize readings, or draft LinkedIn posts. This aligns with research linking effective AI use to self-regulated learning (SRL) practices, where structured prompting and reflection support deeper engagement [12,13]. However, students also acknowledged frequent double-checking of outputs, suggesting a partial reliance balanced by personal judgment. This reflective verification resonates with findings that scaffolded use can enhance metacognition, while unscaffolded reliance may depress higher-order thinking [14].

Beyond academics, AI was also used for personal interests such as discovering restaurants, writing poems, or generating music chords. Here too, students demonstrated ambivalence, valuing convenience but simultaneously expressing uncertainty about accuracy. This pattern reinforces the notion that context matters: without guidance, convenience-driven use risks shallow engagement, but when embedded within SRL frameworks, AI can enhance planning and reflection [12].

Overall, the usage patterns observed in this study capture a balance of creativity, pragmatism, and caution. Students see AI as a flexible companion integrated into daily study and leisure routines, but their reliance is mediated by perceived risks of misuse and by their own ability to evaluate outputs critically. This duality reflects global patterns of adoption [2,8,9] while adding a unique insight from Pakistan: the deliberate use of plagiarism-bypass tools points to gaps in institutional policy and highlights the need for clearer acceptable-use frameworks [11, 28,29].

5.3 Cognitive and Emotional Impact

Four key themes emerged related to the cognitive impacts of AI: over-reliance, retention, executive functioning, and emotion. About 65.7% of survey respondents felt they over-rely on AI, citing reduced creativity and critical thinking, while 37% of interviewees said it hinders learning and admitted they've become "lazier" with regular use. In contrast, 37% felt their retention and planning had improved, though only 26% believed their decision-making had benefited. Several participants formed subconscious emotional bonds with AI, describing it as a "very intelligent friend" they rely on to make decisions, yet denied any emotional connection when asked directly, highlighting a complex relationship. This mirrors the findings of Jose et al. who discuss how AI can be a cognitive amplifier as well as an inhibitor [30]. While incorporating AI into education can be beneficial, to maximize its positive impact, it should be used as an enabler, not a substitute.

5.4 Institutional Relationship with AI

When inquired about the student's perception of the Institutional Relationship with AI, 32.2% of survey respondents said their institutions discourage AI use, while only 27.7% reported encouragement for responsible use. Additionally, 68% of interviewees felt AI should be strictly regulated in academic settings to prevent it from hindering learning or giving certain students an unfair advantage. There was overall wide support for clear guidelines and moderation, especially to prevent over-reliance, cheating, and loss of critical thinking. Some students proposed time-based usage time limits or subject specific restrictions. There was a split between institutional vs. personal regulation: Many believed schools/universities should take the lead in drawing clear boundaries and enforcing academic integrity. Others emphasized the need for personal responsibility, warning that without internal limits, external rules would be ineffective as students will find a way to circumvent otherwise. This reiterates the findings from another study on the impact of AI on educational performance where the conclusion mentions how institutions need to be trained to thoroughly understand how AI should be used and enforce policies to protect student data [31].

6 Limitations and Future Work

Overall, this study provides foundational insights into AI in Pakistan's educational landscape but is limited by its focus on selective private schools in Lahore, a relatively developed city. This focus may skew results toward more privileged groups with greater AI access, reducing generalizability across other regions of Pakistan, including other provinces and rural areas. To address this, we aim to expand geographic coverage by conducting studies in additional provinces and rural contexts, as well as undertaking cross-country comparisons. We also plan to include public and low-income schools to increase institutional diversity. Furthermore, self-selection may have favored students already comfortable with AI, again limiting generalizability. Our reliance on self-reported data also introduces potential bias. Future research can build on this work by expanding the interview sample size to capture a wider range of student perceptions, and by incorporating a more ethnographic approach that includes observations alongside self-reported data. In addition, we hope to conduct longitudinal studies tracking the same students over time to observe how their AI usage and perceptions evolve with technological advancements. Finally, including teachers in the sample could add valuable external perspectives on the student-AI relationship. Despite these limitations, this study offers an important starting point for integrating AI into educational settings in Pakistan.

Ethics Statement

This study adheres to the ACL Ethics Policy and was conducted with full respect for ethical research practices. All participants were informed about the purpose of the study, their participation was voluntary, and consent was obtained. The study design was reviewed to minimize potential psychological or academic risks to students, and anonymity and confidentiality were preserved throughout data collection, analysis, and reporting. The broader impact of this work lies in its potential to inform more inclusive and contextually grounded AI education policies in the Global South, particularly in under-researched contexts such as Pakistan. While this work advocates for thoughtful integration of generative AI tools into educational settings, it also recognizes concerns around over-reliance, cognitive offloading, and ethical use. We encourage institutions and developers to co-design AI systems with students' cognitive well-being in mind. No conflicts of interest are declared by the authors.

References

[1] Clark, A., Chalmers, D.: The Extended Mind. Analysis. 58, 7–19 (1998). https://doi.org/10.1093/analys/58.1.7.

[2] Castillo-Martínez, I.M., Flores-Bueno, D., Gómez-Puente, S.M., Vite-León, V.O.: AI in higher education: a systematic literature review. Frontiers in Education. 9, (2024). https://doi.org/10.3389/feduc.2024.1391485

[3] Gonsalves, C.: Generative AI's Impact on Critical Thinking: Revisiting Bloom's Taxonomy. Journal of Marketing Education. (2024). https://doi.org/10.1177/02734753241305980.

[4] Kundu, A., Bej, T.: Psychological impacts of AI use on school students: a systematic scoping review of the empirical literature. Research and Practice in Technology Enhanced Learning. 20, 030 (2024). https://doi.org/10.58459/rptel.2025.20030.

[5] Zafar, S., Shaheen, F., Rehan, J.: Use of ChatGPT and Generative AI in Higher Education: Opportunities, Obstacles and Impact on Student Performance. iRASD Journal of Educational Research. 5, 01-12 (2024). https://doi.org/10.52131/jer.2024.v5i1.2463.

[6] Younas, M., Ali, S., Mahmood, T., Rizwan Abbas Nasimi, Ashraf, M.K., Akter, S.: Students' Awareness and Perception Regarding the Usage of AI in Education at Government College University, Faisalabad. Journal of Social Sciences Advancement. 5, 80–85 (2024). https://doi.org/10.52223/jssa24-050410-106.

[7] Hutson, J., & Cotroneo, P. (2023). *Praxis and Augmented Creativity: A Case Study in the Use of Generative Artificial Intelligence (AI) Art in the Digital Art Classroom. The International Journal of Technologies in Learning, 31*(1), 113-132. https://doi.org/10.18848/2327-0144/CGP/v31i01/113-132

[8]Belghith, Y., et al.: Testing, Socializing, Exploring: Characterizing Middle-School Students' Interaction with ChatGPT. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). (2024). https://doi.org/10.1145/3613904.3642332

[9]Kubullek, A. K., et al.: Understanding the Adoption of ChatGPT in Higher Education. ACM Digital Library. (2024). https://doi.org/10.1145/3670653.3677507

[10]Walker, F., et al.: They Prefer Humans! Experimental Measurement of Student Trust in ChatGPT. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). (2024). https://doi.org/10.1145/3613905.3650955

[11]Rogers, M. P., et al.: Attitudes Towards the Use (and Misuse) of ChatGPT. ACM. (2024). https://doi.org/10.1145/3626252.3630784

[12]Chiu, T. K. F.: A Classification Tool to Foster Self-Regulated Learning with ChatGPT (SDT × SRL). Educational Technology Research & Development. (2024). https://doi.org/10.1007/s11423-024-10366-w

[13]Lo, C. K.: The Influence of ChatGPT on Student Engagement: A Systematic Review. Computers & Education. (2024). https://doi.org/10.1016/j.compedu.2024.105100

[14]Deng, R., et al.: Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-Analysis. Computers & Education. (2025). https://doi.org/10.1016/j.compedu.2024.105224

[15]Risko, E. F., Gilbert, S. J.: Cognitive Offloading. Trends in Cognitive Sciences. 20, 676–688 (2016). https://doi.org/10.1016/j.tics.2016.07.001

[16]Ward, A. F., Duke, K., Gneezy, A., Bos, M. W.: Brain Drain: The Mere Presence of One's Smartphone Reduces Available Cognitive Capacity. Journal of the Association for Consumer Research. 2, 140–154 (2017). https://doi.org/10.1086/691462

[17]Rahman, R., et al.: AdolescentBot: Opportunities for Chatbots in Sexual and Reproductive Health. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). (2021). https://doi.org/10.1145/3411764.3445694

[18]Brandtzæg, P. B., Følstad, A., Blikberg, S.: When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). (2021). https://doi.org/10.1145/3411764.3445318

[19]Lee, I., et al.: On the Relationship Between Mind Perception and Acceptance of Chatbot Support. Frontiers in Psychology. (2024). https://doi.org/10.3389/fpsyg.2024.1282036

[20]Young, J., et al.: The Role of AI in Peer Support for Young People. arXiv preprint arXiv:2405.02711. (2024). https://doi.org/10.48550/arXiv.2405.02711

[21]Hasanein, A. M., et al.: Drivers and Consequences of ChatGPT Use in Higher Education. Heliyon. 9(11), e20979 (2023). https://doi.org/10.3390/ejihpe13110181

[22]Shahzad, M. F., et al.: ChatGPT Awareness, Acceptance, and Adoption in Higher Education. International Journal of Educational Technology in Higher Education. 21, 59 (2024). https://doi.org/10.1186/s41239-024-00478-x

[23]Park, H.: The Promise and Peril of ChatGPT in Higher Education. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI). (2024). https://doi.org/10.1145/3613904.3642785

[24]Kasinidou, M.: Students' Perceptions of ChatGPT after Educational Intervention.: ACM. (2024). https://doi.org/10.1145/3649165.3690113

[25]Demszky, D., et al.: Can Automated Feedback Improve Teachers' Uptake of Student Contributions? Educational Researcher. 53, 117–129 (2024). https://doi.org/10.3102/01623737231169270

[26]Whitehill, J., LoCasale-Crouch, J.: Automated Evaluation of Classroom Instructional Support with LLMs. Journal of Educational Data Mining. (2024). https://doi.org/10.5281/zenodo.10974824

[27] Kirschner, P. A., & De Bruyckere, P. (2017). The myths of the digital native and the multitasker. *Teaching and Teacher Education*, *67*, 135–142. https://doi.org/10.1016/j.tate.2017.06.001

[28] Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger platform. *Education and Information Technologies*, *25*(2), 975–997. https://doi.org/10.1007/s10639-019-10027-z

[29] Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning.* Center for Curriculum Redesign. Available at: https://curriculumredesign.org/our-work/artificial-intelligence-in-education/

[30]Jose B, Cherian J, Verghis AM, Varghise SM, S M, Joseph S. The cognitive paradox of AI in education: between enhancement and erosion. Front Psychol. 2025 Apr 14;16:1550621. doi: 10.3389/fpsyg.2025.1550621. PMID: 40297599; PMCID: PMC12036037.

[31]Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. Education Sciences, 15(3), 343. https://doi.org/10.3390/educsci15030343

Detecting Math Misconceptions: An AI Benchmark Dataset

Bethany Rittle-Johnson[1], **Rebecca Adler**[1], **Kelley Durkin**[1], **L Burleigh**[2],
Jules King[2], **Scott Crossley**[1],

[1]Vanderbilt University, [2]The Learning Agency,
Correspondence: bethany.rittle-johnson@vanderbilt.edu

Abstract

To harness the promise of AI for improving math education, AI models need to be able to diagnose math misconceptions. We created an AI benchmark dataset on math misconceptions and other instructionally relevant errors, comprising over 52,000 explanations written over 15 math questions that were scored by expert human raters. A data science competition based on the dataset will lead to state-of-the-art large language models for detecting math misconceptions.

1 Background

Student proficiency in math has declined in the past decade, and pandemic-related school disruptions have left lasting impacts on the nation's and world's education systems (National Center for Education Statistics (NCES), 2022). To harness the promise of AI for improving math education, AI models need to be able to diagnose students' potential math misconceptions and other instructionally relevant errors. Misconceptions are "any student conception that produces a systematic pattern of errors" (Smith III et al., 1994). They often form as people attempt to "assimilate. . . new information into their existing conceptual structures" (Stafylidou and Vosniadou, 2004). For example, students often inappropriately generalize concepts and procedures learned about whole numbers to fractions and decimals (i.e., whole number bias; Durkin and Rittle-Johnson, 2012, 2015; Stafylidou and Vosniadou, 2004). Misconceptions interfere with students' ability to learn correct concepts and procedures and can persist for many years (e.g., Byrd et al., 2015). Other instructionally relevant errors include attending to irrelevant information and incomplete solution procedures.

Directly addressing and countering students' misconceptions improves learning outcomes, including on digital learning platforms (Barbieri et al., 2019; Durkin and Rittle-Johnson, 2012; Huang et al., 2008). However, more research and development infrastructure is needed to ensure that teachers and digital learning platforms can integrate information about math misconceptions into product development, research, and instruction. We hosted a workshop that gathered mathematics cognition researchers and digital learning platform developers together to align needs and priorities, which helped guide our work.

One needed tool is benchmark datasets of math misconceptions and other instructionally relevant math errors. This would enable assessment of how well AI models perform, providing an objective way to compare different AI models and platforms, ensuring transparency, accountability, and suitability for use in education (Hodeem, 2024).

1.1 Aims

This research methods paper introduces an AI benchmark dataset on math misconceptions and other instructionally relevant errors. The dataset is the focus of the MAP - Charting Student Math Misunderstandings data science competition hosted on Kaggle, ending in October 2025, to generate large language models that can detect math misconceptions.

2 Sample

The dataset comprises over 52,000 student explanations written over 15 math questions covering key middle-school math topics. We used a dataset from Eedi, a math learning platform based in the U.K., which has been used by over 250,000 teachers in 16,000 schools. All questions and feedback messages were written by math teachers. Each item begins with a multiple-choice question with 1 correct and 3 incorrect options, with incorrect options meant to capture known misconceptions and other instructionally relevant errors as much as

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 20–24
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

possible. We sampled from instances that included a follow-up prompt for an open-ended explanation of why students selected their answer. We selected items that covered core topics in the U.S. middle-school math curriculum, primarily topics in rational numbers and in equations and functions, and for which there were at least 1000 explanation responses available with at least 5 English words (with one exception). Explanations were screened for personally identifiable information before being shared. No demographic information about the participating students was available.

To ensure a meaningful sample of each misconception code and increased explanations for extensive rater training and norming, 14,368 synthetic explanations were generated using Coedit-XL to supplement the 38,095 original explanations (27% synthetic explanations). A maximum of 1 augmented explanation was generated by paraphrasing an authentic student explanation. Coedit-XL tends to provide the correct spelling, punctuation, and capitalization in paraphrasing students' explanations. To better reflect the writing characteristics of authentic student submissions, one spelling error was randomly injected into 50% of the augmented explanations, and 50% of the augmented explanations were fully lowercased. To confirm the realism of synthetic explanations, an expert reviewed a sample of explanations that were partially real and partially synthetic without an indication of the source. Coedit-XL parameters were iterated on until the expert was unable to differentiate the sample.

3 Methods

Students' explanations (both synthetic and original) were scored by human raters using standardized scoring rubrics and procedures. A scoring rubric for each item was developed by three experts in math cognition and misconceptions, drawing on past research on misconceptions as much as possible. The rubric identified criteria for correct explanations and 2-4 potential instructionally-relevant errors, as outlined in Table 1. Each explanation could only receive one code. Raters went through extensive norming prior to independent rating. Raters were primarily undergraduate or graduate students with prior experience teaching or tutoring children in mathematics.

4 Results

The final dataset comprises 52,463 explanations and metadata in tabular format. The dataset contains student ID numbers, item IDs, explanations, and human-assigned codes. Twenty-four percent of explanations were coded by two raters, with high inter-rater reliability (Cohen's Kappa .70-.90). One item with weak inter-rater reliability was dropped. Inter-reliability was also assessed separately for real and synthetic explanations, which resulted in similar Cohen's Kappa values for each item with large enough synthetic explanation sample sizes for reliable statistical evaluation (real vs synthetic Cohen's Kappa differences: 0.005 - 0.15).

Across the 15 items, 27% of explanations had evidence of a potential misconception, and the frequency of particular misconceptions ranged from 0.2% to 35% of explanations. For example, when asked to calculate $\frac{2}{3} \times 5$, 23% of explanations indicated a misconception that the whole number is converted to a fraction with both its numerator and denominator as that whole number. 41% of explanations were correct (range: 18-74%).

This dataset supports the development and evaluation of state-of-the-art large language models that can detect potential misconceptions, including the MAP - Charting Student Math Misunderstandings data science competition hosted on Kaggle, ending in October 2025. Submitted models will be evaluated with the Mean Average Precision @ 3 metric. Winning models will be posted on Kaggle and can be used as baseline scoring models. One potential limitation is that the models may be overly tuned to particular concepts that may be over-represented because synthetic explanations were paraphrased from authentic explanations. To get a better understanding of the models, prediction accuracies can be evaluated for authentic and synthetic explanations, as well as for each misconception code.

Item Topic	**Annotated sample size**	**Error 1**	**Error 2**	**Error 3**
Fraction Representation	6,963	Believes numerator and denominator of a fraction indicate two separate numbers [WNB]	Incomplete steps: fails to simplify fraction [Incomplete]*	
Adding Fractions	3,994	Adding numerators and denominators without finding common denominator [Adding across]	Finds common denominator and adds numerators [Denominator-only change]*	Creates equivalent fractions, and adds numerator and denominator [Incorrect Equivalent fraction addition]
Finding Fraction of Set #1	4,023	Attending to irrelevant feature [Irrelevant]	Incomplete steps: Calculates unit fraction only [Incomplete]	Calculates fraction for the wrong target [Wrong fraction]*
Finding Fraction of Set #2	2,206	–	Incomplete steps: Calculates unit fraction only [Incomplete]	Calculates fraction for the wrong target [Wrong fraction]*
Equivalent fractions	5,204	Attending to irrelevant feature [Irrelevant]	Treating fraction as 2 separate numbers [WNB]	Additive thinking: finds difference [Additive]*
Dividing fractions	4,476	Multiplies to divide by whole number [Mult.]*	Swaps the divisor and dividend [Swap-Dividend]	Flips the dividend (not divisor) and keeps it as division [FlipChange]
Multiplying two Fractions	2,528	Division instead of multiplication [Division]	Subtracts the provided number [Subtracts]*	
Fraction whole number multiplication	4,411	Multiplies numerator and denominator by the whole number (instead of just the numerator) [Duplication]*	Inverts the whole number multiplier [Inversion]	Adds rather than multiplies [Wrong operation]
Solve for y	3,080	Treats y as a missing digit, rather than a variable [Not variable]*	Transforming problem to addition problem [Adding terms]	Applies wrong operation (i.e., multiplies rather than divides) [Inverse]
Decimal Magnitude	3,320	Believes whole numbers are larger than numbers with decimals [Whole numbers larger]*	Believes longer numbers are bigger [Longer is bigger]	Believes zeroes do not add magnitude information [Ignores zeroes][a]

Continued on next page

Item Topic	**Annotated sample size**	**Error 1**	**Error 2**	**Error 3**
Polygon Sides	1,695	Believes there is not enough information to solve problem [Unknowable]*	Does not use correct formula, and instead divides the total interior angle sum by one interior angle [Interior]	Believes a polygon is defined by having a certain number of sides (5 or 6) [Definition]
Subtracting a negative	4,365	Ignores negative signs and adds them back at the end [Tacking]	Incorrect application of two negatives makes a positive [Two negatives is positive]*	
Functional thinking	3,727	Uses the first term of the output as the coefficient of the rule [firstterm]	Calculates the n+1 term, rather than the n+2 term [wrong term]*	
Proportions	968	Reverses proportional reasoning by multiplying instead of dividing [Multiplying by 4]*	Incorrect base rate [Base rate]	
Odds	1,503	Does not understand the range of probability is 0 to 1 [Scale]*	Believes events with probability $\neq 1$ are certain [Certainty]	

Table 1. Error Categories and Frequencies in the Math Misconceptions AI Benchmark Dataset by Item. Notes: *Most frequent error type for each item; [a]Fourth code was: Believes fewer digits after the decimal point, the larger the number is [Shorter is bigger].

5 Conclusion

To harness the promise of AI for improving math education, AI models need to be able to diagnose students' potential math misconceptions and other instructionally relevant errors. We have created an AI benchmark dataset on math misconceptions covering a variety of middle-school math topics that will be publicly available, along with baseline scoring models. Although this dataset is based on explanations primarily from students in the U.K., their explanations align with misconceptions and correct ways of thinking identified in the research literature conducted primarily in the U.S. and Canada. State-of-the-art large language models based on this dataset will support digital learning platforms' ability to detect math misconceptions, and multiple digital learning platforms are interested in adding this capability. Detecting misconceptions is necessary for them to be addressed, and directly addressing and countering students' misconceptions improves learning outcomes (Barbieri et al., 2019; Durkin and Rittle-Johnson, 2012; Huang et al., 2008).

References

Christina A. Barbieri, Dana Miller-Cotto, and Julie L. Booth. 2019. Lessening the load of misconceptions: Design-based principles for algebra learning. *Journal of the Learning Sciences*, 28(3):381–417.

Caroline E. Byrd, Nicole M. McNeil, Dana L. Chesney, and Percival G. Matthews. 2015. A specific misconception of the equal sign acts as a barrier to children's learning of early algebra. *Learning and Individual Differences*, 38:61–67.

Kelley Durkin and Bethany Rittle-Johnson. 2012. The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, 22(3):206–214.

Kelley Durkin and Bethany Rittle-Johnson. 2015. Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction*, 37:21–29.

Hodeem. 2024. Ai: An overview of common llm benchmarks. `https://dev.to/hmcodes/ai-an-overview-of-common-llm-benchmarks-3i7b`.

Tzu-Hua Huang, Yuan-Chen Liu, and Chia-Ya Shiu. 2008. Construction of an online learning system for decimal numbers through the use of cognitive conflict strategy. *Computers Education*, 50(1):61–76.

National Center for Education Statistics (NCES). 2022. NAEP long-term trend assessment results: Reading and mathematics—Age 9. `https://www.nationsreportcard.gov/ltt/?age=9`.

John P. Smith III, Andrea A. diSessa, and Jeremy Rochelle. 1994. Misconceptions preconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3(2):115–163.

Stamatia Stafylidou and Stella Vosniadou. 2004. The development of students' understanding of the numerical value of fractions. *Learning and Instruction*, 14(5):503–518.

Optimizing Opportunity: An Algorithmic Approach to Redistricting for Fairer School Funding

Jordan Abbott
New America Foundation, Education Funding Equity Initiative
abbott@newamerica.org

Abstract

School district boundaries in the United States are not just lines on a map; they are mechanisms that perpetuate deep-seated educational inequities by directly linking school funding to local property wealth. We present a computational framework for optimizing district boundaries to improve resource equity while reducing racial and economic segregation. This study designs a novel two-stage algorithmic process that uses maximally compact plan initialization from spatial clustering and iterative refinement using Markov Chain Monte Carlo (MCMC) optimization. This hybrid approach can reduce required iterations by more than 90%, relative to traditional methods, and allows systematic variation of different numbers of districts. Hard constraints including contiguity, minimum enrollment thresholds, and infrastructure capacity limits are still enforced. Optimization targets three Theil indices measuring property tax capacity disparities, racial segregation, and economic segregation. Across 42 states, results show average state-level improvements of 66.6% in tax-base equality, 47.6% reduction in racial segregation, and 65.0% decrease in economic segregation.

1 Introduction

Public school district borders determine both educational access and taxing jurisdictions, directly impacting resources available to students. Because a significant portion of school district funding is derived from local property taxes, these boundaries create systematic disparities that correlate with racial and economic segregation. While states compensate through progressive funding formulas, high-wealth districts can more easily raise additional local revenue (Kenyon and Munteanu, 2021). As a result, these school systems can easily increase their budgets beyond what lower-wealth districts can match. Nationwide, property taxes reportedly comprised 65% of local revenues in 2021, but can be much higher in some states (Common Core of Data (CCD), n.d). Given the connection between property values and neighborhood affordability, the students that lose out tend to be those from low-income backgrounds.

The problem also has a troubling racial dimension. This funding system is layered on top of generations of policies and government practices that have created and entrenched racial and economic segregation in housing markets (Kuhn et al., 2018). At different phases of America's past and present, this has included redlining and racially discriminatory mortgage lending; court enforcement of racially restrictive covenants; government-funded construction of segregated housing developments; exclusionary zoning policies; and unfair property assessment, among other forms of discrimination. These factors have shaped both the racial composition of neighborhoods and the property values in the taxing jurisdictions from which school districts raise local dollars.

The result is a map of highly segregated residential communities that demonstrate stark economic divides (Reardon and Weathers, 2024). Left unmitigated, the legacies of discriminatory policies have the potential to intersect to create, shape, and enforce new patterns of segregation (Reardon and Owens, 2014). This problem is further reified by the ways in which school district borders function: both as geographic areas that are home to district students, and as the taxing jurisdictions that yield their local funding (Stadler and Abbott, 2024). Because these boundaries determine the students served by a district and the local funding available for its schools, they function to separate students from resources—and from each other.

Furthermore, students from low-income backgrounds, and with other needs and challenges have demonstrably higher funding needs than students from high-income families (Jackson et al., 2015; Jackson and Mackevicius, 2024). While state aid

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 25–33
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

can and should be used as a tool to provide additional support to students with higher needs, this goal is undermined when these funds are eaten up in pursuit of achieving basic funding parity with high wealth districts. This burdens state education budgets with compensating for existing inequity, rather than achieving equity (Gartner, 2023). Given the high degree of alignment between segregated school districts, patterns of residential segregation in the communities they serve, and funding divides, one option is to consider drawing better school district borders.

2 Theoretical Development

This paper presents the first national-scale computational framework for school redistricting by adapting methods from legislative redistricting. We draw on MCMC methods used to efficiently explore high-dimensional solution spaces for boundary realignment (McCartan and Imai, 2023; Fifield et al., 2014), but introduce additional constraints not often considered in legislative contexts, including property tax capacity, infrastructure limitations, and multi-dimensional segregation measures.

Our work builds upon a small but growing body of literature that has explored educational boundary optimization from different but complementary perspectives. The framework designed by Gillani (2023) found that intradistrict segregation could be reduced while maintaining travel times. However, this approach is not designed to address the interdistrict dynamics where nearly two-thirds of all racial segregation occurs (Owens, 2016). Simko (2024) advanced this line of inquiry through a detailed case study of New Jersey, highlighting the importance of crossing district lines to integrate school systems. These valuable analyses centered on demographic integration, with logistical constraints like student capacity and travel times, while keeping the number of districts constant. Our research extends this conversation by shifting both the geographic scale and the central objective. We also expand the solution space, exploring configurations that vary district counts from 25% to 175% of current levels, in doing so, significantly expanding potential equity gains.[1]

This study adopts fiscal equity as a central optimization goal, a dimension not observed in prior demographic-focused studies. The primary contribution of this approach is the integration of parcel-level property tax assessment data, which allows us to directly model the tax base of each potential district. To complement this, we also incorporate Small Area Income and Poverty Estimates (SAIPE), creating a multi-dimensional economic profile of each proposed district. By constructing boundaries around both fiscal equity and demographic balance, our model is designed to create districts with equitable and sustainable local revenue capacity, a vital consideration given that property taxes constitute, on average, 40% of all district funding (CCD n.d.).

Although changing school district boundaries can be politically challenging, policymakers may be unaware of the extent of existing divides or the degree to which they can be mitigated. Further, there is compelling evidence of efficiency savings from consolidating districts (Duncombe and Yinger, 2007; Dodson and Garrett, 2008). In light of this, several states, including Arkansas, Pennsylvania, and New Jersey have recently undertaken efforts to examine the feasibility of district mergers and other boundary changes. This algorithmic approach provides an objective framework to supply legislators with evidence of the potential benefits of redistricting, including fiscal savings, deconcentrated poverty, and integrated school systems.

3 Problem Formulation

Optimization of school district boundaries represents a high-dimensional combinatorial problem where geographic units are assigned to districts while satisfying multiple objectives and constraints. Unlike legislative redistricting, school redistricting must simultaneously consider property tax capacity, demographic integration, and infrastructure capacity. This section formalizes the mathematical framework underlying our optimization approach.

3.1 Multi-Objective Optimization Framework

Census tracts serve as atomic geographic units that must be assigned to districts. Let $G = \{1, 2, ..., n\}$ represent the set of tracts in a state, and let $D = \{1, 2, ..., k\}$ represent the set of districts, where k varies systematically. Each tract $i \in G$ must be assigned to exactly one district $d \in D$, creating a partition of the geographic space.

Optimization seeks to minimize an objective

---

[1] For some, the immediate reaction may be that such changes are unrealistic. However, our core approach posits that considering drastic changes to district counts is necessary to unlock the full potential for equity gains.

function combining three equity dimensions:

$$f(D) = w_1 T_{\text{val}}(D) + w_2 T_{\text{racial}}(D) + w_3 T_{\text{econ}}(D) \tag{1}$$

where T_{val} measures disparities in per-pupil property tax capacity, T_{racial} captures multigroup racial segregation, and T_{econ} quantifies economic segregation based on binary poverty status. Default weights are set to $w_1 = 3.0$, $w_2 = 1.0$, and $w_3 = 1.0$, reflecting a priority on addressing disparities in property tax capacity while maintaining focus on integration objectives.

3.2 Mathematical Formulation of Theil Indices

3.2.1 Theil's T for Disparities in Property Tax Capacity

The Theil T-index captures inequality in assessed property values per pupil across districts:

$$T_{\text{funding}} = \sum_i \left(p_i \cdot \frac{x_i}{\mu} \cdot \log\left(\frac{x_i}{\mu}\right)\right) \tag{2}$$

where:

- $p_i = n_i/N$, the proportion of total students in district i
- n_i = number of children in district i
- N = total children in the state
- x_i = assessed property value per pupil in district i
- μ = state mean assessed value per pupil

This formula directly measures the capacity to raise local revenue, as property assessments form the tax base for school funding. The index equals zero when all districts have identical per-pupil property values, and increases with greater inequality.

3.2.2 Multigroup Theil's H for Racial Segregation

For racial integration, we employ the multigroup entropy-based Theil H-index:

$$T_{\text{racial}} = (E_{\text{state}} - E_{\text{weighted}})/E_{\text{state}} \tag{3}$$

where:

- $E_{\text{state}} = -\sum_r (\pi_r \cdot \log(\pi_r))$, the entropy of racial composition at state level
- π_r = proportion of racial group r in total state enrollment
- $E_{\text{weighted}} = \sum_i (p_i \cdot E_i)$, the enrollment-weighted average of district entropies
- $E_i = -\sum_r (\pi_{ir} \cdot \log(\pi_{ir}))$, the entropy within district i
- π_{ir} = proportion of group r in district i

The index ranges from 0 (perfect integration, where every district mirrors state demographics) to 1 (complete segregation). This multigroup formulation avoids the limitations of binary indices and captures the full complexity of racial composition.

3.2.3 Binary Theil's H for Economic Segregation

Economic segregation uses a similar entropy-based approach with two groups:

$$T_{\text{economic}} = (E_{\text{state}} - E_{\text{weighted}})/E_{\text{state}} \tag{4}$$

Applied to binary poverty status, as defined by SAIPE, this measure captures the concentration of economic disadvantage across districts. The binary formulation is appropriate, given the policy relevance of poverty thresholds for federal program eligibility.

3.3 Constraint Specifications

1. **Geographic Contiguity:** Each district d must form a connected component under rook adjacency (shared edges, not just vertices).
2. **Minimum Population Threshold:** Each district must contain at least m children, where $m = 0.5 \times \min$ (current district resident school-aged population in each state).
3. **Infrastructure Capacity:** For each district d: $\sum(\text{children in tracts assigned to } d) \leq 1.25 \times \sum(\text{capacity } d)$.

4 Data Architecture

4.1 Geographic Foundation

Census tracts serve as the geographic units for our simulation approach, providing a standardized nationwide framework with sufficient granularity to capture local-level variation. We use 2020 census tract boundaries from the Census Bureau, approximately 80,000 tracts. These polygons define our building blocks for spatial optimization.

Tract adjacency relationships are established using rook contiguity, including only shared tract boundaries rather than vertices. The resulting adjacency matrix forms the foundation for contiguity constraint checking and move generation during optimization. Disconnected components (e.g., islands, water boundaries) are connected via minimum distance stitching between nearest tract centroids to ensure graph connectivity.

4.2 Property Valuation Data

Property valuation data is provided by the Center for Geospatial Solutions at the Lincoln Institute of Land Policy, pre-aggregated at the relevant geographic units of analysis. This dataset provides total assessed property values for each unit from the most recent year available, which directly determine local education revenue capacity. Unlike market values or sale prices, assessed values reflect the actual tax base available to districts.[2]

4.3 Demographic Data

4.3.1 Demographic Composition

Demographic data come from two primary sources. Racial and ethnic composition for five categories (White non-Hispanic, Black, Native American, Asian, and Hispanic/Latino) are obtained from the American Community Survey 5-year estimates (2018–22) for the population ages 5–17. Economic status, specifically child poverty rates, are derived from the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. This approach captures all school-age children regardless of enrollment status.

Our methodology is designed to be robust to the noise introduced to tract-level counts by the Census's Disclosure Avoidance System (DAS). As documented by Kenny and et al. (2021), this noise is non-systematic for census tracts. By aggregating multiple tracts to form each simulated district, our models leverage the law of large numbers, substantially diminishing the effects of random error at our scale of analysis.

4.3.2 Enrollment Capacity Estimation

School capacity constraints are derived from the historical maximum enrollment for each school from the past decade, bounding practical capacity without new construction.[3] This sets an upper bound for the number of children that can be assigned to a simulated district, ensuring that redistricting respects facility constraints.

4.4 Data Integration

4.4.1 School District Mapping

Existing school district boundaries require careful processing due to the complex structure of American educational governance. We include only districts with assessment, demographic, and poverty status data for more than 75 percent geographic coverage by the underlying census tracts.[4] Similarly, states with less than 75 percent geographic coverage are excluded from our analysis. Those included have on average, 92.6 percent coverage.

Where elementary and secondary districts overlap, elementary districts are assigned to their corresponding unified or secondary district to avoid double-counting. In cases where only elementary or secondary districts exist, we use those boundaries and their corresponding data directly. This process yields approximately 10,500 school districts with sufficient demographic, property assessment, and spatial data for analysis.

5 Algorithmic Framework

The optimization framework employs a two-stage approach: spatial clustering for initialization, simulated annealing for refinement, followed by systematic variation across district counts. This section details the technical implementation of each stage and the mechanisms for constraint enforcement.

5.1 Stage 1: SKATER Initialization

Optimization begins with SKATER (Spatial 'K'luster Analysis by Tree Edge Removal), which generates geographically coherent initial district configurations through constrained graph partitioning (AssunÇão et al., 2006). By starting from maximally compact configurations rather than random assignments, we reduce the required iterations for convergence from hundreds of thousands (typical in redistricting literature) to approximately 2.5

---

[2] Our model optimizes for the potential tax base (assessed property value) rather than actual tax revenues, as we do not incorporate current tax rates. The assumption is that a large-scale reorganization would likely necessitate a recalibration of tax rates, making the underlying tax base the more stable and relevant metric for long-term fiscal capacity.

[3] Enrollment data are used only as a proxy for existing infrastructure capacity. We geocode schools to census tracts and use the maximum historical enrollment for each school to estimate available seats. School-age children counts from the census are used for demographic and population metrics. We multiply the capacity estimate by 1.25 as a conservative buffer, recognizing that not all resident children enroll in public schools.

[4] Coverage here refers to the spatial overlap of census tracts containing the necessary data (property assessment, demographic, and poverty) with existing school district boundaries. A small number of states and some rural areas exhibit higher rates of missing data and are thus excluded.

times the number of census tracts in each state. This reduction in computational expense allows us to complete the first school system redistricting analysis that is national in scope.

SKATER constructs a minimum spanning tree from the tract adjacency graph using edge weights based on scaled geographic coordinates. Each resulting partition forms a contiguous district, eliminating the need for post-hoc contiguity repair that plagues random initialization approaches often used in MCMC optimization.

5.1.1 Capacity Repair Mechanism

When SKATER produces initial configurations that violate the capacity constraint, a repair mechanism attempts to restore feasibility before optimization. For each violating district, tracts are evaluated for reassignment to neighboring districts with available capacity. The repair process attempts up to 50 chained explorations of 2,000 iterations each.

5.1.2 Status Quo Fallback Strategy

In cases where the SKATER initialization does not satisfy our constraints, and the chained repair strategy is unable to resolve the issue, we implement a fallback initialization approach. The system instead starts the optimization engine from the configuration of census tracts most similar to the status quo configuration of school districts. Because census tracts are not conterminous with existing school districts, we assign each tract to its geographic majority overlap district. This fallback mechanism ensures that the optimization can proceed, resolving the issue of invalid starting points.

5.2 Stage 2: MCMC Optimization

Following SKATER initialization, MCMC-based simulated annealing refines boundaries to minimize our previously defined objective function, optimizing on our criteria while maintaining all constraints. At each iteration, the algorithm selects a tract for potential reassignment. Border tracts are identified and preferentially selected, as they represent the only tracts that can change districts while maintaining contiguity. The selected tract is proposed for reassignment to a randomly chosen adjacent district.

The algorithm starts with a high temperature, its willingness to accept worse solutions, and it gradually becomes more selective over time, reducing this acceptance rate by 1 percent after each step, using the cooling formula $T(t+1) = T(t) \times 0.99$. Beginning with acceptance to any move ($T_0 = 1.0$), the process continues until it has made 2.5 times as many successful changes as there are geographic units in the redistricting plan. Each move validates compliance with the constraints outlined in section 3.3.

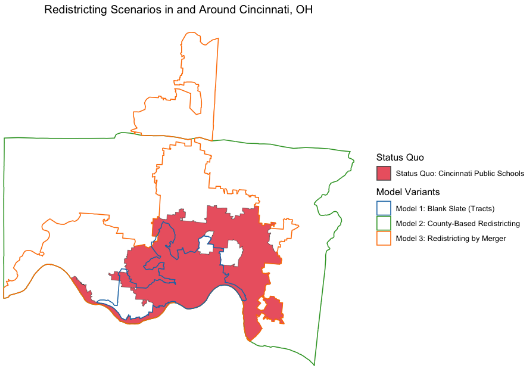

Redistricting Scenarios in and Around Cincinnati, OH
Status Quo
🟥 Status Quo: Cincinnati Public Schools
Model Variants
☐ Model 1: Blank Slate (Tracts)
☐ Model 2: County-Based Redistricting
☐ Model 3: Redistricting by Merger

Figure 1: Example plans generated by each model in and around Cincinnati.

5.3 Three Model Variants

The algorithmic framework is applied to two distinct redistricting models and a programmatic county-level merge, each offering different tradeoffs between optimization flexibility and implementation feasibility.

5.3.1 Model 1: Blank-Slate Redistricting (Tract-Level Optimization)

This model uses census tracts as atomic units, providing maximum flexibility to create optimal boundaries. This model can completely reconfigure districts without regard to existing boundaries. It establishes the theoretical frontier for equity improvements.

5.3.2 Model 2: County-Based Redistricting

This model implements a programmatic, county-based consolidation, assigning all tracts within each county to a single district. The constraints of section 3.3 are relaxed for illustrative purposes. This simulation serves as a baseline to show what simple administrative consolidation achieves versus algorithmic optimization.

5.3.3 Model 3: Redistricting by Merger (Optimized District Consolidation)

This model uses existing school districts as atomic units, preserving current boundaries while allowing mergers. The same SKATER-optimization framework operates on a district adjacency graph rather than tract-level data. This provides more politically feasible solutions that maintain district identities while still pooling resources.

Both optimization models undergo systematic variation of current district counts. Each produces tract-to-district assignments with complete Theil index calculations, enabling direct comparison of equity impacts. The tract-level model demonstrates maximum theoretical improvements. The consolidation model balances feasibility with equity gains.[5] The county benchmark validates the value of optimization over simple administrative boundaries. Figure 2 compares the per pupil tax base equity improvements in Maryland.

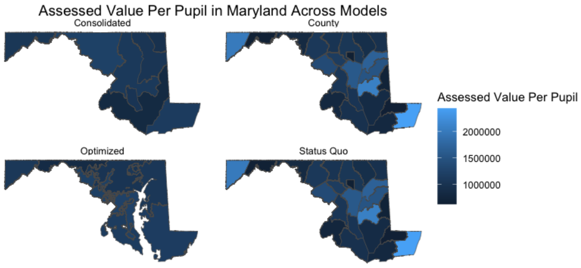


Figure 2: Per pupil tax base equity improvements in Maryland across models.

6 Output Specification

Each configuration for each model includes complete unit-to-district assignment vectors alongside a comprehensive metric suite: the three Theil indices (tax capacity, race, and poverty status), Polsby-Popper compactness scores, and additional metrics. This structure enables systematic comparison across varied numbers of districts, revealing that changing district count can improve resource distribution. The format remains consistent across all three models, which facilitates direct comparison of their relative performance.

6.1 Pareto Frontier Construction

Rather than selecting a single "optimal" solution, we identify the set of Pareto-efficient configurations that represent different trade-offs among competing objectives. A proposal is included only if no other configuration performs better on all objectives simultaneously. The Pareto selection evaluates across four criteria: minimizing the three Theil indices and maximizing geographic compactness (Polsby-Popper).

The frontier reveals critical trade-offs that cannot be resolved through optimization alone. Some configurations achieve significant funding equity but maintain racial segregation, while others integrate diverse populations at the cost of funding disparities.

6.2 Configuration Selection

While the Pareto frontier presents all efficient options, practical implementation requires selecting a single configuration. Normalized scores are combined using policy-determined weights that reflect our optimization metrics and compactness. The default weights prioritize tax base equity while maintaining focus on integration objectives and geographic coherence. The weighted score for each configuration equals the sum of each normalized measure multiplied by its corresponding weight. The configuration with the lowest weighted score is selected as the recommended plan for each state. This selection is performed only among Pareto-efficient configurations, ensuring the chosen plan is not dominated by any alternative.

The framework's key strength is its flexibility to accommodate different policy priorities. Stakeholders can adjust weights to explore how different priorities affect optimal configurations. This approach transforms a complex multi-objective optimization problem into a structured decision process. Rather than claiming to identify a single "best" solution, we provide a menu of high-quality options and a transparent mechanism for selection based on explicit policy priorities.

7 Results and Conclusion

Our strategic boundary optimization across 42 states reveals substantial potential for improving educational equity. Figure 3 demonstrates that the three models demonstrate that purposeful redrawing of district lines can significantly reduce property tax disparities in every state while simultaneously decreasing racial and economic segregation between districts.[6] As referenced in Table 1, the

---

[5] A potential critique is that aggregating smaller geographic units (like tracts or districts) into larger ones will mathematically reduce measured segregation by definition, as it averages over local variations. While this is true, the magnitude of the reduction achieved through our optimization far exceeds what would be expected from simple aggregation, demonstrating the value of purposeful boundary drawing.

[6] Figures detailing Theil improvements across states for racial and economic segregation can be found in Appendix A.

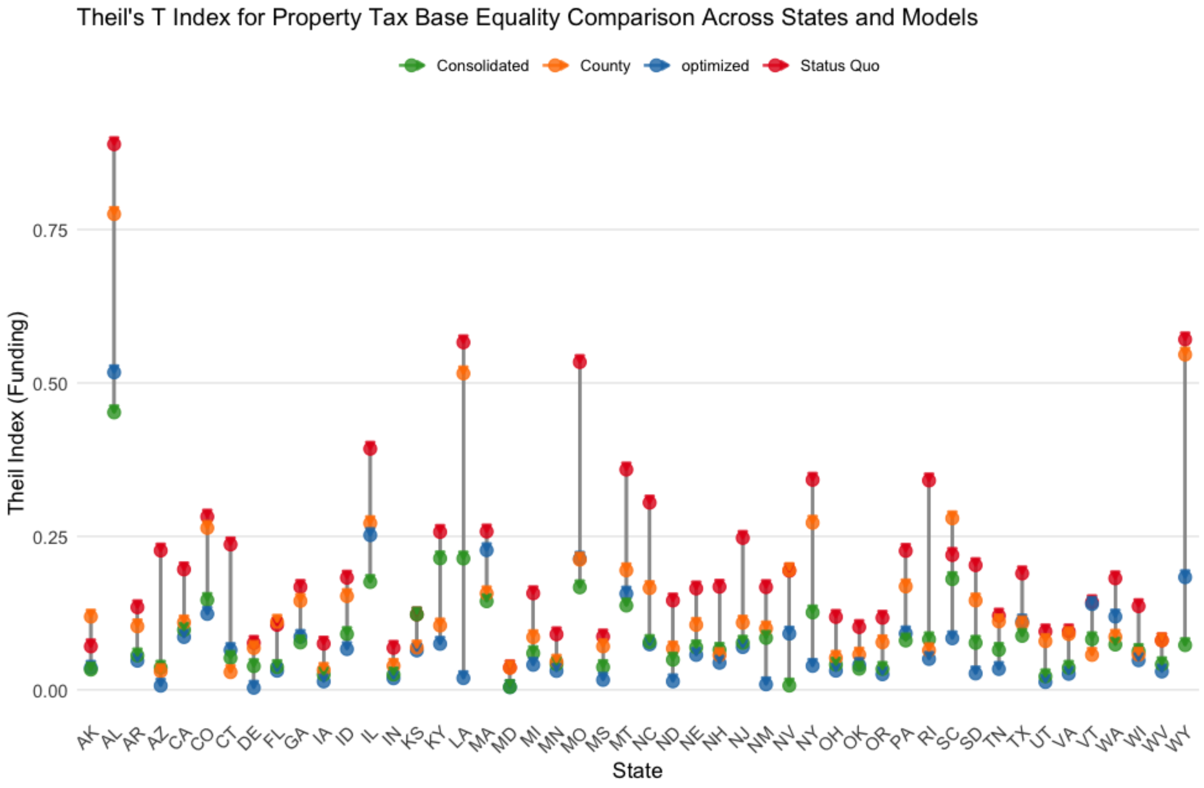


Figure 3: Theil Index improvement for property tax disparities across 42 states. This figure spans the full width of the page to show detail across all states.

Model	Property Tax Equity Improvement	Racial Integration Improvement	Economic Integration Improvement
Blank-Slate Redistricting	66.6%	47.6%	65.0%
County-Based Redistricting	39.0%	40.7%	57.2%
Redistricting by Merger	63.0%	48.2%	54.6%

Table 1: Equity Improvements by Redistricting Model. *Note:* Applied to 42 states. County-based results reflect 37 states that would see boundary changes.

Blank-Slate approach achieves average improvements of 66.6% in tax-base equality, 47.6% reduction in racial segregation, and 65.0% decrease in economic segregation. Most notably, these improvements can be achieved purely through boundary changes, without any student or family having to move. The Merger model offers a potentially more politically feasible alternative, keeping district identities intact while still delivering meaningful improvements.

While our local search approach cannot guarantee global optimality, the framework provides policymakers with concrete evidence of redistricting's potential benefits. The primary barrier to implementation remains political feasibility, as communities maintain strong attachments to existing districts. Additionally, using rook contiguity rather than actual road networks and historical enrollment maximums for capacity estimates may not capture all practical constraints like transportation barriers or current infrastructure conditions. Future work could address these limitations by incorporating dynamic demographic modeling, actual transportation networks, and mechanisms to predict post-redistricting property value adjustments.

Despite these constraints, this study establishes the first national-scale framework for school redistricting. By demonstrating that significant equity improvements are technically achievable across diverse state contexts, we provide an objective foundation for policy discussions about using boundary change as a tool for educational equity. The flexibility of our multi-objective optimization approach allows stakeholders to explore trade-offs transparently, transforming a complex challenge into a structured decision process grounded in empirical evidence.

References

R. M. AssunÇão, M. C. Neves, G. Câmara, and C. Da Costa Freitas. 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811.

Marvin E. Dodson and T. A. Garrett. 2008. Inefficient education spending in public school districts: A case for consolidation? *Contemporary Economic Policy*, 22(2):270–280.

William Duncombe and John Yinger. 2007. Does school district consolidation cut costs? *Education Finance and Policy*, 2(4):341–375.

Benjamin Fifield, Kosuke Imai, and James Olmsted. 2014. Automated redistricting simulation using markov chain monte carlo. pages 593–604.

Matthew Gartner. 2023. State education budgets and equity. *Education Policy Analysis Archives*.

Beeferman D. Vega-Pourheydarian C. Overney C. Van Hentenryck P. Roy D. Gillani, N. 2023. Redrawing attendance boundaries to promote racial and ethnic diversity in elementary schools. *Educational Researcher*, 52(6):348–364.

C. Kirabo Jackson, Rucker C. Johnson, and Claudia Persico. 2015. The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *NBER Working Paper Series*.

C. Kirabo Jackson and Claire Mackevicius. 2024. What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics*, pages 412–446.

C. Kenny and et al. 2021. The impact of the u.s. census disclosure avoidance system on redistricting and voting rights analysis. *Science Advances*.

Daphne A. Kenyon and Bethany Munteanu. 2021. State-by-state disparities in school district revenue. In *Lincoln Institute of Land Policy*.

Moritz Kuhn, Moritz Schularick, and Ulrike I. Steins. 2018. The historical process of racial and economic segregation. *Working Paper*.

Cory McCartan and Kosuke Imai. 2023. Sequential monte carlo for sampling balanced and compact redistricting plans. *Annals of Applied Statistics*, 17(4):3300–3323.

Ann Owens. 2016. Inequality in children's contexts: Income segregation of households with and without children. *American Sociological Review*, 81(3):456–488.

Sean F. Reardon and Ann Owens. 2014. 60 years after brown: Trends and consequences of school segregation. *Annual Review of Sociology*, 40:199–218.

Sean F. Reardon and Ericka Weathers. 2024. Reardon, s. f., weathers, e. s., fahle, e. m., jang, h., kalogrides, d. *American Sociological Review*, 89(06):971–1010.

Tyler Simko. 2024. School desegregation by redrawing district boundaries. *Nature Scientific Reports*.

Zahava Stadler and Jordan Abbott. 2024. Crossing the line: Segregation and resource inequality between america's school districts. *New America*.

A State-Level Theil Improvements

This appendix provides additional visualizations of the state-level improvements for racial and economic segregation metrics, complementing Figure 3 in the main text.

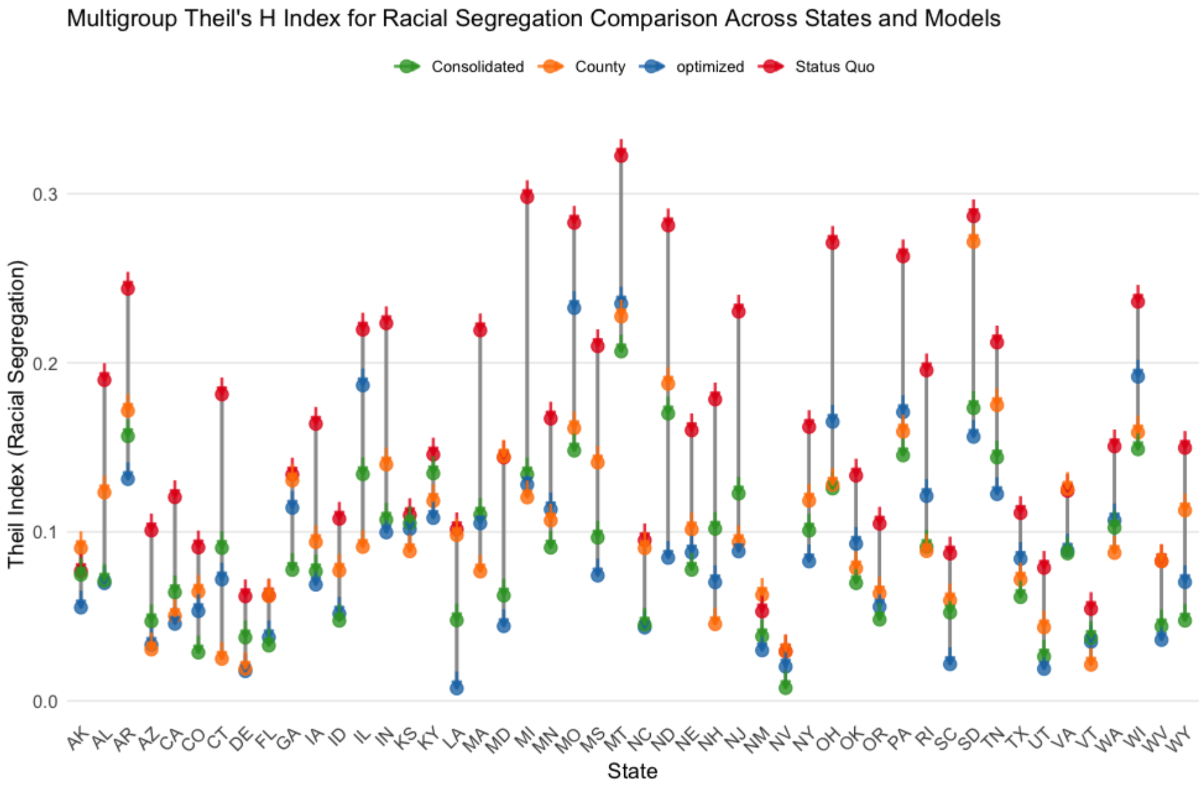


Figure 4: Theil Index improvement for racial segregation across 42 states.

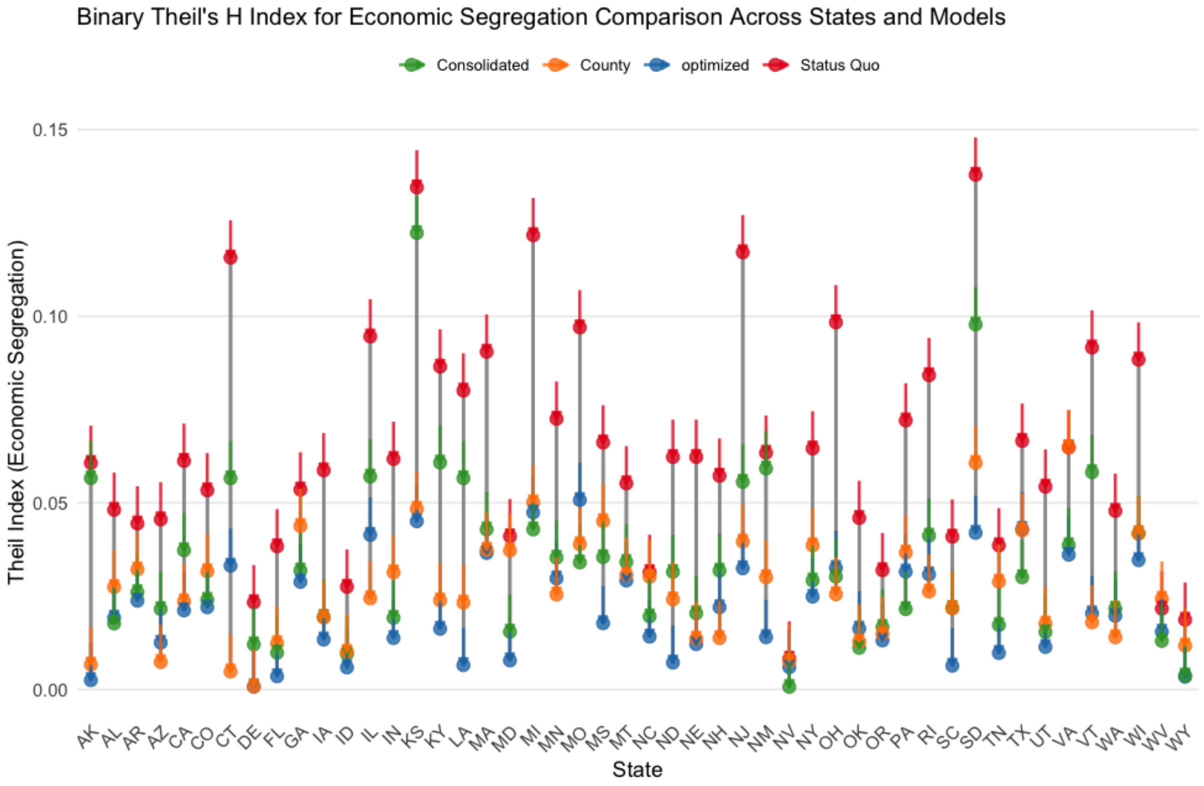


Figure 5: Theil Index improvement for economic segregation across 42 states.

Automatic Grading of Student Work Using Simulated Rubric-Based Data and GenAI Models

Yiyao Yang[1], Yasemin Gulbahar[2]

{yy3555[1], yg2918[2]}@tc.columbia.edu

Teachers College, Columbia University, New York, NY, USA

Abstract

Grading assessment in data science faces challenges related to scalability, consistency, and fairness. Synthetic dataset and GenAI enable us to simulate realistic code samples and automatically evaluate using rubric-driven systems. The research proposes an automatic grading system for generated Python code samples and explores GenAI grading reliability through human-AI comparison.

Keywords

Automated Assessment, Generative AI, Rubric-Based Grading, Inter-Rater Reliability, Human-AI Assessment Comparison, Hybrid Assessment Systems

1 Introduction

Digital technologies have significantly influenced educational assessment, leading to a growing interest in the automatic grading of student work. Traditionally, evaluating student submissions, especially coding and open-ended responses, has been labor-intensive and often subjective for educators. Manual grading tends to be inconsistent, biased, and a considerable time investment, particularly in large-enrollment courses. Automatic grading systems, primarily designed to address these challenges, aim to enhance efficiency, consistency, and objectivity in the assessment process, streamlining educational workflows and providing timely student feedback. Automatic grading is subject to the nature of text, code, and evaluating program codes might use different approaches for automated grading, where human cognition and human experience shed light on the process, and we will also assess the reliability of GenAI coding grading through a series of comparisons between human and GenAI evaluations.

Automated Grading of Student Work

The development of automatic grading systems is deeply rooted in advancements in artificial intelligence (AI), particularly in the fields of Natural Language Processing (NLP) and Machine Learning (ML) (V. Nikhil et al., 2025; Kumar et al., 2024; Alqahtani et al., 2023). While early systems relied on rule-based approaches and statistical models to analyze surface-level features such as grammar, spelling, and word count, more recently, deep learning techniques, including models like BERT and RoBERTa, have enabled more sophisticated semantic analysis, allowing systems to better understand the content and coherence of written responses (Ayaan & Ng, 2025; Faseeh et al., 2024; Bayer et al., 2022). The emergence of Large Language Models (LLMs) further promises to revolutionize this domain by offering enhanced capabilities in processing complex sentences, establishing relationships between text elements, and even discerning the intent behind writing.

The benefits of automatic grading systems are substantial and widely discussed in the literature. Foremost among these is the significant reduction in grading time and workload for instructors, freeing instructors to focus on more personalized student interactions and curriculum development (Messer et al., 2025). Automated systems also provide unparalleled consistency and objectivity, applying uniform criteria across all submissions and minimizing human biases that can inadvertently affect grades (Vetrivel et al., 2025). These advantages, which incl • ude scalability, rapid feedback, and reduced bias, have been extensively synthesized in recent systematic reviews (Marcelo Guerra Hahn et al., 2021), underscoring their central role in modern online and large-scale learning environments.

Crucially, these systems offer immediate feedback to students, a feature essential for reinforcing learning and enabling prompt self-correction. Previous research suggests that these practices lead to better long-term retention and skill development (Demszky et al., 2023; Wilson et al., 2021). Furthermore, the scalability of automatic grading makes it a crucial tool for large classes and online learning environments (Lin et al., 2024; Messer et al., 2023).

Despite these advantages, previous studies also highlight several challenges and limitations associated with automatic grading (Padó et al., 2023). A primary concern is the inherent difficulty in replicating the nuanced judgment of human graders, especially concerning subjective aspects of writing such as creativity, originality, critical thinking, and subtle rhetorical devices like humor or irony. Critics argue that over-reliance on algorithms might inadvertently incentivize students to adopt formulaic writing styles that appeal to the machine's evaluative criteria rather than foster genuine intellectual development. Moreover, while automated systems excel at quantitative feedback, they often struggle to provide the rich, qualitative, and constructive suggestions that human graders can offer, which are essential for deep learning and improvement (Figueras et al., 2025; Bato & Pomperada, 2025; Fagbohun et al., 2024).

Ethical considerations and student perception are also critical in the discourse (Farazouli, 2024). Concerns about algorithmic bias, where systems might inadvertently perpetuate or amplify existing societal biases in their training data, are frequently raised. It can lead to unfair or inequitable assessments for specific demographic groups (Baker & Hawn, 2021; Kordzadeh & Ghasemaghaei, 2021). Besides, some studies indicate that students may perceive automated grading as less fair or trustworthy than human evaluation, potentially undermining their confidence in the integrity of the scoring process (Vetrivel et al., 2025; Chai et al., 2024). Addressing these issues requires transparent system design, rigorous validation, and, in many cases, a hybrid approach that integrates human oversight (Kern et al., 2022). Looking ahead, automatic grading is moving towards advanced hybrid systems that combine the strengths of AI with human insight. Researchers focus on how AI can help evaluate more complex skills and adapt to individual learning needs, emphasizing personalized educational paths. A key development area involves integrating these grading tools with existing Learning Management Systems (LMS) to create smooth and effective educational environments. By thoroughly capturing human behavior, these systems can continuously learn from human input, leading to a more efficient, objective, and supportive learning experience for students through detailed and instant feedback, all while carefully managing the associated complexities.

Hence, with the increasing integration of data science and coding instruction into educational curricula, scalable and equitable assessment of student-generated code is becoming prominent. Assignments that involve code development, data analysis, and interpretation pose challenges for large-scale instruction due to the complexity and subjectivity in grading. Although criteria provide a standardized basis for assessment, manual scoring is time-consuming and inconsistent. It emphasizes the requirement for scalable, reliable, and pedagogy-matching grading solutions for education. GenAI and LLMs could develop the perception of code, whereas synthetic student data methods enable the recreation of student submissions in controlled environments, negating student privacy concerns. Moreover, automated grading systems can behave differently depending on the nature of the text and the machine learning approach used to evaluate. Grounded in prior studies, our research proposes an auto-grading framework that integrates GenAI with synthetic data and evaluates the reliability of automated grading in data science by examining differences between human and AI grading outcomes, intending to enhance the efficiency and effectiveness of coding assessment practices.

Hence, our study aims to address the following research questions (RQs):
RQ1: How effectively do rubric-based GenAI grading outputs align with human ratings across all programming code samples regarding total scoring consistency?
RQ2: What are the methodological strengths and practical limitations of GenAI-based rubric grading systems with respect to reliability, scalability, and fairness in programming assessment, and how can educators be guided to integrate such systems effectively into grading practices?

2 Research Methodology

Based on academic and behavioral characteristics, we used GPT-5 to generate 100 synthetic student profiles as the original dataset. These profiles include features such as the number of hours studied per day, lecture attendance rate in percentage, average quiz score, assignment score, final exam score, class participation level, number of hours of internet usage per day, and average number of sleeping hours per day. Each profile was uniquely identified by a distinct student ID.

Based on the synthetic education dataset, we have defined a regression task (supervised task) for further code samples simulation and human-AI grading comparison.

TASK: The Regression Task (Supervised Task)

This task is a supervised regression problem that aims to predict a continuous numerical value, the final exam outcome, based on various student behavioral and academic features. The goal is not just to make accurate predictions, but to build a model that can be easily understood, allowing us to identify which specific student behaviors and indicators have the most significant impact on student final scores. It is different from a classification task, which would predict a discrete category like "pass" or "fail." Instead, the response variable final exam outcome is a numerical variable, such as "92.5". When generating the AI code, you'll need to consider several key details:

Dataset: The input data is in a CSV file named synthetic_education_data.csv. The task includes loading, preprocessing, and analyzing this data.

Response Variable: The column representing the final exam outcome is the response variable we want to predict. You will need to identify this column in the dataset.

Feature Variables: The other columns containing the student behavioral and academic information are the features or independent variables. These will be used to train the model.

Model: Since the goal is interpretability, a good starting point would be models like linear regression, decision trees, or random forests. While more complex models like neural networks might be more accurate, they are often less transparent about predictions.

Evaluation Metrics: The code should use regression-specific metrics to evaluate the model's performance. Common metrics include: Mean Squared Error (MSE): Measures the average of the squared differences between the predicted and actual values. A lower MSE indicates more accurate predictions. Root Mean Squared Error (RMSE): The square root of the MSE, expressed in the same units as the response variable, making it easier to interpret. R-squared (R^2): Indicates how well the model's predictions fit the actual data, ranging in $(-\infty, 1]$, with values closer to 1 indicating that the model explains more of the variability in the outcome. A negative R^2 suggests that the model performs worse than a simple mean predictor.

Based on the task descriptions, a total of 25 synthetic Python code samples were generated by GPT-5, differing in syntax, formatting, and comments. The prompt we used was: “Could you generate 25 distinct Python solutions for the Regression Task by simulating 25 different students who have diverse levels of expertise and performance in coding skills, educational data analytics, and data mining methods? ”

Evaluation Rubric: Each code sample was assessed using a detailed analytic rubric with 20 evaluation criteria to assess a broad range of coding competencies aligned with learning outcomes in data science education: comments used, number of lines, number of libraries, number of variables, number of visualizations, error-free, clear structure, organized, data cleaning, outlier checking, optimized solution, code complexity, interpretation quality, code readability, predictable variable names, visual readability, code reusability, data accessibility, resource efficiency, and overall quality.

Each code sample received a complete rubric-based score evaluation, and the total score for each code sample was computed by summing the 20 criteria. Based on the 20 criteria, scores ranged from 1 to 5 for each criterion, and with a total score out of 100 for each code sample. GenAI grading was conducted using GPT-5 via OpenAI, guided by the 20 criteria grading scheme with human cognition to enhance reliability and evaluation alignment. To compare

GenAI and human grading results, a human grader evaluated those generated 25 code samples based on the same rubric with 20 evaluation criteria. To better understand the score distribution of GenAI versus Human grading result comparison, we visualized total scores using a multi-line radar and a scatter plot. These visualizations revealed a broad distribution of grades, supporting the diversity in the code samples simulation. Inter-rater agreement between human and AI grading results was analyzed using intraclass correlation coefficients (ICC), Cohen's Kappa, and Cronbach's α to evaluate reliability and consistency between human and AI evaluators. The research establishes a reproducible framework for rubric-based automatic code grading, incorporates realistic grading variability, and evaluates the reliability of AI-based scoring, contributing to the development of hybrid assessment systems that balance efficiency with instructional quality in STEM education.

3 Data Analysis & Results

All data analyses and Python code are stored in a private GitHub repository (*2025-NCME-AIME-Con-Yiyao-Yang*; Yang, 2025), available upon request. The summary statistics of rubric-based GenAI versus human grading across 25 submitted code samples of the regression task (Table 1) indicate a generally consistent trend, but a systematically lower scoring pattern by GenAI. Among all 25 different code samples, the mean score of GenAI grading (81.00) is lower than that of human grading (86.91), with median scores of 80.85 and 86.53, respectively. The score ranges show that both GenAI (74.02 – 91.88) and human (73.47 – 95.79) raters have captured the full spectrum of code quality, although human gradings exhibit higher variability (range = 22.32, IQR = 4.53) compared to GenAI ratings (range = 17.86, IQR = 2.63).

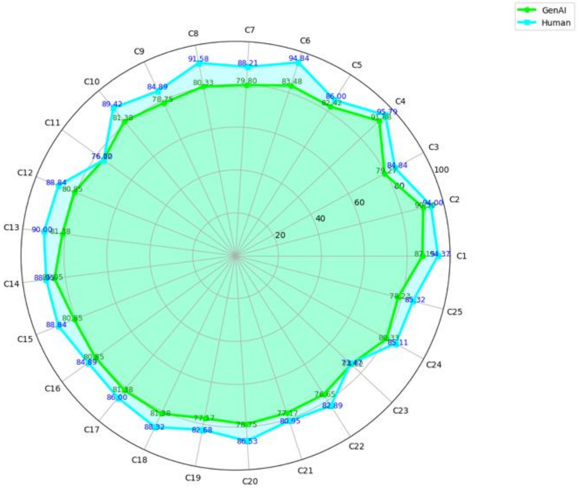
Table 1
Summary Statistics of Rubric-Based GenAI versus Human Grading of 25 Generated Code Samples

	GenAI	Human
Mean	81.00	86.91
Median	80.85	86.53
Range	17.86	22.32
IQR	2.63	4.53
Max	91.88	95.79
Min	74.02	73.47

Note. All values are reported to two decimal places.

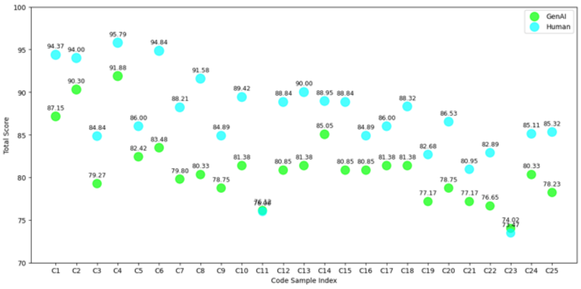
Examining the score distributions, GenAI exhibits a narrower range than human raters, reflecting compressed score variability. Across the 25 code samples, human grading generally assigned higher scores than GenAI grading, with only two tasks (Code Sample # 11 and # 23) receiving similar grades from both evaluators. Visual comparisons, including the multi-line radar plot (Figure 1) and the scatter plot (Figure 2), further confirm that while GenAI grading follows the overall performance trend, it predominantly underestimates scores relative to human evaluation.

Figure 1
GenAI versus Human Grading: Multi-Line Radar



Note. "C" denotes "Code Sample".

Figure 2
GenAI versus Human Grading: Scatter Plot



Note. "C" denotes "Code Sample".

Overall, GenAI moderately captures relative performance trends but tends to underestimate scores compared to human grading. It provides a useful foundation for consistency checks and supports the development of semi-automated grading workflows. The evaluation of rubric-based GenAI grading systems highlights both methodological strengths and practical constraints. In terms of reliability, the intraclass correlation coefficient (ICC (2,1) ≈ 0.51) indicates moderate consistency between GenAI and human grading results, suggesting that GenAI reasonably captures relative performance trends, though absolute score alignment remains limited. The low Cohen's Kappa (≈ 0.01) highlights minimal exact agreement on total scores, indicating that categorical consistency between GenAI and human grading outcomes is extremely low. In contrast, the high Cronbach's α (≈ 0.90) demonstrates strong internal consistency across rubric criteria, indicating coherent scoring patterns within the multi-criteria evaluation framework.

Regarding scalability, GenAI efficiently processes large volumes of code samples, producing rapid and reproducible scores without the temporal and cognitive variability of human raters, as a key advantage in large programming courses. For instructional alignment, the moderate total-score reliability suggests that GenAI is best used as a complement, rather than a replacement for human judgment. Educators may use GenAI for first-pass grading, trend identification, and efficient formative feedback, while maintaining human oversight for final scoring decisions. Iterative refinement of rubric prompts can further improve alignment, enabling a collaborative hybrid human-AI grading workflow.

4 Conclusion

Taken together, the findings underscore the necessity of re-evaluating grading practices in programming education. Previous research indicates that human graders often show considerable variability in scoring the same programming assignments, with both inter-rater disagreement and intra-rater inconsistency, suggesting that the notion of a "gold standard" in human grading may be inherently flawed (Messer et al., 2025). A shared rubric alone is insufficient to guarantee consistent evaluation, and additional measures such as assessor training and alternative grading practices are needed to improve reliability. In this context, our research further demonstrates that rubric-based GenAI grading offers a practical complement: While GenAI auto-grading cannot replace human judgment, it can efficiently perform an initial assessment of coding assignments, after which human evaluators can review and adjust the grading results. Such a collaborative human-AI workflow leverages the efficiency of automated scoring while preserving the refined judgment of human graders, providing an effective approach to scalable, semi-automated hybrid assessment of programming tasks. By combining the efficiency of GenAI with the experience and judgement of human evaluators, we can ensure assessment fairness while giving educators the space to guide students meaningfully in data science education, encouraging and inspiring them to grow as passionate programmers and to blossom as inquisitive learners and reflective thinkers, guided by curiosity, courage, and the joy of discovery.

5 References

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*,*19*(8). https://doi.org/10.1016/j.sapharm.2023.05.016

Ayaan, A., & Ng, K.-W. (2025). Automated Grading using Natural Language Processing and Semantic Analysis. *MethodsX*, *14*, 103395–103395. https://doi.org/10.1016/j.mex.2025.103395

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bato, B. E., & Pomperada, J. R. (2025). Automated grading system with student performance analytics. *Technium Romanian Journal of Applied Sciences and Technology*, *30*, 58–75. https://doi.org/10.47577/technium.v30i.12871

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, 3544558. https://doi.org/10.1145/3544558

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1221177

Dorottya Demszky, Jing, L., Hill, H. C., Jurafsky, D., & Piech, C. (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*, *46*(3), 016237372311692-016237372311692. https://doi.org/10.3102/01623737231169270

Fagbohun, O., Iduwe, N. P., Abdullahi, M., Ifaturoti, A., & Nwanna, O. M. (2024). Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *2*(1), 1–8. https://doi.org/10.51219/jaimld/oluwole-fagbohun/19

Farazouli, A. (2024). Automation and Assessment: Exploring Ethical Issues of Automated Grading Systems from a Relational Ethics Approach. *Postdigital Science and Education*, 209–226. https://doi.org/10.1007/978-3-031-58622-4_12

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, *12*(21), 3416. https://doi.org/10.3390/math12213416

Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., & Rossitto, C. (2025). Promises and breakages of automated grading systems: a qualitative study in computer science education. *Education Inquiry*, 1–22. https://doi.org/10.1080/20004508.2025.2464996

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., &. Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*(10), 100591. https://doi.org/10.1016/j.patter.2022.100591

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and Future Research Directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kumar, M. V., Vichitra Sivaji, B V V Padmavathi, N. Jothi, B. Muthulakshmi, & V R. Yasu Bharathi. (2024). *Natural Language Processing Techniques for Enhancing Automated Essay Scoring Systems*. 1–5. https://doi.org/10.1109/ic3tes62412.2024.10877454

Lin, Z., Yan, H., & Zhao, L. (2024). Exploring an effective automated grading model with reliability detection for large-scale online peer assessment. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12970

Marcelo Guerra Hahn, Margarita, S., de-la-Fuente-Valentín, L., & Burgos, D. (2021). A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. *IEEE Access*, *9*, 108190–108198. https://doi.org/10.1109/access.2021.3100890

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2023). Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Transactions on Computing Education*, *24*(1), 3636515. https://doi.org/10.1145/3636515

Messer, M., Neil, Kölling, M., & Shi, M. (2025). How Consistent Are Humans When Grading Programming Assignments? *ACM Transactions on Computing Education*. https://doi.org/10.1145/3759256

Nikhil, V., Annamalai, R., & Jayapal, S. (2025). NLP-Driven Approaches to Automated Essay Grading and Feedback [Review of *NLP-Driven Approaches to Automated Essay Grading and Feedback*]. In T. Murugan, K. Periasamy, & A. M. Abirami (Eds.), *Adopting Artificial Intelligence Tools in Higher Education*. Taylor & Francis Group. https://doi.org/10.1201/9781003470304-5

Padó, U., Yunus Eryilmaz, & Kirschner, L. (2023). Short-Answer Grading for German: Addressing the Challenges. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00383-w

Vetrivel, S. C., Arun, V. P., Ambikapathi, R., & Saravanan, T. P. (2025). Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments [Review of *Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments*]. In T. Murugan, K. Periasamy, & A. M. Abirami (Eds.), *Adopting Artificial Intelligence Tools in Higher Education Student Assessment*. Taylor & Francis Group. https://doi.org/10.1201/9781003470304

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208. https://doi.org/10.1016/j.compedu.2021.104208

Cognitive Engagement in GenAI Tutor Conversations: At-scale Measurement and Impact on Learning

Kodi Weatherholtz[1], **Kelli Millwood Hill**[1], **Kristen DiCerbo**[1], **Walt Wells**[1],
Phillip Grimaldi[1], **Maya Miller-Vedam**[1], **Charles Hogg**[1], and **Bogdan Yamkovenko**[1]
[1]Khan Academy,
Correspondence: kodi@khanacademy.org

Abstract

We developed and validated a scalable LLM-based labeler for classifying student cognitive engagement in GenAI tutoring conversations. Higher engagement levels predicted improved next-item performance, though further research is needed to assess distal transfer and to disentangle effects of continued tutor use from true learning transfer.

1 Introduction

Student engagement is a key predictor of learning outcomes, but not all engagement is equally beneficial. Behavioral engagement (e.g., time-on-task, number of tasks completed) captures participation, whereas cognitive engagement refers to the quality of mental effort devoted to sense-making. Theory and evidence consistently link cognitively productive behaviors–such as self-explaining, elaborating, and drawing inferences–to superior learning, beyond surface-level participation alone (Sinatra et al., 2015; Greene and Azevedo, 2010).

Measuring cognitive engagement at scale remains challenging. Conventional learning-analytics behavioral measures (e.g., time on task, retry rates) are convenient and can correlate with learning outcomes. But they provide limited insight into the metacognitive and inferential processes that underpin knowledge abstraction, integration and transfer (Winne, 2020). By contrast, text-based GenAI tutoring produces rich discourse traces of students' questions, explanations, and reasoning. These traces enable rubric-guided text mining—including LLM-as-judge approaches—to operationalize cognitive engagement directly within authentic tutoring conversations.

The present research has two main aims:

1. **At-scale labeling.** To develop and validate a reliable and scalable procedure for labeling student cognitive engagement in GenAI math tutoring conversations.
2. **Impact evaluation.** To measure the impact of different levels of *sustained* cognitive engagement on learning outcomes, focusing on near transfer to subsequent performance on the same skill. (Distal transfer to mixed-skill practice or assessments is of interest but out of scope.)

2 Background

The ICAP framework for cognitive engagement. The ICAP framework is a central account of how the depth of cognitive engagement shapes learning. This framework outlines four levels of engagement—Interactive » Constructive » Active » Passive—and predicts deeper learning as learners progress from *passive* activities (e.g., listening, reading) to *active* behaviors (e.g., highlighting/underlining, copying steps, selecting answers), to *constructive* generation (e.g., self-explaining, making inferences, creating concept maps/diagrams), and ultimately to *interactive* co-construction of knowledge with a peer (e.g., dialogue where partners question, explain, and build on each other's contributions) (Chi and Wylie, 2014).

A central prediction of the ICAP framework is that *constructive* engagement facilitates robust knowledge abstraction and transfer. The *interactive* level refers to dyads in which both peers engage constructively, generating and amplifying learning benefits for both participants. This contrasts with, for instance, a *passive-constructive* dyad, in which one partner generates constructive ideas while the other primarily listens or affirms without reciprocally elaborating or extending the reasoning. In the latter case, the constructive participant might benefit from the dialogue, but the passive participant is predicted to benefit minimally, if at all (Chi and Menekse, 2015).

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 40–48
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

The ICAP framework was initially introduced to characterize the affordances of learning activities (Chi, 2009) and has since been used extensively to analyze engagement in *peer-to-peer* contexts (e.g., (Vosniadou et al., 2023; Stump et al., 2017; Chi et al., 2017; Hsiao et al., 2022; Chi et al., 2018)). Empirical applications typically (a) compare activity types hypothesized to induce different engagement levels, or (b) use labor-intensive, human coding of classroom observations/transcripts to identify ICAP behaviors in situ. These approaches clarified which activities tend to promote deeper learning, but they pose two limitations for large-scale analytics: the unit of analysis is often coarse (activity episodes rather than specific turns of talk), and expert-level human annotation does not scale.

Adapting ICAP to student–tutor GenAI conversations. Text-based GenAI tutoring changes both the data available for analysis and the interactional context. On the data side, it produces detailed discourse traces—student turns and tutor responses—that can be examined at the level of individual contributions rather than entire activities. This granularity makes it possible to apply rubric-guided text mining approaches to authentic tutoring conversations, enabling scalable labeling methods (e.g., LLM-as-judge).

On the interactional side, GenAI tutoring differs fundamentally from peer collaboration. Whereas the ICAP framework's notion of *interactive* engagement emphasizes co-construction among peers of similar ability, GenAI tutoring is structured as a novice–expert dialogue. The GenAI tutor is, by design, a constructive partner with expert-level domain knowledge. What varies is how the learner contributes to the dialogue.

Although much attention has been given to tutor behaviors—such as eliciting reasoning, correcting misconceptions, and calibrating when to provide or withhold information (VanLehn, 2011; Koedinger et al., 2012; Graesser et al., 2012; Chi et al., 2001)—the more fundamental issue for whether and how much learners benefit from intelligent tutoring may be how learners engage in these dialogues: whether they participate passively, actively, or constructively.

Present approach and study roadmap. Study 1 adapts the ICAP framework for the student–GenAI tutor setting by developing and validating a rubric-guided, LLM-as-judge procedure for labeling student cognitive engagement at scale. Study 2 then examines how within-learner variation in cognitive engagement relates to learning, as measured by near transfer to subsequent performance on the next item attempt on the same skill (knowledge component). These analyses use student panel data (multiple tutoring conversations and corresponding item attempts per learner) and employ models with rich fixed effects to remove confounds due to time-invariant aspects of student ability, variation in skill difficulty, and local practice context.

3 Study 1: At-scale labeling

3.1 Methods

Operationalization. We extended the ICAP framework to student–tutor (novice-expert) dialogues by classifying student contributions as passive, active, or constructive (see Table 1 for a summary of operational definitions of these levels). In brief, *passive* turns reflect perfunctory acknowledgments of the tutor's information (e.g., yes/no answers) or answer-mining behavior. *Active* turns typically involve the learner mechanically following procedural guidance from the tutor without contributing new information. And *constructive* turns involve the learner adding new information by asking targeted questions, explaining reasoning, making inferences, or engaging in other behaviors that deepen the conversation and show ownership of learning.

Expert coding. Two human experts applied this rubric to a non-random coverage sample of conversations ($n = 106$), selected to reflect a diverse range of student and tutor discourse behaviors. While each student turn in a conversation can be classified independently, the expert-labeling effort focused on holistically classifying the dominant pattern of student cognitive engagement across their turns in a conversation (see examples in Table A1). After the experts labeled the conversations, we measured inter-rater reliability, resolved disagreements, and refined the rubric until high agreement was achieved.

LLM-as-judge labeling. We then converted the final rubric to an LLM-as-judge prompt (Table A2) and used OpenAI's Batch API (model snapshot = `gpt-4o-2024-05-13`) to machine-label both the human-labeled threads and a larger random set. We iterated on the prompt (e.g., few-shot examples, instruction refinements) until suitable human/LLM agreement was reached.

PASSIVE	**ACTIVE**	**CONSTRUCTIVE**	**INTERACTIVE**
The Tutor fully drives the conversation, with the Learner passively following along, **guessing, mining for the answer, or providing perfunctory inputs (e.g., ‘yes’, ‘no’, ‘idk’, ‘help’)**	The Tutor drives the conversation, but the Learner is actively involved. This often looks like the Learner **mechanically following the Tutor’s procedural guidance but without contributing new information** or without demonstrating deeper exploration / questioning.	The Learner takes ownership of the problem solving by **asking targeted questions, explaining their thinking, making inferences**, generalizations or connections to other material, or generally building on the Tutor’s guidance.	Only applies in peer-to-peer situations where learners are **co-constructing** knowledge. Does not apply when learners interact with an (ostensibly) “all knowing” other / GenAI tutor.

Table 1: Operational definitions of student cognitive engagement in GenAI tutoring conversations (adapted from ICAP framework).

Inter-rater agreement

For human-human and human–LLM rater pairs

	N threads	**Both raters agree**	**Counts**				**Krippendorff’s α**
			TP	TN	FP	FN	
Human expert #1 vs. #2							
Passive	107	93%	16	83	2	6	0.755
Active	107	85%	28	63	12	4	0.667
Constructive	107	88%	40	54	4	9	0.754
Human expert #1 vs. LLM-judge							
Passive	344	90%	142	167	23	12	0.796
Active	344	87%	84	214	25	21	0.688
Constructive	344	94%	56	266	6	16	0.796

Table 2: Summary of human/human and human/LLM-judge agreement.

3.2 Results

Inter-rater agreement is shown in Table 2. Overall agreement exceeded 84% for both human/human and human/LLM pairs. Krippendorff’s alpha ranged from 0.67 to 0.80, where values above 0.6 indicate good agreement and above 0.8 indicate very strong agreement. Human/LLM agreement was slightly higher than human/human agreement, suggesting the LLM judge performed comparably to human experts, if not slightly better.

3.3 Discussion

The LLM-as-judge approach for labeling student cognitive engagement from tutoring discourse transcripts achieved expert-level agreement while outputting structured data and rationales. This enables both high-throughput labeling and downstream analytics (e.g., linking cognitive engagement to outcomes), complementing traditional learning-analytics pipelines that rely on clickstream data and behavioral measures.

4 Study 2: Impact evaluation

4.1 Data and samples

The global sample from which we derived our causal impact analysis samples is summarized in Table 3. Specifically, we curated Khan Academy platform telemetry data and school district administrative data for students in six US public school districts who had access to Khanmigo – Khan Academy’s GenAI tutor – for the 2024-2025 SY. The platform data were limited to students (n = 9,129) with non-zero use of Khanmigo as a math tutor during the Fall and early Spring semesters (2024-08-01 to 2025-03-01). For these students, we curated question performance for all Khan Academy math items attempted during the eligibility window, totalling nearly 9 million question attempts, along with the conversation text of all Khanmigo threads that took place on those items (about 200K threads). We ran all of the tutoring conversations through a series of machine labelers – including the cognitive engagement labeler described above – to extract discourse features.[1]

4.2 Within-learner panel creation

To estimate near transfer while controlling for prior skill knowledge, we required sustained (multi-turn) threads and at least one prior and one subsequent attempt on the same knowledge component. We created two learner-level panels: Panel 1 includes learners with at least one qualifying thread in each of Passive/Active/Constructive; Panel 2 includes learners with qualifying threads in at least two of the three levels. This yields roughly 200 learners/5k threads (Panel 1) and 2k learners/35k threads (Panel 2).

Table 4 presents a descriptive summary of the panels. The rate of Constructive threads is low – an inherent property of these data. The analysis threads are by definition ‘sustained’ / multi-turn conversations: on average, there are 5 back-and-forth turns between the learner and the GenAI Tutor. The vast majority (>90%) of threads started before the learner attempted the math question (i.e., non-punitive pre-submission help); hence we expect answer correctness to be biased upwards on these items. Further, there is a non-zero rate (variable by student cognitive engagement) at which the Tutor (undesirably) reveals the final answer to the problem, which further biases correctness upwards. For knowledge transfer analysis, we examine learner performance on the next skill item. Ideally, that next item would not involve use of the Tutor. Deriving a sample with that restriction is not feasible with these data (given that about two-thirds of treatment items are followed by Tutor use). We adjust for this property of the data in the causal impact analyses; but there remains a limitation that what appears to be a treatment effect that generalizes to future performance could be explained (in part, or in full) by the future Tutor use.

4.3 Statistical methods

To examine whether deeper cognitive engagement is associated with improved learning, we modeled student performance on the next item practiced within the same skill. Specifically, we fit a fixed effects logistic regression:

$$
\begin{aligned}
\text{logit } \Pr(Y_{s,k,i+1} = 1) = \beta_1 \text{ Engagement}_{s,k,i} +& \\
\alpha_{\text{student}(s)} +& \\
\alpha_{\text{skill}(k)} +& \\
\alpha_{\text{next\_item\_had\_tutor}(s,k,i+1)} +& \\
\alpha_{\text{prev\_item\_had\_tutor}(s,k,i-1)} +& \\
\alpha_{\text{prev\_item\_credited}(s,k,i-1)} +& \\
\alpha_{\text{incoming\_skill\_mastery}(s,k)} +& \\
\alpha_{\text{tutoring\_conversation\_length}(i)} +& \\
\varepsilon_{s,i}&
\end{aligned} \tag{1}
$$

This specification models the probability that learner (s) answers the subsequent item ($i + 1$) on the same skill (k) correctly as a function of their observed engagement with the GenAI tutor on the current item (i). The key treatment variable is the level of cognitive engagement (Passive, Active, or Constructive) exhibited by the student during the tutoring exchange. By estimating a logistic regression with an extensive set of fixed effects (α terms), the model seeks to isolate within-student, within-skill variation in performance that is plausibly attributable to differences in cognitive engagement level.

The inclusion of student fixed effects ($\alpha_{student}$) accounts for all stable (time-invariant) learner characteristics that might otherwise confound the estimates, such as baseline ability, motivation, or demographic background. Skill fixed effects (α_{skill}) absorb persistent differences in the inherent difficulty of skills, ensuring that comparisons are not biased by heterogeneity in task complexity. Additional item-level controls capture contextual features of practice: whether the previous or subsequent item involved tutor use, whether the prior item was credited as correct, the learner’s incoming mastery estimate for the skill, and the number of conversational interactions within the thread. Together, these fixed effects substantially reduce between-student and between-skill variation, narrowing the identifying variation to comparisons within the same learner–skill trajectory.

The implicit assumption of this modeling approach is that, conditional on the fixed effects and included controls, variation in observed engagement level is orthogonal to other unobserved deter-

---

[1] Other labelers included deterministic labelers for linguistic complexity, reading ease, and sentiment; an LLM-judge labeler for whether the tutor revealed the final answer without the student doing the work first; and an embedding-based machine labeler for the Tutor’s degree of repetitiveness during a dialogue.

	N students	N question attempts	N question attempts with tutor use	% attempts with tutor use	% off topic	% passive	% active	% constructive
Total	9,129	8,853,983	211,229	2.4	5.5	72.2	21.6	0.7
By district								
District 1	5,834	7,971,201	173,806	2.2	5.9	74.5	19.1	0.5
District 2	1,987	268,099	12,516	4.7	4.3	62.5	32.1	1.2
District 3	257	128,332	7,403	5.8	2.5	60.7	35.5	1.3
District 4	500	108,028	6,743	6.2	1.2	63.4	33.9	1.4
District 5	216	211,523	6,202	2.9	7.1	61.5	30.0	1.3
District 6	335	166,800	4,559	2.7	1.7	58.0	35.4	5.0

Table 3: Summary of the global sample and by-district breakdown.

					Thread Characteristics						Surrounding Context					
											% Correct			Tutor Use		
ICAP Level	Learners	Threads	% of Total Threads	Skills	In a Completed Task	Median Position in Task	Median Thread Length	Median User Word Count	Thread Started Before Question Attempt	Tutor Revealed Final Answer	Prev Item	Current Item	Next Item	Prev Item	Current Item	Next Item
Panel 1: Each student has all three P/A/C levels																
Passive	191	2,257	47%	705	70%	2	5	1.0	94%	20%	71%	74%	74%	57%	100%	64%
Active	191	2,215	46%	689	80%	2	5	4.0	94%	13%	76%	83%	79%	59%	100%	66%
Constructive	191	375	8%	271	82%	3	5	10.5	95%	6%	80%	87%	82%	61%	100%	62%
Total	191	4,847	100%	954	75%	2	5	2.0	94%	16%	74%	79%	77%	58%	100%	65%
Panel 2: Each student has at least two of the three P/A/C levels																
Passive	2,116	23,278	68%	1,221	65%	2	5	1.0	93%	18%	69%	70%	69%	51%	100%	61%
Active	2,158	10,548	31%	1,158	74%	2	5	4.0	93%	11%	72%	81%	75%	53%	100%	61%
Constructive	253	503	1%	343	81%	3	5	9.2	92%	7%	79%	82%	81%	56%	100%	57%
Total	2,168	34,329	100%	1,362	68%	2	5	1.5	93%	15%	70%	73%	71%	52%	100%	61%

Table 4: Descriptive summary of analysis samples (Panel 1: learners with all three engagement levels; Panel 2: learners with at least two engagement levels). Includes thread characteristics, tutor interaction patterns, correctness, and tutor use on adjacent items.

minants of performance on the next item. In other words, after holding constant learner ability, skill difficulty, prior correctness, and local tutoring context, remaining differences in next-item correctness are attributed to differences in engagement mode. While this strengthens the rigor of inference relative to pooled or random-effects models (Angrist & Pischke, 2009), the design remains observational: unmeasured time-varying factors such as transient motivation, fatigue, or classroom context could still bias estimates.

4.4 Results

The regression results are reported in Table 5. We show cumulative stepwise addition of fixed effects to highlight how the treatment estimates changed as the amount of within-variation was controlled. Figure 1 shows the estimated marginal contrasts for the full/saturated regression model for each sample after converting the log-odds estimates to probabilities. We observe a significant stair-step effect: learners are more likely to answer the next skill item correctly (by 4-5 percentage points) after a sustained Active tutoring conversation, relative to sustained Passive, and even more likely (10ppt) after sustained Constructive.

Note that while the results trended toward Constructive > Active, the estimate for this contrast was not statistically significant. The base rate of constructive threads is very low, resulting in an imprecise measure, as seen in the wide confidence intervals in Figure 1.

4.5 Discussion

Our analyses indicate that constructive cognitive engagement with a GenAI tutor is associated with substantively better performance on subsequent skill practice, compared to passive engagement, highlighting the promise of scalable, log-based evaluation of tutor effects. Still, given the nonrandom treatment, low base rates, and outcome scope, we temper causal claims and detail limitations and

Table 5: Fixed-effect logistic regression results for the impact of cognitive engagement on knowledge transfer (i.e., correctness on next item within skill).

	Student Panel 1							Student Panel 2						
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Sustained Active	0.29 *** [0.14, 0.44]	0.30 ** [0.07, 0.52]	0.27 * [0.05, 0.49]	0.27 * [0.05, 0.49]	0.24 * [0.02, 0.46]	0.20 + [-0.03, 0.43]	0.20 [-0.05, 0.45]	0.30 *** [0.23, 0.37]	0.32 *** [0.25, 0.40]	0.29 *** [0.22, 0.36]	0.28 *** [0.21, 0.35]	0.27 *** [0.20, 0.35]	0.24 *** [0.17, 0.32]	0.27 *** [0.19, 0.34]
Sustained Constructive	0.51 ** [0.18, 0.84]	0.67 * [0.13, 1.21]	0.65 * [0.11, 1.18]	0.63 * [0.10, 1.17]	0.61 * [0.08, 1.15]	0.55 + [-0.01, 1.10]	0.72 + [-0.02, 1.46]	0.49 *** [0.21, 0.77]	0.71 *** [0.35, 1.08]	0.67 *** [0.31, 1.04]	0.65 *** [0.28, 1.02]	0.63 *** [0.26, 1.00]	0.57 ** [0.19, 0.94]	0.61 ** [0.21, 1.01]
Model statistics														
Num. Obs.	4,706	3,408	3,408	3,408	3,408	3,408	3,403	33,098	32,022	32,022	32,022	32,022	32,022	32,015
R^2	0.069	0.178	0.199	0.202	0.232	0.232	0.240	0.087	0.145	0.163	0.164	0.168	0.198	0.199
R^2 Adj.	0.000	-0.105	-0.085	-0.083	-0.054	-0.054	-0.058	-0.007	-0.001	0.017	0.018	0.021	0.051	0.050
R^2 Within	0.004	0.004	0.004	0.004	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.002	0.002	0.002
R^2 Within Adj.	0.003	0.003	0.003	0.003	0.002	0.001	0.001	0.002	0.003	0.003	0.002	0.002	0.002	0.002
RMSE	0.41	0.41	0.40	0.40	0.40	0.39	0.39	0.43	0.42	0.41	0.41	0.41	0.40	0.40
Fixed effects included (stepwise)														
FE: student (kaid)		X	X	X	X	X	X	X	X	X	X	X	X	X
FE: skill_id			X	X	X	X	X	X	X	X	X	X	X	X
FE: next_item_on_skill_used_km				X	X	X	X				X	X	X	X
FE: prev_item_on_skill_used_km					X	X	X					X	X	X
FE: prev_item_on_skill_is_credited						X	X						X	X
FE: skill_fpm_level_before						X	X						X	X
FE: n_interactions							X							X

Notes: $+ p < .10$, $* p < .05$, $** p < .01$, $*** p < .001$. Each cell shows the coefficient (top) and 95% CI (bottom). Highlighted columns are fully saturated specifications, per Equation 1.

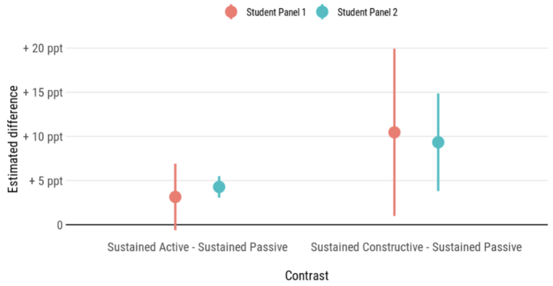


Figure 1: Estimated marginal contrasts from fixed effects regressions for the probability of answering the next skill item correct following 'sustained' Active or Constructive cognitive engagement relative to Passive engagement.

threats to validity.

Observational design. The treatment is learners' observed engagement level, not random assignment. Despite rich fixed effects that account for time-invariant within-learner and within-skill properties, unobserved confounds remain possible (e.g., transient motivation).

Subsequent tutor use. About two-thirds of treated items are followed by tutor use on the next item. Models adjust for this, but residual bias is plausible; what looks like transfer may partly reflect continued support.

Base-rate imbalance. Constructive threads are rare (<1%), limiting precision and external validity across subgroups/skills.

Outcome scope. We evaluated near transfer (next item on the same skill). Whether gains persist to mixed-skill practice or distal assessments is left for future work.

5 Conclusions

We developed a rubric-grounded LLM-as-judge that can validly classify cognitive engagement in GenAI tutor conversations at scale, enabling new forms of learning analytics that attend to the quality of student thinking. In large-scale field data, sustained Active and especially Constructive engagement were associated with higher odds of success on the next item in the same skill (near transfer), even after rigorous controls. Future work should test whether these effects extend to distal outcomes and persist without continued tutor support. Improving tutor designs to increase constructive moves is a promising path to better learning with GenAI tutors.

5.1 References

Acknowledgments

We are grateful to Michelene Chi and Kurt Van Lehn for feedback on the rubric we developed to classify student cognitive engagement in text-based tutoring interactions.

References

Michelene T. H. Chi. 2009. Active–constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1):73–105.

Michelene T. H. Chi, Joshua Adams, Elizabeth B. Bogusch, Colleen Bruchok, Seokmin Kang, Matthew Lancaster, Roger Levy, Katherine McEldoon, Glenda S. Stump, Ruth Wylie, Dongchen Xu, and David L. Yaghmourian. 2018. Translating the icap theory of cognitive engagement into practice. *Cognitive Science*, 42(6):1777–1832.

Michelene T. H. Chi, Seokmin Kang, and David L. Yaghmourian. 2017. Why students learn more from dialogue- than monologue-videos: Analyses of peer interactions. *Journal of the Learning Sciences*, 26(1):10–50.

Michelene T. H. Chi and Muhsin Menekse. 2015. Dialogue patterns that promote learning. In Lauren B. Resnick, Christa S. Asterhan, and Sherice N. Clarke, editors, *Socializing Intelligence through Academic Talk and Dialogue*, chapter 21, pages 263–274. American Educational Research Association, Washington, DC.

Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

Michelene T. H. Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.

Arthur C. Graesser, Mark W. Conley, and Andrew Olney. 2012. Intelligent tutoring systems. In Karen R. Harris, Steve Graham, and Tim Urdan, editors, *APA Educational Psychology Handbook, Vol. 3: Application to Learning and Teaching*, pages 451–473. American Psychological Association.

Jeffrey A. Greene and Roger Azevedo. 2010. The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, 45(4):203–209.

Jo-Chi Hsiao, Ssu-Kuang Chen, Wei Chen, and Sunny S. J. Lin. 2022. Developing a plugged-in class observation protocol in high-school blended stem classes: Student engagement, teacher behaviors and student–teacher interaction patterns. *Computers & Education*, 178:104403.

Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. The knowledge-learning-instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798.

Gale M. Sinatra, Benjamin C. Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13.

Glenda S. Stump, Na Li, Seokmin Kang, David Yaghmourian, Dongchen Xu, Joshua Adams, Katherine L. McEldoon, Matthew Lancaster, and Michelene T. H. Chi. 2017. Coding dosage of teachers' implementation of activities using icap: A video analysis. In *Promoting Spontaneous Use of Learning and Reasoning Strategies*. Routledge.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

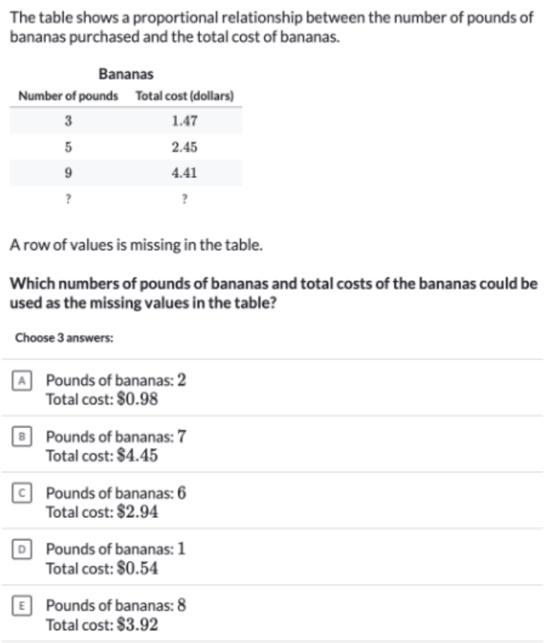
Stella Vosniadou, Michael J. Lawson, Erin Bodner, Helen Stephenson, David Jeffries, and I. Gusti Ngurah Darmawan. 2023. Using an extended icap-based coding guide as a framework for the analysis of classroom observations. *Teaching and Teacher Education*, 128:104133.

Philip H. Winne. 2020. Learning analytics for self-regulated learning. In *Handbook of Learning Analytics*, 2 edition. Society for Learning Analytics Research (SoLAR), Vancouver, BC.

A Supplementary Tables

Target math problem for conversations below.

Correct answers = A, C, and E.



The table shows a proportional relationship between the number of pounds of bananas purchased and the total cost of bananas.

Bananas

Number of pounds	Total cost (dollars)
3	1.47
5	2.45
9	4.41
?	?

A row of values is missing in the table.

Which numbers of pounds of bananas and total costs of the bananas could be used as the missing values in the table?

Choose 3 answers:

☐ A Pounds of bananas: 2
Total cost: $0.98

☐ B Pounds of bananas: 7
Total cost: $4.45

☐ C Pounds of bananas: 6
Total cost: $2.94

☐ D Pounds of bananas: 1
Total cost: $0.54

☐ E Pounds of bananas: 8
Total cost: $3.92

Student cognitive engagement	Role	Text
Active The learner makes a generic help request (passive) and then mechanically follows the tutor's procedural guidance.	Student	help
	Tutor	Let's solve this together! First, what do you notice about the relationship between the pounds of bananas and their total cost in the table? How can we use this to find the missing values?
	Student	find the total cost of 1 pound
	Tutor	Great idea! Using the table, can you calculate how much 1 pound of bananas costs?
	Student	$0.49
	Tutor	That's correct! Each pound costs $0.49. How can you use the cost per pound to find the total cost for different numbers of pounds?
	Student	multiply
	Tutor	Exactly! By multiplying the cost per pound ($0.49) by the number of pounds, you can find the total cost. Can you try this for the options given and see which ones fit the pattern?
	Student	2 x $0.49 is 0.98
	Tutor	Great job! You found that 2 pounds cost $0.98. Now, try using the same method to calculate the total cost for the other options. Which ones match the pattern?
Constructive The learner summarizes the problem, follows the tutor's guidance, errs on the first substep, gets feedback, retries while explaining their reasoning, and then solves the problem.	Student	The table says 3 pounds would be $1.47. 5 pounds cost $2.45 and 9 pounds cost $4.41. But we have to find the proportional relationship.
	Tutor	You're doing well! To find the proportional relationship, calculate the cost per pound by dividing the total cost by the number of pounds for each given data point. What do you find when you do this?
	Student	So if I divide all of the pounds by the cost it would be 2.04 but we have to find the actual cost every banana.
	Tutor	It looks like there might be a small mix-up. To find the cost per pound, you should divide the total cost by the number of pounds, not the other way around. Try calculating it again using this method. What do you find?
	Student	If I divide each of the costs by the number of pounds, it is exactly 0.49
	Tutor	Great! Now that you know the cost per pound is $0.49 per pound, you can use this to check which options maintain this consistent cost. Which options do you think match this cost per pound?
	Student	So A C E would be exactly like $0.49 because when I divide all of them it would be $0.49
	Tutor	You've got it! Options A, C, and E all maintain the consistent cost of $0.49 per pound. Great work on verifying the proportional relationship! If you have any more questions, feel free to ask.

Table A1: Example student-tutor conversations.

Instructions
You are a Critic evaluating a Student's cognitive engagement during a back-and-forth conversation with a Tutor. Your task is to assess the Student's cognitive engagement in the conversation, using the rubric provided. **Think before you respond.** Read the entire conversation first; then use the rubric to provide a holistic judgment of the student's contributions to the conversation. You must output your evaluation as a **nested JSON object** as defined below.

Cognitive Engagement Rubric

Level: Constructive – The Student is actively and constructively involved in the conversation. The Student takes ownership of problem-solving by asking targeted questions about the content, building on the Tutor's guidance, explaining their thinking, and/or making inferences, generalizations or connections to other material.
Rating: 8–10
Indicators:

- **High-specificity help requests:** The Student asks targeted questions that deepen conceptual understanding of the problem (e.g., "Why do we use a negative sign in this equation?", "How is a diamond different from a square?").
- **Substantive responses:** The Student builds on the Tutor's responses with further explanation, reasoning, or problem-solving steps (e.g., "I see, we multiply here because it's a percentage increase, not an absolute number.").
- **Deep engagement with feedback:** The Student revises their approach, while also asking follow-up questions or explaining how they modified their thinking (e.g., "I see why my approach didn't work—so I should have used the distributive property?").
- **Deep ownership of learning:** The Student actively guides the conversation by asking insightful questions, proposing alternative methods, or pointing out potential mistakes (e.g., "Could we also solve this using substitution?").

Level: Active . . .

Level: Passive . . .

Level: Off-topic . . .

How to Use the Rubric:
The rubric defines behavioral indicators for four levels of cognitive engagement (off topic, passive, active, and constructive) along a numeric scale from 1–10. Determine which level best describes the Student's contributions to the conversation. And then determine a numeric rating based on how well the conversation fits that level. The numeric ratings are not 1:1 with the levels: for example, the Passive range is from 2–4 and Active is from 5–7. These numeric ranges allow room to distinguish conversations where the Student is, for instance, exclusively Passive (rating=2) from conversations where the Student demonstrates a mixture of Passive and Active contributions (which could be rated 3–5 depending on the mixture).

Output Format:
You must output a **nested JSON object** where the key is "icap" for the ICAP Framework, and the value is an object containing:

1. `rating`: An integer rating (1–10).
2. `evidence`: A numbered list of instances from the Student's behavior in the conversation that justify the rating.
3. `off_topic_count`: An integer denoting the number of Student turns that are best classified as "off topic".
4. `passive_count`: An integer denoting the number of Student turns that are best classified as "passive".
5. `active_count`: An integer denoting the number of Student turns that are best classified as "active".
6. `constructive_count`: An integer denoting the number of Student that are best classified as "constructive".

If a type of engagement doesn't occur, still include its object: e.g., "constructive_count: 0". The sum total of `passive_count`, `active_count`, and `constructive_count` values must equal (NEVER EXCEED) the total number of Student turns in the conversation.

Output Example:

```
{
  "icap": {
    "rating": 5,
    "evidence": "(1) Student provides surface-level answers. (2) Student follows
        steps mechanically with no elaboration or reflection.",
    "off_topic_count": 0,
    "passive_count": 1,
    "active_count": 3,
    "constructive_count": 0
  }
}
```

Example Scenarios [redacted few-shot prompting]

Table A2: Complete rubric and instructions for the LLM-as-judge engagement evaluator.

Chain-of-Thought Prompting for Automated Evaluation of Revision Patterns in Young Student Writing

Tianwen Li[1], **Michelle Hong**[1],**Lindsay Clare Matsumura**[1], **Elaine Wang**[3]
Diane Litman[1,2], **Zhexiong Liu**[2],**Richard Correnti**[1]

[1]Learning Research and Development Center, University of Pittsburgh
[2]Department of Computer Science, University of Pittsburgh
[3]RAND Corporation
`{tianwen.li,mih196,lclare,dlitman,rcorrent}@pitt.edu`
`zhexiong@cs.pitt.edu`, `ewang@rand.org`

Abstract

This study explores the use of Chain-of-Thought prompting with ChatGPT-4.1 as an approach for identifying revision patterns in young students' argument writing. ChatGPT-4.1 shows substantial agreement with human coders on evidence-related revision patterns and moderate agreement on explanation-related ones. Implications for CoT prompting for writing evaluation are discussed.

1 Introduction

Revision is a crucial component of the writing process (Hayes, 1996; Fitzgerald, 1987). However, young students struggle with effective revision (Graham et al., 1995; Wang et al., 2020). A well-established approach to improving writing skills is to provide formative feedback targeting various aspects of the writing process, as it builds knowledge of the criteria for successful writing (Stein and Matsumura, 2008; Matsumura et al., 2023). Few assessments directly measure revision quality in terms of how successfully students implement feedback to meet particular writing goals. Instead, it is typically inferred from the overall quality of the revised text using holistic or trait-based scoring. However, such measures do not provide direct insight into the specific revision actions taken or into their effectiveness in meeting writing goals. Therefore, there is a need for assessments that directly capture holistic revision patterns (e.g., adding non-text-based evidence) to reveal how effectively students revise between drafts in response to specific writing goals (Correnti et al., 2024). Such assessments can serve two purposes: providing formative information to support the development of students' writing skills and providing data for research. These purposes require different validity investigations, but both involve reliability (Correnti et al., 2022).

One of the major barriers to developing and implementing direct measures of revision quality is the time-consuming nature of the evaluation process, which has been well documented in educational research. In studies that examine students' revision, researchers have relied on manual human coding to track and evaluate changes between drafts (e.g., Cho and MacArthur, 2010; MacArthur and Graham, 1987; Wang et al., 2020). The quality of revisions is often examined across four aspects: the type of revision (surface- or content-level), the type of operation (e.g., addition, deletion), the impact on meaning (meaning-changing or meaning-preserving), and the impact on text quality (increase or decrease). This is a multi-step process that is too labor-intensive for busy teachers, who would also need specialized training. Moreover, the cost of training and hiring raters to assess revision patterns in essays is prohibitively expensive for writing researchers. Therefore, efficient approaches are needed to assess revision patterns.

Recent advancements in large language models (LLMs) have shown significant promise in evaluating student writing quality (Li et al., 2024; Pack et al., 2024; Seßler et al., 2025; Tang et al., 2024; Tate et al., 2024) and can be an efficient approach to assessing revision patterns. The present study explores the potential of one of the most popular LLMs, ChatGPT 4.1, for identifying revision patterns in students' text-based argument writing. We are interested in exploring the integration of the chain-of-thought (CoT) prompting strategy to improve the performance of automatic evaluation of revision patterns. The CoT prompting is frequently used to evaluate student responses to mathematics and science tasks, as specifying the intermediate reasoning steps leading to the final answer can improve LLM performance in automatic scoring within those fields (e.g., Cohn et al., 2024; Lee

October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

et al., 2024). In contrast, prompting strategies for writing evaluation often rely on zero-shot to few-shot prompting by including scoring rubrics and sometimes related essay examples (e.g., Pack et al., 2024; Tate et al., 2024), but they rarely specify intermediate evaluation steps. Akin to prompting strategies employed in math and science evaluation, we aim to extend current research by exploring whether CoT can improve ChatGPT-4.1's performance in assessing revision quality.

To address this goal, this study applies two different prompting strategies: the baseline strategy of few-shot prompting and few-shot CoT prompting. We then examine the reliability of ChatGPT-4.1's predictions from two perspectives: the internal consistency in generating the same output across multiple runs and the accuracy in predicting revision patterns that align with human coding. The reliability of automated scores has most often been evaluated by comparing machine-generated ratings with human ratings, which are often considered the gold standard. We extend the investigation of reliability by examining the consistency of ChatGPT-4.1's ratings. Internal consistency has received less attention in prior research, yet it is important given evidence that LLMs can produce inconsistent ratings (Tang et al., 2024; Tran et al., 2024). Therefore, this paper addresses the following questions:

1. How internally consistent is ChatGPT-4.1 in assessing revision patterns in students' text-based argument writing across different prompting strategies?
2. How accurate is ChatGPT-4.1 in assessing revision patterns in students' text-based argument writing across different prompting strategies?

2 Data

In this section, we describe the dataset of students' essays, outline the taxonomy of revision patterns used to code revision quality, and explain the human coding process.

2.1 Dataset of student essays

The corpus for this study is drawn from a larger project of eRevise+RF, an automated writing evaluation system designed to support young students' argument writing and revision (Correnti et al., 2024; Liu et al., 2023; Liu et al., 2025). eRevise+RF is developed to score responses and provide feedback to students on the Response-to-Text Assessment (RTA). The RTA aims to assess the quality of students' ability to reason about texts in their writing and to use text evidence to support their claims (Correnti et al., 2012; Correnti et al., 2013). To administer the system, the teacher reads the text aloud to students, poses planned questions, and defines selected vocabulary at specific points in the article to ensure that all students comprehend the material before writing. In this study, each student completed one of two RTAs: one task was based on an article about the United Nations Millennium Villages Project (MVP) to fight poverty in Kenya, and another one was based on an article about the benefits and costs of space exploration (SPACE) (Appendix A).

After students submit their first drafts, the system uses NLP features generated during the automatic scoring of students' initial essays (including the number of pieces of evidence, the specificity of evidence, the concentration of evidence, and word count) to select appropriate revision goals and related feedback based on the quality of evidence in each draft. Revision goal 1 emphasizes adding additional evidence from the text. Revision goal 2 instructs students to add details to existing evidence to increase specificity. Revision goal 3 guides students to explain their evidence and connect it to the claims (Correnti et al., 2020; Wang et al., 2020). After receiving a tailored revision goal, students revise their essays accordingly.

The dataset was collected from 330 students in grades 4th through 8th in Louisiana and Pennsylvania. It contains a total of 330 essay pairs, including both initial and revised drafts. Among these essay pairs, 172 were written in response to the MVP article, while 158 were written in response to the SPACE article.

2.2 Taxonomy of revision patterns for argument writing

The taxonomy of revision patterns for argument writing is adapted from Wang et al.'s (2020) qualitative study, which examined how students revised their writing in response to the aforementioned revision goals and the feedback generated by eRevise, the earlier version of eRevise+RF. The revision patterns identified by Wang et al. (2020) were reorganized and consolidated around three guiding questions: 1) Do the revisions focus on content? 2) Do the revisions effectively address the targeted goal? and 3) To what extent do the revisions sub-

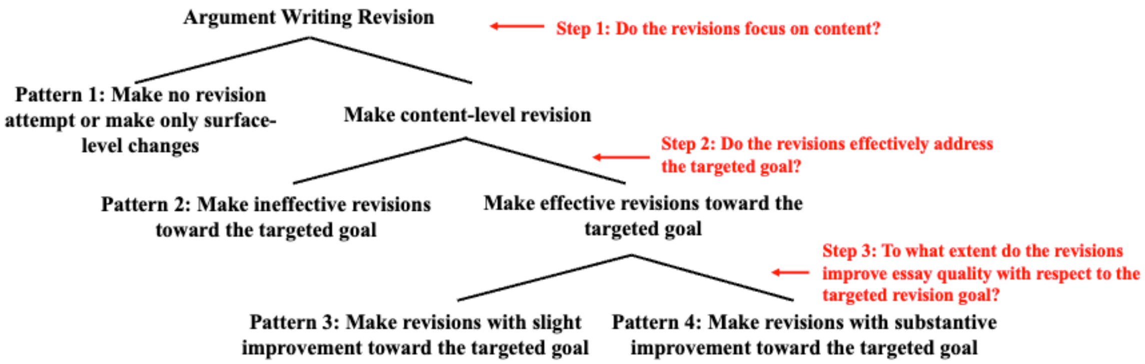

Argument Writing Revision
Step 1: Do the revisions focus on content?
Pattern 1: Make no revision attempt or make only surface-level changes
Make content-level revision
Step 2: Do the revisions effectively address the targeted goal?
Pattern 2: Make ineffective revisions toward the targeted goal
Make effective revisions toward the targeted goal
Step 3: To what extent do the revisions improve essay quality with respect to the targeted revision goal?
Pattern 3: Make revisions with slight improvement toward the targeted goal
Pattern 4: Make revisions with substantive improvement toward the targeted goal

Figure 1: Taxonomy of revision patterns. This taxonomy reflects the general framework for categorizing revision patterns. The manifestations of revision patterns for each goal are presented in Appendix B.

stantially improve essay quality with respect to the targeted revision goal? Based on these questions, four main revision patterns were generated and organized into a taxonomy, as shown in Figure 1. The four revision pattern codes reflect a progression from the least to the most effective type of revision patterns, which represents an ordinal scale. In addition, these revision pattern codes are mutually exclusive, which means coders should assign only one code to each essay pair. As students were assigned different revision goals, the description of each revision pattern for each goal and related examples are presented in detail in Appendix B.

2.3 Human coding of revision patterns

Assessing revision quality is a complex process. To increase interrater reliability between two human coders, we standardized the coding procedure by breaking down the task of assigning revision patterns into a series of manageable steps developed from the three guiding questions (Figure 1). All changes between the first and second drafts were counted as revisions. The procedure was structured as a three-step sequential filtering approach, with each step building on the previous one to progressively focus the analysis on a smaller, more meaningful set of revisions:

Step 1. Examine all revisions to determine whether any content-level changes are presented. If all changes are surface-level (e.g., mechanical issues of writing), the revision pattern is Pattern 1: Make no revision attempt or make only surface-level changes.

Step 2. Further examine the revisions identified as content-level to determine whether those revisions effectively address the targeted goal. If none of the revisions are effective, the revision pattern is Pattern 2: Make ineffective revisions toward the targeted goal.

Step 3. Focus on the revisions that effectively address the revision goal, and consider both the quantity of these revisions and the overall quality of the first draft to determine whether there is a substantive improvement in overall essay quality. This step aims to select whether the revision pattern should be Pattern 3 of making revisions with slight improvement or Pattern 4 of making revisions with substantive improvement. This three-step coding process was developed into a decision-making flowchart presented in Appendix C.

All the essay pairs were double-coded by two human coders. Discrepancies were discussed and resolved between coders to establish the benchmark for the comparison with coding by ChatGPT-4.1. We assessed interrater reliability with two metrics: exact agreement, calculated using confusion matrices, and quadratic weighted kappa (QWK). The interpretation of Kappa follows the guideline proposed by Landis & Koch (1977): values below 0 indicate poor agreement; 0.01–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1, almost perfect agreement. Human coders engaged in identifying revision patterns for each of the three goals (i.e., evidence, details, and explanations). For the revision goal of adding evidence, the exact agreement between the two coders was 87.5%, with a QWK of 0.87, indicating almost perfect agreement. For the revision goal of adding details, the exact agreement was 84.91%, and the QWK was 0.91, indicating almost perfect agreement. For the final revision goal of adding explanations, the exact agreement was 79.67%, and the QWK was 0.77, indicating

	Goal 1: Evidence	**Goal 2: Details**	**Goal 3: Explanations**
Pattern 1: Make no revision attempt or make only surface-level changes	2	3	19
Pattern 2: Make ineffective revisions toward the targeted goal	14	22	60
Pattern 3: Make revisions with slight improvement toward the targeted goal	44	17	32
Pattern 4: Make revisions with substantive improvement toward the targeted goal	74	17	26
Total	134	59	137

Table 1: Distribution of revision patterns for each revision goal

substantial agreement. The distribution of human-coded revision patterns for each revision goal is shown in Table 1.

3 Methods

We used ChatGPT-4.1 to assign revision pattern codes to each essay pair (i.e., first and second draft). To evaluate the model's internal consistency in assessing revision quality, each essay pair was processed three times using the same prompt. To minimize randomness in the output, the temperature was set to 0. Two types of prompts were tested in this study: few-shot prompting and few-shot CoT prompting. In the following section, we provide an overview of these prompting methods (see Appendix D for details).

3.1 Few-shot prompting

Few-shot prompting served as the baseline, in which ChatGPT-4.1 was directly asked to assign one of the revision pattern codes to each essay pair. The prompt consisted of three parts. First, it introduced the RTA by providing the source text and the related writing prompt. Second, it specified the student's revision goal and presented the list of revision pattern codes associated with that goal. We also include the descriptions of each code and an essay pair to exemplify the pattern. Finally, the student's first and second drafts were provided, and the model was prompted to output one revision pattern code.

3.2 Few-shot chain-of-thought prompting

The CoT prompting approach was developed based on the human coding process described in the previous section. While most of the information included in the baseline prompt remained the same, the only change was substituting the list of revision patterns with stepwise guidance for selecting a pattern. Given the three revision goals, more detailed instructions tailored to each goal were developed.

3.3 Evaluation metrics

To address RQ1, we ran each type of prompt three times to evaluate ChatGPT-4.1's consistency in generating the same code. We then calculated the QWK for each pairwise comparison and averaged the results to determine internal consistency. To address RQ2, we used a majority vote approach to determine the final code assigned by ChatGPT-4.1 across the three runs and computed exact agreement and QWK to evaluate the agreement between ChatGPT-4.1 coding and human coding. We further computed per-class F1 scores for each revision goal, which enabled a direct comparison between baseline prompting and CoT prompting in their ability to identify different revision patterns.

4 Results & Discussion

4.1 RQ1: Consistency of revision pattern predictions

Both the baseline prompting strategy and the CoT prompting strategy exhibited almost perfect consistency across the three runs, with QWK values exceeding 0.90 for each revision goal (Table 2). This finding demonstrates that ChatGPT-4.1 is highly reliable in assigning the same revision patterns to essays when prompted multiple times. Moreover, consistency remained stable across prompting strategies, suggesting that the internal consistency of ChatGPT-4.1 in assessing revision quality is not affected by different prompting strategies.

	Baseline		CoT	
	Exact Agreement	QWK	Exact Agreement	QWK
Goal 1: Evidence	90.30%	0.94	90.30%	0.94
Goal 2: Details	89.83%	0.96	88.14%	0.94
Goal 3: Explanations	89.78%	0.94	92.70%	0.96

Table 2: Internal consistency across two prompting strategies

	Baseline		CoT	
	Exact Agreement	QWK	Exact Agreement	QWK
Goal 1: Evidence	64.18%	0.56	64.93%	0.63
Goal 2: Details	67.80%	0.76	61.02%	0.80
Goal 3: Explanations	40.88%	0.39	52.55%	0.54

Table 3: Prediction accuracy across two prompting strategies

4.2 RQ2: Accuracy of revision patterns predictions

As shown in Table 3, with the baseline prompting strategy, we observed an exact agreement of 64.18% and a QWK of 0.56 for revision goal 1, indicating moderate agreement with human coding. For revision goal 2, the exact agreement was 67.80% with a QWK of 0.76, indicating substantial agreement with human coding. For revision goal 3, the exact agreement was 40.88% with a QWK of 0.39, indicating only fair agreement with human coding.

We further tested the CoT prompt, and the results showed that it improved accuracy in predicting revision patterns. For goal 3, the exact agreement rose from 40.88% to 52.55%, and the QWK increased from 0.39 to 0.54, representing a moderate level of agreement. For goal 1, the QWK slightly increased from 0.56 to 0.63, indicating substantial agreement with human coding, while the agreement remains substantial for revision goal 2.

Across both prompting strategies, revision goal 3 (adding explanations) consistently showed the lowest accuracy in predicting revision patterns. Similarly, in human coding, interrater reliability was lowest for revision goal 3 (QWK=0.77). Assessing the quality of newly added explanations (goal 3) is more subjective than assessing the quality of added evidence (goal 1) or related details (goal 2). The quality of evidence can be directly checked against the source text. By contrast, explanations of how evidence supports claims vary widely in length, clarity, logic, and persuasiveness, which requires more nuanced judgment. Simply including the definition of revision patterns in the prompt does not capture the judgments made by humans during coding the quality of explanations. As a result, ChatGPT-4.1 struggled to assign revision patterns that aligned with human coding when working on goal 3. This finding suggests that the performance of LLMs co-varies with the level of human agreement in coding educational materials (Cohn et al., 2024; Wang et al., 2023). In other words, when human coders demonstrated higher interrater reliability, ChatGPT-4.1 also achieved higher accuracy in predicting revision patterns.

We further computed F1 scores to gain insight into how baseline prompting and CoT prompting performed differently on identifying revision patterns, with particular attention to revision goal 3. As shown in Table 4, the CoT prompt increased the weighted-average F1 score from 0.40 to 0.51. When examining the per-class F1 scores for each revision pattern, the CoT improved performance in predicting ineffective revisions of explanation (pattern 2), achieving an F1 score of 0.63, nearly double that of the baseline model (F1 = 0.34). This higher score reflects CoT's ability to capture more true instances of ineffective revisions while reducing misclassifications of other revision types as ineffective. This is an important improvement in prediction accuracy, as the evaluation of ineffective

	Goal 1: Evidence		Goal 2: Details		Goal 3: Explanations	
	Baseline	CoT	Baseline	CoT	Baseline	CoT
Pattern 1	0.40	0.29	0.67	0.75	0.52	0.62
Pattern 2	0.46	0.46	0.63	0.73	0.34	0.63
Pattern 3	0.52	0.58	0.63	0.57	0.42	0.24
Pattern 4	0.74	0.74	0.78	0.77	0.44	0.48
Weighted average F1 score	0.63	0.65	0.68	0.70	0.40	0.51

Table 4: F1 scores across two prompting strategies

explanation revision is the most complex in the human coding process. This complexity arises primarily from the pedagogical knowledge required to recognize the diverse forms of ineffective explanation attempts. Instead of adding explanations that clearly connect evidence to claims, young students often insert personal comments, empty explanations, summaries of the evidence, or elaborations that do not strengthen the argument. Moreover, students' ineffective explanations are not always presented in a single pattern; rather, they frequently appear as a mix of multiple inadequate attempts in their revision, sometimes even accompanied by partial but effective explanations. With the baseline prompt, when an essay contained both effective and ineffective revisions of exlanations, the selection of a revision pattern often appeared arbitrary, as no clear major pattern emerged. By contrast, with the CoT prompt, ChatGPT-4.1 was instructed to evaluate the quality of explanation revisions first at the sentence level and then transition to the essay level by considering the quantity of effective explanations shown within the revision. The inclusion of standardized evaluation steps in the CoT prompting, which makes explicit the considerations human coders apply during coding, likely contributed to the accuracy of identifying the revision pattern that applied ineffective explanations.

5 Conclusions

Revision is a crucial component in writing development, yet many young students struggle to revise effectively (Wang et al., 2020). Accurately evaluating the revision quality (e.g., identifying revision patterns) is a key step in providing targeted feedback that supports the growth of their revision skills. However, such evaluation is time-consuming for human coders. Therefore, this study investigates the potential of ChatGPT-4.1 as an alternative tool for identifying revision patterns across various writing goals. Our findings demonstrate that ChatGPT-4.1 is highly consistent in predicting the same revision patterns across multiple runs and shows strong potential for effectively identifying patterns that align with human coders. Similar to studies that explore CoT prompting in the automated scoring of math and science tasks (e.g., Lee et al., 2024), we also found that including intermediate evaluation steps improves the accuracy of predicting revision patterns, particularly those under the goal of adding explanations. Specifying evaluation steps makes the nuanced judgments of human coders more explicit, which likely contributed to this improvement. Moving forward, we suggest that researchers and teachers carefully reflect on and document their writing evaluation processes, standardize these steps, and transform them into a sequence of manageable subtasks or decision points. Such practices may better support collaboration with LLMs in scoring tasks more broadly.

6 Limitations

First, our study focused on a specific writing evaluation task of assessing revision quality among young students. Future research should apply CoT prompting strategies across diverse writing evaluation tasks, such as holistic scoring or trait-based scoring in different writing genres to examine whether CoT can outperform baseline models. Second, the assessment of revision introduced in this study is designed primarily for the purpose of providing feedback by teachers in the classroom, and we only focused on testing the reliability of ChatGPT-4.1 scoring. Although we demonstrated its potential in identifying revision patterns, future research should investigate the validity of the assessment to ensure that it captures the meaningful dimensions of student revision or develop a more comprehensive format based on it.

Acknowledgments

The research was supported by the National Science Foundation Award #2202347. The opinions expressed are those of the authors and do not represent the views of the foundation.

References

Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learning and instruction*, 20(4):328–338.

Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 23182–23190.

Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

Richard Correnti, Lindsay Clare Matsumura, Laura S Hamilton, and Elaine Wang. 2012. Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment*, 17(2-3):132–161.

Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, Diane Litman, Zahra Rahimi, and Zahid Kisa. 2020. Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, 55(3):493–520.

Richard Correnti, Lindsay Clare Matsumura, Elaine Lin Wang, Diane Litman, and Haoran Zhang. 2022. Building a validity argument for an automated writing evaluation system (erevise) as a formative assessment. *Computers and Education Open*, 3:100084.

Rip Correnti, Elaine Lin Wang, Lindsay Clare Matsumura, Diane Litman, Zhexiong Liu, and Tianwen Li. 2024. Supporting students' text-based evidence use via formative automated writing and revision assessment. In *The Routledge international handbook of automated essay evaluation*, pages 221–243. Routledge.

Jill Fitzgerald. 1987. Research on revision in writing. *Review of educational research*, 57(4):481–506.

Steve Graham, Charles MacArthur, and Shirley Schwartz. 1995. Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology*, 87(2):230.

Andrew F Hayes. 1996. Permutation test is not distribution-free: Testing h: ρ= 0. *Psychological Methods*, 1(2):184.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. 2024. Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 365–380, Mexico City, Mexico. Association for Computational Linguistics.

Zhexiong Liu, Diane Litman, Elaine Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. erevise+ rf: A writing evaluation system for assessing student essay revisions and providing formative feedback. *arXiv preprint arXiv:2501.00715*.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. Predicting the quality of revisions in argumentative writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.

Charles A MacArthur and Steve Graham. 1987. Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of special education*, 21(3):22–42.

Lindsay Clare Matsumura, Elaine Lin Wang, Richard Correnti, and Diane Litman. 2023. Tasks and feedback: An exploration of students' opportunity to develop adaptive expertise for analytic text-based writing. *Assessing Writing*, 55:100689.

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.

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.

Mary Kay Stein and Lindsay Clare Matsumura. 2008. Measuring learning for teacher instruction. *Measurement Issues and Assessment for Teaching Quality*, page 179.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14).

Tamara P Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can ai provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, Lindsay Clare Matsumura, et al. 2024. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2):304–335.

Elaine Lin Wang, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. erevis (ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44:100449.

Rose E Wang, Pawan Wirawarn, Noah Goodman, and Dorottya Demszky. 2023. Sight: A large annotated dataset on student insights gathered from higher education transcripts. *arXiv preprint arXiv:2306.09343*.

A RTA writing task

MVP writing task: The author described how the quality of life can be improved by the Millennium Villages project in Sauri, Kenya. Based on the article, did the author convince you that "winning the fight against poverty is achievable in our lifetime"? Explain why or why not with 3 to 4 examples from the text to support your answer.

SPACE writing task: Consider the reasons given in the article for why we should and should not fund space exploration. Did the author convince you that "space exploration leads to long-term benefits" that justify the cost? Give reasons for your answer. Support your reasons with 3 to 4 pieces of evidence from the text.

B Revision Pattern Codes

Revision Goal 1: Adding more text-based evidence

1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level but uses ineffective revision strategies that fail to improve the quality of evidence used in the essay. These strategies include adding explanations instead of adding new evidence and adding new evidence that is not based on the text.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health...". Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health...". Therefore, that new technique can be used to more people and help them to be healthy.

3. Make revisions with slight improvement toward the targeted goal: The revision involves adding only one piece of evidence or adds a list of keywords of various evidence in the second draft; however, the draft as a whole still contains only two or fewer pieces of evidence.

Example: Draft 1: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth. Draft 2: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable. Space exploration can even solve problems. It can also monitor land, track corps, stop wars, provide motivations to nations.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding more than one piece of evidence from different parts of the text to support the main claim.

Example: Draft 1: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth. Draft 2: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth. Another reason why space exploration is desirable is how scientist use monitors to check astronauts health. My next reason is, in addition ,the race led to significant investment and progress in American education ,especially in math and science. this shows that by looking outward into space,we also improved life here on earth. Finally, Over 46.2 million Americans (15%) live in poverty. The investment in space exploration argue that 19 billion dollars is not to much.

Revision Goal 2: Adding details to the pieces of evidence used in the essay

1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level but uses ineffective revision strategies that fail to improve the specificity of evidence used in the essay. These strategies include adding explanations instead of adding new evidence, and adding new evidence and details that are not based on the text.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health...". Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health...". Therefore, that new technique can be used to more people and help them to be healthy.

3. Make revisions with slight improvement toward the targeted goal: The revision involves adding details to only one piece of evidence in the second draft or introduces a new piece of evidence with limited detail. However, the rest of the evidence in the second draft continues to lack specificity and details.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have solve hunger and poverty. These include better exercise machines, better airplanes, and better weather forecasting." That is why I believe that we should find space exploration more desirable. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have solve hunger and poverty." For example, the text states that Satellites that circle Earth can monitor land and the atmosphere. They can track and measure the conditions of crops, soil, and rainfall. We can use this information to improve the way we produce and distribute food. That is why I believe that we should find space exploration more desirable.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding details to more than one piece of evidence in the second draft, so the evidence becomes more specific.

Example: Draft 1: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth. Draft 2: I am convinced that space exploration is desirable because space exploration helps us remain a creative society. It makes us strive for better technologies and scientific knowledge. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth. Another reason why space exploration is desirable is how scientist use monitors to check astronauts health. My next reason is, in addition ,the race led to significant investment and progress in American education ,especially in math and science. this shows that by looking outward into space,we also improved life here on earth. Finally, Over 46.2 million Americans (15

Revision Goal 3: Explain the evidence and connect to the claims 1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level; however, the changes do not improve the quality of the explanation of how the evidence supports the claim. Ineffective revision strategies include adding new evidence but not addressing the revision goal of adding the explanation, adding personal comments instead of explaining, providing empty explanations, paraphrasing existing evidence without explaining how the evidence supports the claim, or elaborating on the evidence without explaining how the evidence supports the claim.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.

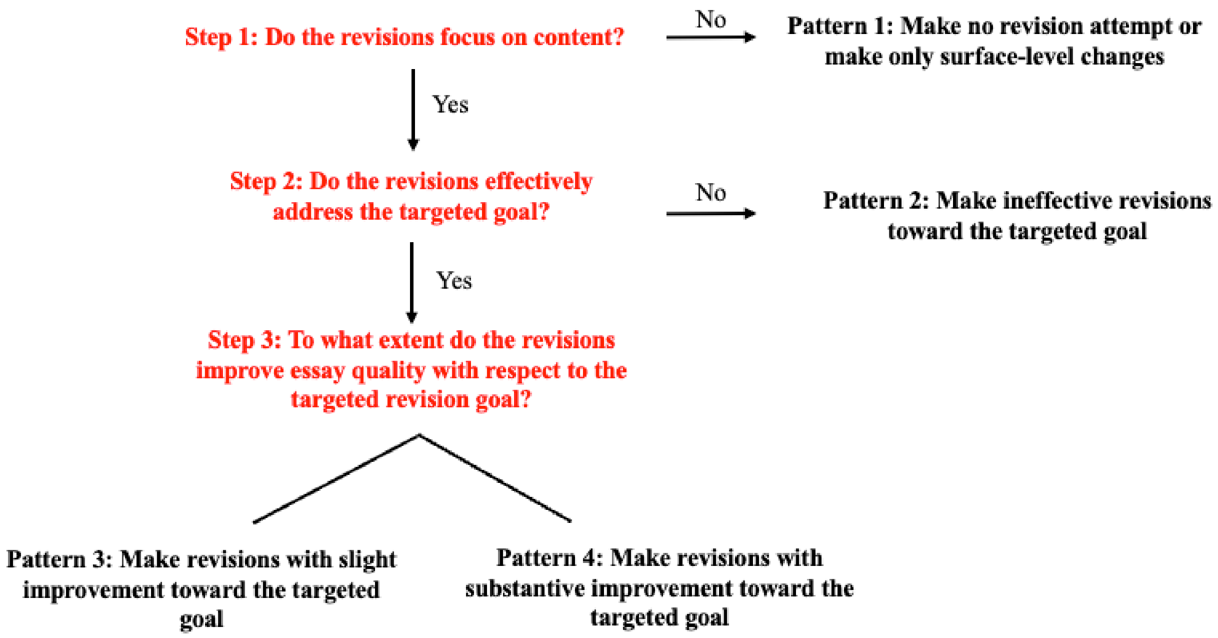
3. Make revisions with slight improvement toward the targeted goal: The revision involves adding a brief explanation to one piece of evidence to show how it supports the main claim or reuses the same explanation for multiple pieces of evidence. However, how each distinct piece of evidence supports the claim remains unclear, and some evidence may be left unaddressed.

Example: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding multiple explanations to existing evidence, clearly showing how the various pieces of evidence support the claim, thereby making the argument more persuasive.

Example: Draft 1: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. Draft 2: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. so that means that they need money to have the stuff to look and see what is going to happen in the future and there is a machine in space to see what the weather is going to be so they need money for that. It is important because like what is there is a tornado unexpected so they will not know how cold or what is going to happen there might be snow coming and we do not know. (a strong explanation) Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. They also need money to have satellite see if we did not have a satellite we would not know when a tornado would come. I think we should keep giving them money because they are keeping us safe by making a satellite and telling us on the news so we can get the info so we should keep giving they money.

C Human Coding Flowchart


Step 1: Do the revisions focus on content?
No
Pattern 1: Make no revision attempt or make only surface-level changes
Yes
Step 2: Do the revisions effectively address the targeted goal?
No
Pattern 2: Make ineffective revisions toward the targeted goal
Yes
Step 3: To what extent do the revisions improve essay quality with respect to the targeted revision goal?
Pattern 3: Make revisions with slight improvement toward the targeted goal
Pattern 4: Make revisions with substantive improvement toward the targeted goal

D Prompt Designs for ChatGPT-4.1 (Using Goal 3 as an Example)

Baseline prompt Students from 4th grade to 8th grade are currently working on text-based argument essays. After submitting their first draft, students received feedback and revised their second draft. Here is the text and writing task:

<Text + Writing Prompt>

You are a writing teacher who works with students from 4th grade to 8th grade. Students are working on the revision goal of explaining how the evidence supports the main claim. Your task is to compare the first and second drafts to identify the major revision patterns in the students' work in response to the revision goal. Below are four revision patterns; select the one that best represents the student's revision.

1. Make no revision attempt or make surface-level revision: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions. If changes in the second draft exceed 35 words, it is not considered a surface-level revision. Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make revisions with ineffective strategies toward the targeted revision goal: The student attempts a content-level revision; however, the changes do not improve the quality of the explanation of how the evidence supports the claim. Ineffective revision strategies include adding new evidence but not addressing the revision goal of adding the explanation, adding personal comments instead of explaining, providing empty explanations, paraphrasing existing evidence without explaining how the evidence supports the claim, or elaborating on the evidence without explaining how the evidence supports the claim. Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.(personal comments instead of explaining how existing evidence supports the claims).

3. Make slight improvement toward the targeted revision goal: The student adds a brief explanation to one piece of evidence to show how it supports the main claim or reuses the same explanation for multiple pieces of evidence. However, how each distinct piece of evidence supports the claim remains unclear, and some evidence may be left unaddressed. Example: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy.

4. Make substantive improvement toward the targeted revision goal: The student adds multiple explanations to existing evidence, clearly showing how the various pieces of evidence support the claim, thereby making the argument more persuasive. Example: Draft 1: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. Draft 2: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. so that means that they need money to have the stuff to look and see what is going to happen in the future and there is a machine in space to see what the weather is going to be so they need money for that. It is important because like what is there is a tornado unexpected so they will not know how cold or what is going to happen there might be snow coming and we do not know. (a strong explanation) Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. They also need money to have satellite see if we did not have a satellite we would not know when a tornado would come. I think we should keep giving them money because they are keeping us safe by making a satellite and telling us on the news so we can get the info so we should keep giving they money (a strong explanation).

Output one of the following category numbers: **1; 2; 3; 4**

CoT prompt Students from 4th grade to 8th grade are currently working on text-based argument essays. After submitting their first draft, students received feedback and revised their second draft. Here is the text and writing task:

<Text + Writing Prompt>

You are a writing teacher who works with students from 4th grade to 8th grade. Students are working on the revision goal of explaining how the evidence supports the main claim. Your task is to compare the first and second drafts to identify the major revision patterns in the students' work in response to the revision goal. To determine the appropriate revision pattern, follow these steps:

Step 1: Determine whether the revision is surface-level. Surface-level revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions. If changes in the second draft exceed 35 words, it is not considered a surface-level revision. If the revision pattern is surface-level, output pattern number "1." If not, proceed to step 2. Example of Pattern 1: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

Step 2: Assess the quality of the newly added content after each piece of evidence. Assign one of the following codes to indicate the effectiveness of each explanation attempt: 1. Personal Commentary: An elaboration that is about personal reactions, evaluations, or feelings. For example, after presenting evidence that Malaria is a disease common in Africa, the student added, "it is pretty cool I never knew that and I am afraid of getting it." 2. Empty explanation: This type of explanation attempt is overly broad, vague, and does not have content, though it may look like an explanation in form. For example, the student added, "Based on what I provided, this is why I agree." 3. Paraphrase: A revision attempt where the explanation merely rewords the evidence rather than explains it. For example, after presenting the evidence " Those in favor of space exploration argue that 19 billion dollars is not too much and satellites can solve the problem of poverty," students added, "we should fund space exploration because 19 billion dollars is not too much and provide more food for people in poverty." 4. Elaboration of Evidence but no

Connection to the claim: Explain the evidence by discussing the implications or providing more information, but it lacks reasoning of why the evidence supports the claim. For example, after presenting the evidence that " Malaria is common in Africa," the student added, “this is a preventable illness, just need people to donate some money, and children can live.” This is an elaboration of evidence, but no explanation of how this example supports the claim. 5. Linked Claim-Evidence: A revision that explains why the provided evidence supports the main claim. For example, to support the claim that we should support space exploration, the student added the explanation, “Malaria is common in Africa. And space exploration can develop new medicine and equipment. These are good evidence because space exploration can save people lives and prevent Malaria”

Step 3: Determine the quality and quantity of explanations added in the second draft If the second draft only contains explanations coded as “personal comments”, “empty explanation”, “paraphrase”, and “elaboration of the evidence”, output 2.

Example of Pattern 2: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.(personal comments instead of explaining how existing evidence supports the claims).

If the second draft contains one piece of explanation category as Linked Claim-Evidence, or reuses the same Linked Claim-Evidence for multiple pieces of evidence, output 3.

Example of Pattern 3: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. If the second draft contains at least two Linked Claim-Evidence, output pattern number “4.” Example of Pattern 4: Draft 1: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. Draft 2: They should get paid because 19 billion dollars a year for exploration. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. so that means that they need money to have the stuff to look and see what is going to happen in the future and there is a machine in space to see what the weather is going to be so they need money for that. It is important because like what is there is a tornado unexpected so they will not know how cold or what is going to happen there might be snow coming and we do not know. (a strong explanation) Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. They also need money to have satellite see if we did not have a satellite we would not know when a tornado would come. I think we should keep giving them money because they are keeping us safe by making a satellite and telling us on the news so we can get the info so we should keep giving they money (a strong explanation).

Apply the aforementioned evaluation steps and reason step by step. Output one of the following category numbers: **1; 2; 3; 4**

Predicting and Evaluating Item Responses Using Machine Learning, Text Embeddings, and LLMs

Evelyn S. Johnson, Hsin-Ro Wei, Tong Wu, and Huan Liu
Riverside Insights

Abstract

This study compares machine learning, text embeddings, and large language models (LLMs) for generating synthetic responses to field test items for a social-emotional assessment that uses Likert-scale responses. Using accuracy metrics and item response theory (IRT) calibration, results show that machine learning and embeddings more closely mirror student data than LLMs. Findings highlight synthetic data's promise while underscoring the need for continued validation.

1 Introduction and Background

Developing high quality assessment items requires rigorous field testing, yet this process is time consuming and costly. Traditional calibration using item response theory (IRT) typically requires hundreds of examinees per item to estimate difficulty, discrimination, and guessing parameters with acceptable precision. This burden is particularly challenging in educational settings where Likert-type surveys are widely used. Such instruments tend to measure sensitive or hard-to-predict constructs (e.g., social-emotional skills, behavioral ratings) and must often pass district level approval for wording and focus, further slowing the process of field-testing. These constraints underscore the need for alternative strategies that can accelerate item validation without compromising psychometric rigor.

Synthetic data has emerged as a promising solution for assessment developers. By generating artificial responses that approximate the distributions and relationships found in real datasets, researchers can reduce reliance on large-scale human field trials. Psychometric research demonstrates that large language models (LLMs) can serve as synthetic respondents, producing item parameter estimates that correlate highly with those derived from human data, though often with narrower variability (Liu et al., 2025). These findings suggest that artificial respondents may augment or partially substitute for actual student responses in item development.

Although LLMs provide one pathway for generating synthetic responses, they are not the only approach under investigation. A more traditional starting point has been machine learning (ML), which relies on historical student response data to predict responses to new items. However, ML models often struggle with unseen items, since new questions cannot be calibrated until sufficient student data is available. To address this limitation, researchers have explored text-aware methods that incorporate semantic information from item content. For example, Khan et al. (2025) introduced Text-LENS, which integrates text embeddings from a transformer encoder. This approach matched baseline ML performance on known items but substantially outperformed it when predicting responses to novel items (Khan et al., 2025). Such embedding-based methods offer a middle ground, more flexible than conventional ML yet more efficient than large-scale LLM simulations.

LLM-based approaches, in contrast, provide a different kind of advantage. By simulating students across ability levels, LLMs can produce synthetic response distributions that reflect difficulty trends and distractor functioning (Benedetto et al., 2024; Shridhar et al., 2023). While not perfect substitutes, these models allow test developers to "pre-pilot" items at scale, discarding poor candidates before committing resources to costly field testing.

ML, text embedding, and LLM approaches reveal a spectrum of tradeoffs. ML methods that rely on prior response patterns may be most

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 66–70
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

effective when field testing items that are structurally and conceptually like those previously administered. Text embedding may provide flexibility when items target new constructs within an existing domain, enabling models to generalize without extensive retraining. LLMs, while computationally intensive, may be necessary when item pools are entirely new or when the goal is to approximate the variability and reasoning patterns of real student responses. The choice among methods may depend on the purpose of item development as well as the need to balance efficiency, fidelity, and generalizability in the context of field testing.

2 Study Purpose

The purpose of this study is to compare three approaches: ML, text embedding, and LLMs, in generating synthetic responses to 10 field test items from the Devereux Student Strengths Assessment (DESSA), a standardized self-report of social-emotional competence. We evaluate the accuracy of the synthetic responses by comparing and calibrating them with IRT to compare estimated item parameters and thresholds to those derived from actual student data.

3 Methods

3.1 Sample

The sample consists of student data (N = 3,982) from an administration of the DESSA high school student self-report form (Robitaille et al., 2025). Students responded to 40 scored items and 10 field test items using a five-point Likert scale ranging from 0 (*Never*) to 4 (*Almost Always*). The demographic information of the sample is summarized in Table 1.

3.2 Measures

The DESSA (Robitaille et al., 2025) is a 50-item, standardized, norm-referenced self-report behavior rating scale for students in 9th – 12th grades that yields T-scores (M=50, SD=10) that are reported into three descriptive categories, “Need for Instruction”, “Typical”, or “Strength”. The DESSA measures six social emotional competencies, Optimistic Thinking, Self-Awareness, Social Awareness, Responsible Decision Making, Relationship Skills, and Self-Management.

Category	**Group**	**N**	**%**
Gender	Female	1995	50.10
	Male	1987	49.90
Grade	9th	1180	29.63
	10th	1139	28.60
	11th	883	22.17
	12th	780	19.59
Race/ Ethnicity	American Indian	49	1.23
	Asian	214	5.37
	Black	1068	26.82
	Hispanic	910	22.85
	Hawaiian	6	.15
	White	1947	48.90
	Other	159	3.99
Region	Midwest	203	5.10
	Northeast	970	24.36
	South	2764	69.41
	West	45	1.1

Table 1. Demographic Information of the Sample.

3.3 Approach

We compared three methods for generating synthetic responses to the 10-field test items from the DESSA. First, we trained a Random Forest classifier on real student responses, using stratified sampling to balance classes. The model was fit to training data (n = 3,186; ~80% of the dataset) and then used to predict synthetic responses for the test set (n = 796; ~20% of the dataset). Accuracy was tracked both at the macro level and for each item.

Next, we used embeddings derived from the item text and response options to inform predictions. These embeddings were incorporated into a predictive model that mapped semantic similarity and structural features to likely student responses. As with the ML approach, predictions were generated for the 10 field-test items, with evaluation against actual student responses.

Finally, we used FLAN-T5, an instruction-tuned transformer model, to simulate student responses. The model was prompted with DESSA item stems and Likert response options, framed as if it were a high school student completing a social-emotional self-report survey. Prompts included general instructions to reflect variability in responses rather than always producing the same option to approximate realistic distributions. In addition, we applied Low-Rank Adaptation (LoRA) fine-tuning on the training data, using real student responses to field-test items as supervised pairs, which allowed the model to better align with the rating scale and item content.

3.4 Graded Response Model

The Graded Response Model (GRM; Samejima, 1968), was employed to analyze polytomous scored items intended to measure varying levels of a latent trait. The GRM is appropriate for items with ordered categorical response options, such as those found in Likert-type scales. The probability of endorsing a response category is calculated as the difference between cumulative logistic functions across thresholds. Item calibration and model estimation were conducted using the mirt package in R (Chalmers, 2012), facilitating a robust evaluation of item functioning and trait estimation.

To evaluate the fidelity of synthetic response data generated for field items, the Pearson correlation coefficient was computed between the student-generated and synthetic response vectors (Cohen, 1988). To evaluate and compare item parameter estimates derived from synthetic data generation methods, a free calibration was conducted using the actual response dataset. From this calibration, forty item parameters were extracted and designated as anchor items. These parameters were fixed across three separate calibration conditions. The remaining ten field-test items were calibrated independently using synthetic response data. This procedure ensured that all item parameters were aligned on a common measurement scale, allowing for valid comparisons across different synthetic methods.

4 Results

We first examined the accuracy of synthetic responses generated by each approach. At the macro level, ML achieved the highest test accuracy (.62), followed closely by text embeddings (.61), while the LLM approach showed lower performance (.55). These differences were consistent across most items, with the ML and text embedding models producing comparable results, and the LLM yielding weaker alignment with observed student responses.

At the micro level, item-specific test accuracies further illustrated these trends (see Table 2). Machine learning predictions for individual field test items ranged from .50 to .71, with higher accuracy observed for items Q15 (“respect a person’s right to have a different opinion?”) and Q35 (“make others feel welcome or included?”). Text embedding results were similar, with item accuracies ranging from .48 to .71, again showing strength on items Q15 and Q35, but lower performance on Q25 (“recognize your emotions?”), and Q50 (“have a teacher or other adult at school you can talk to?”).

LLM performance was consistently lower across items, with test accuracies clustering in the .49 to .63 range. Across all ten items, both ML and text embedding methods maintained consistent predictive performance, whereas the LLM tended to underpredict or misalign with actual student response patterns (Figure 1).

Item	**ML**		**Text Embed**		**LLM**	
	Train	Test	Train	Test	Train	Test
7	1.00	0.63	0.88	0.61	0.57	0.56
10	1.00	0.62	0.90	0.61	0.54	0.54
15	1.00	0.71	0.89	0.71	0.52	0.48
20	1.00	0.66	0.87	0.64	0.61	0.62
25	1.00	0.53	0.86	0.53	0.53	0.53
30	1.00	0.62	0.87	0.59	0.57	0.57
35	1.00	0.67	0.87	0.65	0.62	0.59
40	1.00	0.62	0.87	0.60	0.63	0.62
45	1.00	0.67	0.89	0.66	0.53	0.49
50	1.00	0.50	0.86	0.48	0.49	0.49

Table 2. Item level accuracies across ML, text embedding and LLM approaches.

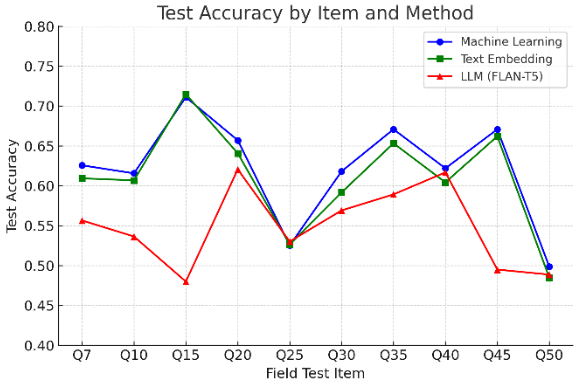


Figure 1: Accuracy levels per item across ML, text-embedding and LLM approaches.

4.1 IRT Calibration

To assess whether synthetic responses could recover psychometric characteristics of the DESSA field test items, we calibrated the 10 field items under a GRM using synthetic data from each method and compared the resulting discrimination and threshold parameters with those obtained from real student responses (Table 3).

Statistic	Method	Mean	SD	Bias	RMSE	r
a	Actual	1.86	0.36			
	ML	3.16	0.63	1.30	1.86	0.76
	Text	3.02	0.57	1.16	1.43	0.88
	LLM	5.06	0.01	3.21	10.41	-0.33
b1	Actual	-2.64	0.49			
	ML	-2.59	0.41	0.05	0.06	0.87
	Text	-2.47	0.43	0.17	0.06	0.93
	LLM	-2.32	0.03	0.32	0.33	-0.02
b2	Actual	-1.67	0.48			
	ML	-1.84	0.53	-0.17	0.05	0.95
	Text	-1.73	0.51	-0.07	0.01	0.98
	LLM	-1.65	0.01	0.02	0.21	-0.07
b3	Actual	-0.38	0.35			
	ML	-0.41	0.39	-0.03	0.00	0.99
	Text	-0.42	0.40	-0.04	0.01	0.98
	LLM	-0.30	0.01	0.08	0.12	0.36
b4	Actual	0.88	0.31			
	ML	0.87	0.30	-0.01	0.01	0.92
	Text	0.81	0.30	-0.08	0.02	0.91
	LLM	0.73	0.01	-0.15	0.10	0.48

Table 3. Estimation of Graded Response Model Item parameters (10 field items).

Across items, the ML and text embedding approaches showed broadly similar correspondence with actual parameters; neither consistently outperformed the other. By contrast, the LLM simulations exhibited weaker alignment with thresholds from actual student data and greater instability across items, echoing their lower classification accuracy. Overall while all three methods produced plausible synthetic responses, the ML and embedding approaches better preserved psychometric fidelity relative to the LLM.

5 Conclusion

This investigation highlights both the promise and the limitations of synthetic data for accelerating assessment development. Across the three synthetic data approaches applied to the DESSA field items, conventional machine learning slightly outperformed the text-embedding model, and both exceeded the LLM in aligning with observed student responses and IRT-derived item parameters. These findings indicate that ML and embeddings can plausibly support early item evaluation and calibration, while current LLM outputs appear less reliable for parameter recovery for assessments like the DESSA. Continued investigation on novel field test items aligned with different purposes (e.g., similar items for new constructs within the same domain) will inform when to use different approaches to generate synthetic data. Overall, synthetic approaches hold promise for reducing reliance on costly field testing, but continued investigation, with larger item sets, additional benchmarks, and rigorous IRT comparisons, is needed before they can be used with confidence in operational assessment.

Acknowledgments

We would like to thank our colleagues at Riverside Insights for their feedback on this work.

References

Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., & Buttery, P. (2024). *Using LLMs to simulate students' responses to exam questions*. Findings of the Association for Computational Linguistics: EMNLP 2024, 11351–11368. https://doi.org/10.18653/v1/2024.findings-emnlp.663

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Khan, A., Li, N., Shen, T., & Rafferty, A. N. (2025). *Just read the question: Enabling generalization to new assessment items with text awareness*. arXiv preprint arXiv:2507.08154. https://doi.org/10.48550/arXiv.2507.08154

Liu, Y., Bhandari, S., & Pardos, Z. A. (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

Robitaille, J., Johnson, E.S., LeBuffe, P. A., & Naglieri, J. (2025). *Devereux Student Strengths Assessment (DESSA)–High School Edition*. Riverside Insights.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores 1. *ETS Research Bulletin Series*, *1968*(1), i-169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Shridhar, K., Bicer, H., & Liu, Q. (2023). Generating and evaluating tests for K–12 with LM item-response simulators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2210–2224). https://doi.org/10.18653/v1/2023.emnlp-main.135

A Appendices

none

B Supplementary Material

none

Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity

Yue Huang
Measurement Incorporated
`yueh@udel.edu`

Joshua Wilson
School of Education
University of Delaware
`joshwils@udel.edu`

Abstract

This study evaluates large language models (LLMs) for automated essay scoring (AES) in comparison with a traditional feature-based system (PEG) and human ratings. Drawing on 541 essays from Grades 3–4, we examined three generative LLMs (including GPT-4o, Gemini 2.5 Flash and Claude Sonnet 4) under baseline, context-enhanced, and few-shot chain-of-thought prompting strategies. Results show that carefully designed prompting, particularly context-enhanced few-shot chain-of-thought, substantially improved LLM performance, approaching PEG in human–machine agreement and human–human agreement. Fairness analyses revealed that PEG produced larger disparities for English language learners (ELLs), while LLMs showed smaller but still persistent subgroup bias. Beyond these findings, the study contributes recent evidence on fairness and validity in LLM-based AES and extends research to younger students, a group rarely examined in prior work. Together, these results highlight both the promise and the challenges of integrating LLMs into educational assessment.

1 Introduction

Automated essay scoring (AES) refers to the use of computational methods and/or AI techniques to evaluate student-generated writing and assign scores in place of or alongside human raters (Uto and Okano, 2020). In the field of educational measurement, automatic scoring has become a cutting-edge approach to evaluating written content without manual grading. This strategy is especially valuable in large-scale assessments and classroom contexts where manual scoring is impractical (Latif and Zhai, 2024; Susanti et al., 2023). Early AES systems were built on *natural language processing* (NLP) and machine learning methods with handcrafted features and large labeled datasets (Uto, 2021). More recently, deep learning models such as recurrent and transformer-based architectures have improved scoring performance by leveraging text embeddings and contextual representations (El-Massry et al., 2025). Still, these approaches face limitations in interpretability, scalability, and fairness.

The rise of *large language models* (LLMs) offers new potential for AES. Pretrained on massive corpora, LLMs capture sophisticated semantic and discourse-level features, eliminating the need for manual feature engineering. With careful prompt design or fine-tuning, these models can be guided to produce not only holistic scores but also rubric-aligned scores. However, critical challenges remain: model outputs can vary depending on prompt engineering, and concerns about fairness, subgroup performance, and transparency persist (Huang et al., 2025).

This paper addresses these gaps by systematically evaluating several LLM-based AES approaches and comparing them with a more traditional feature-based AES system. We examined: (1) accuracy, measured by their alignment with human ratings; (2) fairness, focusing on differences between English language learners (ELLs) and non-ELLs. By analyzing prompting strategies across different LLMs and subgroup outcomes, we contribute empirical evidence to guide both AES research and educational assessment practice.

2 Related Work

2.1 AES in Educational Assessment

Research on automated essay scoring (AES) has a long history in educational assessment, beginning with the initial version of Project Essay Grade (PEG; Page, 1966) scoring engine. PEG relied on surface-level textual features, such as word counts, sentence length, and syntactic structures, combined with statistical modeling to predict scores. Subsequent systems, such as e-rater, extended this approach by incorporating more linguistically in-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 71–83
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

formed features grounded in NLP, including grammar, usage, mechanics, style, and organizational development (Hussein et al., 2019). These early systems demonstrated that linguistic features, when coupled with statistical models, could produce levels of agreement with human raters comparable to inter-rater reliability, establishing the feasibility of automated scoring for constructed responses, short answers, and essays.

With advances in artificial intelligence, AES shifted toward feature-rich regression and classification models trained on large corpora during the mid-2010s. The adoption of deep neural networks, particularly recurrent architectures (such as LSTM) and convolutional neural networks (CNN), enabled models to capture sequential dependencies in student writing (Dong et al., 2017; Taghipour and Ng, 2016). The emergence of transformer-based large language models (LLMs), including BERT and XLNet, further advanced AES by leveraging contextual embeddings that outperformed prior deep learning methods (Rodriguez et al., 2019; Uto, 2021). Building on this foundation, Yang et al. (2020) introduced R^{2}BERT, a BERT-based model that combined regression and ranking objectives, achieving strong performance on the widely used Automated Student Assessment Prize (ASAP) dataset (Hamner et al., 2012). Extending this line of work, Xie et al. (2022) proposed Neural Pairwise Contrastive Regression (NPCR), a contrastive learning approach that modeled score differences across essay pairs and achieved state-of-the-art results. These models reduced reliance on manual feature engineering and improved generalizability across essay tasks.

Despite notable advances, AES still faces important limitations. First, systems struggle to capture the full range of features that characterize high-quality writing, particularly when holistic scoring and rubric-based analytic scoring demand different forms of feature representation (Kumar and Boulanger, 2021). Elements such as rhetorical intent, coherence, and creativity remain especially difficult to model (Huang et al., 2025). Second, much empirical research relies heavily on benchmark datasets such as the ASAP corpus (focusing on essays from students in Grades 7–8), which facilitate standardized comparisons but offer limited insight into writing at earlier developmental stages where challenges are most acute. Finally, fairness continues to be a major concern. Feature-based AES systems frequently inherit biases present in the human ratings used for training, leading to systematically different outcomes for subgroups such as ELLs (Wilson and Huang, 2024). Ensuring equitable scoring across diverse student populations therefore remains a central challenge for AES in educational measurement.

2.2 Generative Large Language Models for AES

The recent development of generative LLMs such as GPT-4 and Llama-3 (referred to hereafter as GPT-family models for simplicity) has demonstrated remarkable capabilities in language understanding, reasoning, and text generation. Unlike earlier encoder-based LLMs, GPT-family models adopt decoder-based, autoregressive architectures (Minaee et al., 2025). This design enables them to generate coherent and contextually rich text, capture nuanced semantic relationships, and adapt flexibly to varied writing genres and proficiency levels, which holds particular promise for evaluating essays in ways that attend not only to surface features but also to deeper rhetorical and logical structures.

A growing body of research has examined the performance of generative LLMs for AES, though findings remain mixed. Results vary depending on prompting strategies, fine-tuning methods, and system adaptation (Huang et al., 2025). Proprietary models such as GPT-3.5 and GPT-4 show reasonable performance with few-shot prompting, especially when combined with rubric descriptions, explicit task instructions, and/or chain-of-thought (CoT) reasoning (Mansour et al., 2024; Quah et al., 2024; Wei et al., 2022). However, they often underperform compared to fine-tuned models and raise concerns about transparency and replicability. Designing prompts that ensure reliability remain an open challenge.

Open-source models such as Llama-3 introduce new opportunities. Research by Ormerod and Kwako (2024) demonstrated that smaller open-source models, when fine-tuned, can achieve performance comparable to traditional best-performing models while running on modest hardware. This approach enhances transparency and allows researchers to integrate explainable AI methods, addressing some of the limitations of closed GPT-family models. Yet, fine-tuning requires technical expertise, and performance still lags behind state-of-the-art models on benchmark datasets.

As with earlier LLMs, the adoption of GPT-family models raises broader concerns about fairness and validity (Huang et al., 2025). Few studies have examined subgroup differences, and those that exist focus mainly on multilingual learners (e.g., Tate et al., 2024). Ethical concerns are also mounting, particularly around data privacy, consent, and intellectual property.

Overall, research on LLM-based AES is still emerging. Current evidence suggests that while LLMs can approximate human scoring with careful prompt design or fine-tuning, their performance remains inconsistent across contexts, and fairness outcomes are underexplored. The next phase of research must therefore integrate technical advances with principles of educational measurement to ensure that LLM-based automated scoring is both effective and equitable. The present study contributes to this effort by examining LLM scoring across student subgroups, specifically ELLs versus non-ELLs.

3 Research Questions

This study evaluates three large language models (LLMs) alongside a traditional feature-based AES system (PEG) to examine their alignment with human ratings and their fairness for ELLs compared to non-ELLs, under three different prompt engineering strategies.

RQ1: How do prompt engineering strategies affect human–machine agreement across LLMs, PEG, and human raters?

RQ2: Do LLMs exhibit performance differences or subgroup bias between ELL and non-ELL students?

4 Methods

4.1 Sample

This study draws on data from an evaluation of an automated writing evaluation system in Grades 3–5 in a U.S. school district in school year 2017–2018. The district implemented the system in conjunction with a Common Core–aligned English language arts curriculum to support writing instruction for all students. A subsample of 541 de-identified essays from third and fourth graders (N = 233 and 308, respectively) written between April 1 to May 31, 2018, was analyzed; each grade responded to a separate grade-level informative essay task. For the writing tasks, Grade 3 students read two short texts about national parks—one emphasizing their value for recreation, wildlife protection, and science, and the other highlighting challenges such as pollution and overcrowding—and were asked to write an informative essay explaining what national parks are and why they matter. Similarly, Grade 4 students read texts introducing invertebrates and describing the features, habitats, and life cycle of crabs, and were asked to write an informative essay about the key characteristics of crabs and how they live. Essays were scored by six approaches (see details below). Ten percent of the essays (N = 57) were randomly double scored by a second human rater. ELLs comprised 32% of third graders and 46% of fourth graders.

4.2 Measures

Six scoring approaches were evaluated: (1) human rater 1, (2) human rater 2 (10% of the sample), (3) PEG, (4) LLMs with baseline CoT prompting, (5) LLMs with context-enhanced CoT prompting, and (6) LLMs with context-enhanced + few-shot CoT prompting. Three LLMs were considered: GPT-4o, Gemini 2.5 Flash and Claude Sonnet 4.

Human raters were professional scorers employed by the company operating the automated writing evaluation system. They received extensive training and were continuously monitored through rater management systems designed to ensure scoring accuracy and consistency. Human raters applied a six-trait rubric assessing development of ideas, organization, style, sentence fluency, word choice, and conventions. Each trait was scored on a 1–5 scale, and a holistic score was obtained by summing the six traits (range = 6–30).

The most recent PEG scoring engine has advanced substantially beyond its earlier, simpler versions. Current PEG scores are produced using a proprietary model that integrates more than 800 linguistic features with deep learning algorithms, trained on a large corpus of historical student essays from the same grade band and curriculum-aligned tasks.

Figure 1 presents the flowchart for the three prompting strategies. In the baseline CoT condition, prompts included the scoring task instructions, essay task description, rubric details, and a CoT component guiding the model to reason step by step about how to apply the rubric. The system was then asked to generate a score and provide the scoring output as specified. For the context-enhanced CoT strategy, one additional component was introduced: the model was assigned the role of an experienced essay rater familiar with the writing proficiency levels of third- and fourth-grade stu-

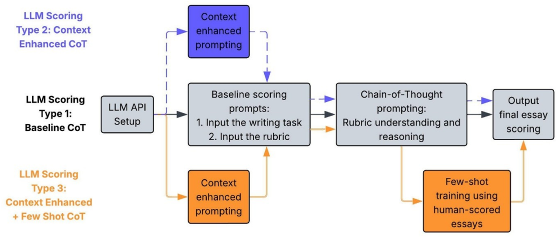

LLM Scoring Type 2: Context Enhanced CoT
Context enhanced prompting
LLM Scoring Type 1: Baseline CoT
LLM API Setup
Baseline scoring prompts:
1. Input the writing task
2. Input the rubric
Chain-of-Thought prompting:
Rubric understanding and reasoning
Output final essay scoring
LLM Scoring Type 3: Context Enhanced + Few Shot CoT
Context enhanced prompting
Few-shot training using human-scored essays

Figure 1: Flow Chart for the Design of Prompt Engineering Strategies Using LLMs

dents. For the context-enhanced + few-shot CoT strategy, another component was added: the model was provided with five sample essays along with their corresponding trait-level and holistic scores to guide its scoring decisions. All prompts were executed iteratively for each individual essay through API interactions with the models using Python.

4.3 Data Analysis

For RQ1, we evaluated human–machine agreement by calculating *Quadratic Weighted Kappa* (QWK) and *exact/adjacent agreement rates* between each scoring method (**S**: PEG and all LLMs across prompting strategies, as well as human rater 2) and human scores (**H**, from rater 1). QWK, a widely used reliability index, ranges from 0 to 1, with higher values indicating stronger alignment between two sets of ratings. Exact agreement reflects the proportion of cases where the AES score matches the human score exactly, while adjacent agreement reflects cases where the AES score is within ±1 point of the human score. Together, these measures provide complementary perspectives on model accuracy relative to human raters.

For RQ2, QWKs were calculated separately for ELL and non-ELL students to assess subgroup-specific performance. To further evaluate fairness, we applied Litman et al.'s (2021) metrics:

- *Overall Score Accuracy* (OSA): Measures whether AES scores are equally accurate across groups by regressing squared error $(S - H)^2$ on student group. A significant positive coefficient indicates systematic differences in accuracy between groups.

- *Overall Score Difference* (OSD): Assesses whether AES scores are consistently higher or lower than human scores for different groups using the absolute difference $|S - H|$ as the outcome. Significant differences suggest systematic over- or under-prediction for a subgroup.

- *Conditional Score Difference* (CSD): Extends OSD by controlling for student proficiency (approximated by human scores). Two models are compared—with and without student group. A significant increase in R^2 indicates that group membership affects AES accuracy beyond proficiency, signaling potential subgroup bias.

Based on RQ1 findings (see Results section), only LLM scores generated with context-enhanced + few-shot CoT prompting—the highest-performing strategy overall—were subjected to fairness analyses (OSA, OSD, CSD).

5 Results

5.1 RQ1–Comparisons across Scoring Approaches

Figure 2 shows QWK values for total scores. Human–human agreement was the highest benchmark (QWK = .91), with PEG next in line (QWK = .76). The LLMs, while trailing PEG, demonstrated a clear upward trend across prompting strategies: GPT-4o improved from .46 under baseline CoT to .72 with context-enhanced + few-shot prompting, Gemini 2.5 Flash rose from .43 to .60, and Claude Sonnet 4 from .30 to .69. These results indicate that structured prompts, especially those combining context and few-shot examples, substantially strengthen the alignment of LLM-generated scores with human ratings.

Trait-level analyses (Figure 3) reveal similar patterns. PEG maintained strong agreement across all traits (QWK = .61–.74), consistently falling between human–human agreement (.77–.86) and LLM performance. Among the LLMs, GPT-4o

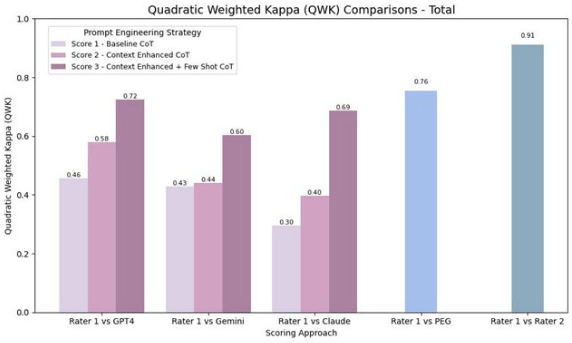


Figure 2: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

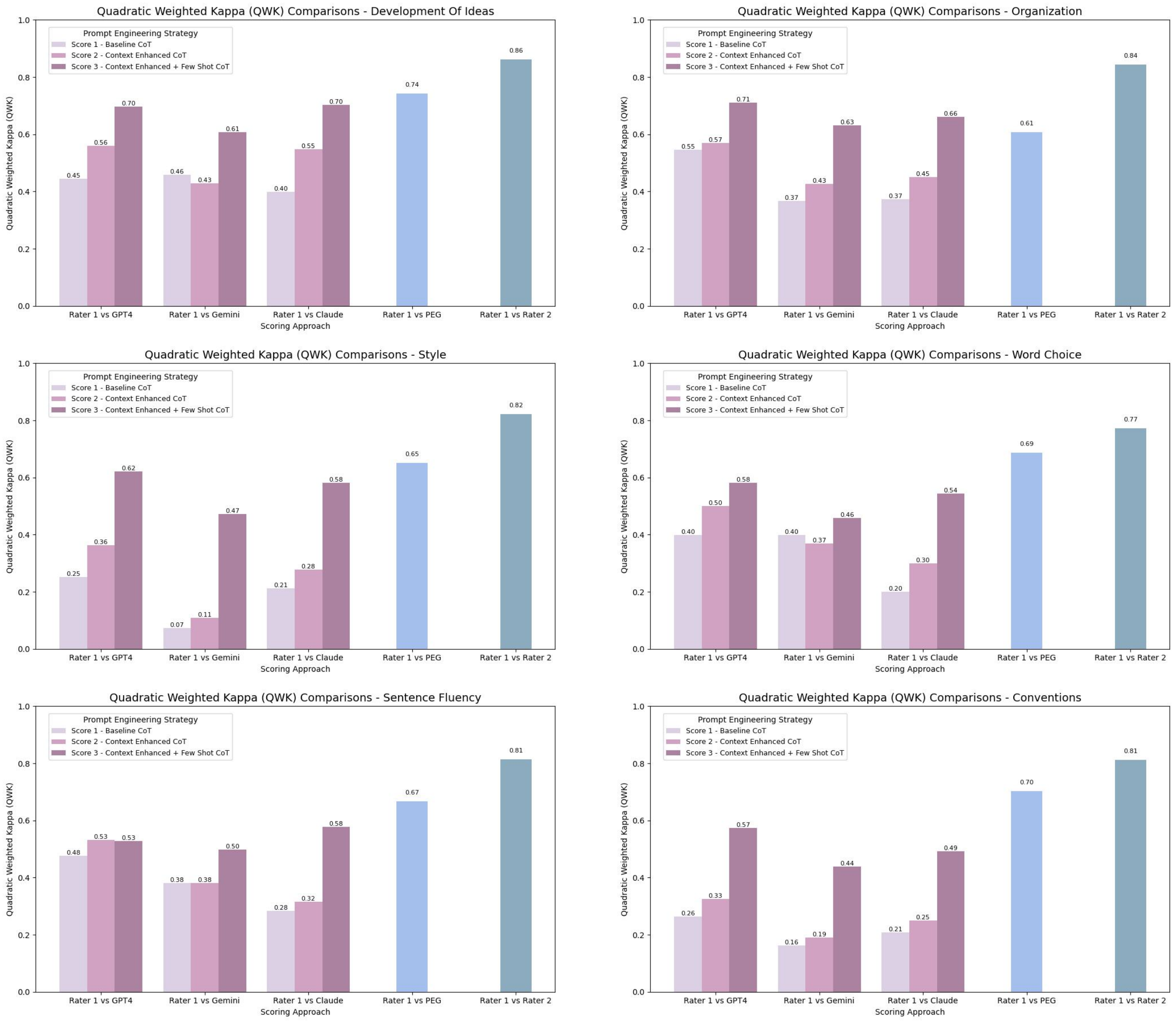


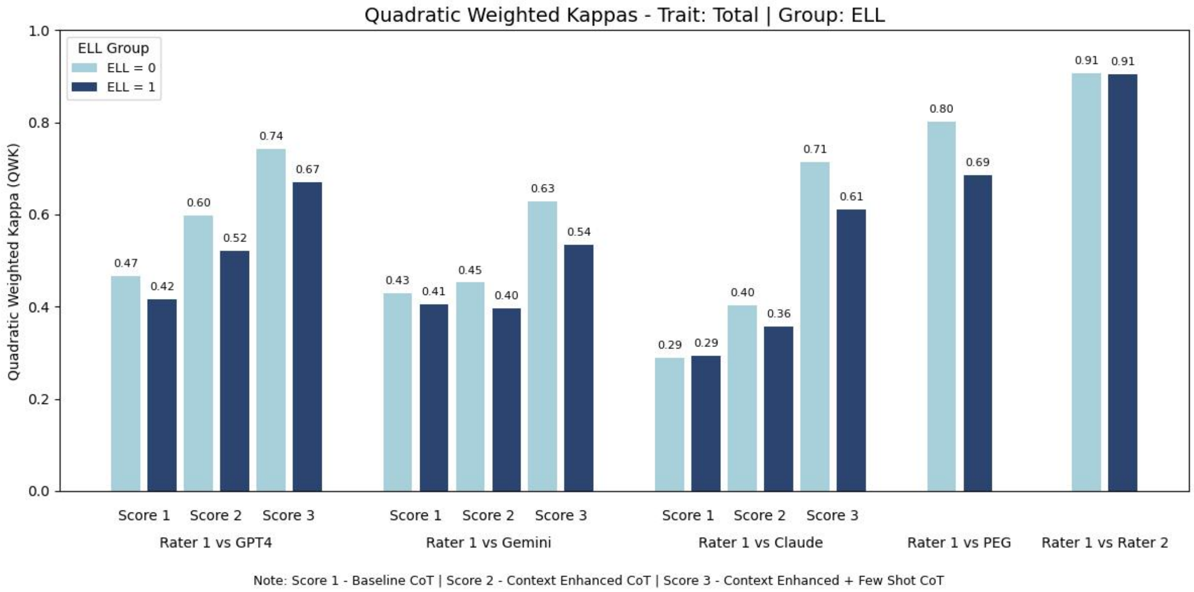
Figure 3: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait

again showed the highest alignment, particularly for development of ideas (.70) and organization (.71), while Claude Sonnet 4 performed competitively for development of ideas (.70). Gemini 2.5 Flash generally lagged behind, though its agreement improved under structured prompting.

Patterns in exact and adjacent agreement (see Appendix A) further support these findings. Exact agreement was highest for human–human (.28 for total score) and PEG (.21), with LLMs showing smaller but improving proportions as prompting strategies became more structured (e.g., GPT-4o rising from .14 to .20). Adjacent agreement was consistently stronger for total scores and trait scores. For example, human–human reached .63 in total score, PEG achieved .44, and LLMs again improved with prompting, with GPT-4o and Claude Sonnet 4 approaching PEG's level for traits including development of ideas, organization, style and word choice. Overall, these results suggest that while PEG remains the most reliable automated scorer, LLMs (particularly GPT-4o) can achieve meaningful gains through contextually enriched, few-shot prompting, with the largest improvements seen on traits tied to style and conventions.

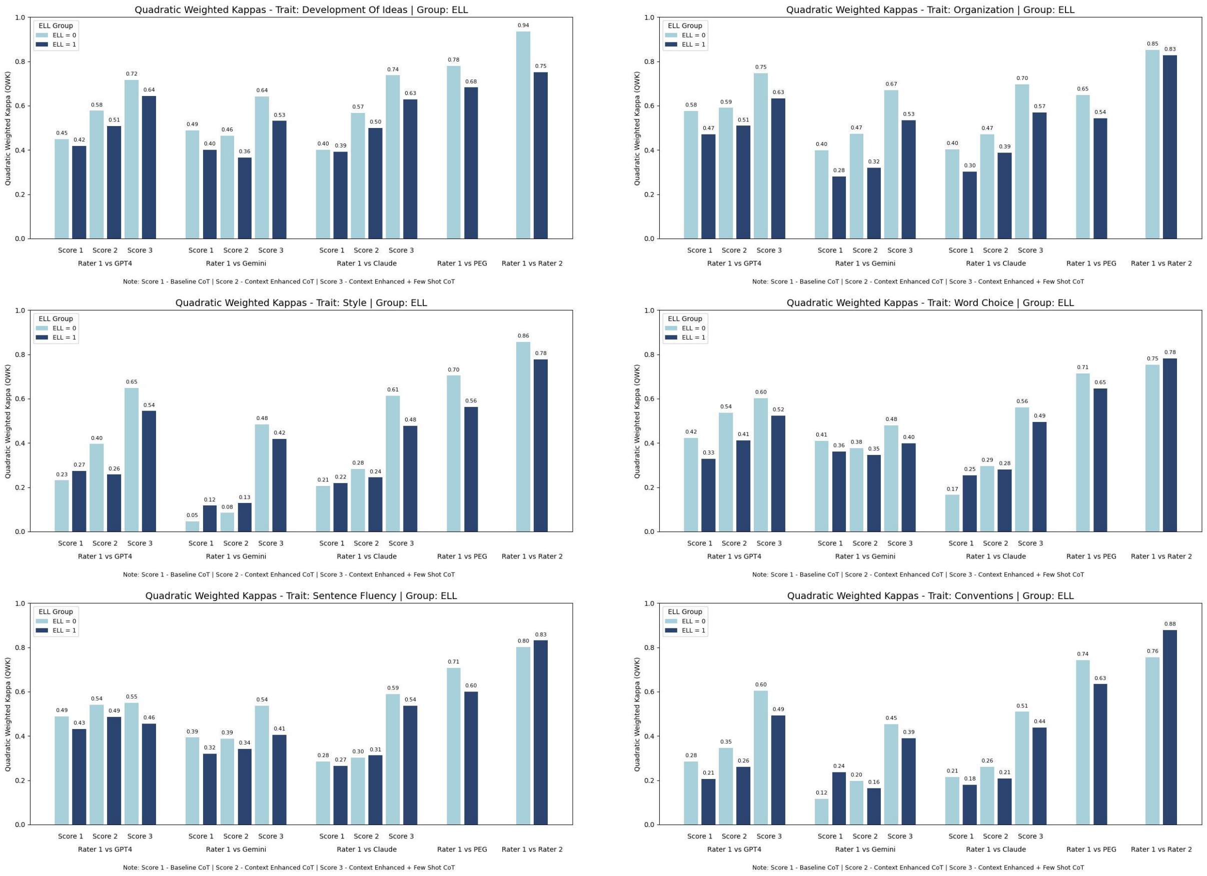
5.2 RQ2–Fairness across ELL Group

Figures 4 and 5 show QWK comparisons by ELL status. Across nearly all models and traits, agreement between AES scores and human ratings was higher for non-ELLs than for ELLs, indicating modest subgroup disparities. For total scores, GPT–human agreement reached .74 for non-ELLs versus .67 for ELLs under context-enhanced + few-shot prompting, Claude–human agreement achieved .71 versus .61, and Gemini–human agreement .63 versus .54, while PEG-human agreement


Quadratic Weighted Kappas - Trait: Total | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Figure 4: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs by ELL Status – Total Score


Quadratic Weighted Kappas - Trait: Development Of Ideas | Group: ELL

Quadratic Weighted Kappas - Trait: Organization | Group: ELL

Quadratic Weighted Kappas - Trait: Style | Group: ELL

Quadratic Weighted Kappas - Trait: Word Choice | Group: ELL

Quadratic Weighted Kappas - Trait: Sentence Fluency | Group: ELL

Quadratic Weighted Kappas - Trait: Conventions | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Figure 5: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs by ELL Status – by Trait

Metric	Human Rater 2	PEG	GPT-4o (Context+Few-Shot CoT)	Gemini (Context+Few-Shot CoT)	Claude (Context+Few-Shot CoT)
Total Score					
OSA – ELL	1.047	**6.293***	-0.077	0.658	1.502
OSD – ELL	-0.014	**0.693***	-0.211	-0.240	-0.354
CSD – ΔR^2	0.001	0.006	**0.010***	**0.009***	**0.016***
Trait 1 – Development of Ideas					
OSA – ELL	**0.260***	0.146	-0.008	0.072	0.080
OSD – ELL	**0.262***	**0.133***	-0.098	-0.073	-0.039
CSD – ΔR^2	0.064	0.004	**0.012***	0.006	0.006
Trait 2 – Organization					
OSA – ELL	-0.003	**0.198***	0.061	0.043	0.040
OSD – ELL	-0.146	0.105	0.004	-0.079	**-0.120***
CSD – ΔR^2	0.033	0.001	0.002	**0.008***	**0.018***
Trait 3 – Style					
OSA – ELL	0.128	**0.214***	-0.026	-0.115	-0.031
OSD – ELL	-0.043	**0.124***	-0.001	-0.019	-0.113
CSD – ΔR^2	0.003	0.002	**0.006***	**0.008***	**0.026***
Trait 4 – Word Choice					
OSA – ELL	0.040	0.137	0.028	0.056	0.080
OSD – ELL	0.073	**0.124***	-0.005	-0.027	-0.024
CSD – ΔR^2	0.000	0.002	**0.007***	**0.008***	**0.009***
Trait 5 – Sentence Fluency					
OSA – ELL	-0.077	**0.200***	-0.004	0.097	0.014
OSD – ELL	-0.072	**0.117***	-0.054	-0.034	-0.023
CSD – ΔR^2	0.008	0.001	**0.016***	**0.009***	**0.009***
Trait 6 – Conventions					
OSA – ELL	-0.208	**0.165***	0.030	-0.054	-0.013
OSD – ELL	-0.088	0.090	-0.057	-0.007	-0.036
CSD – ΔR^2	0.019	0.000	**0.012***	**0.006***	**0.009***

Table 1: Fairness Evaluation Results by ELL Status. Values are coefficients for OSA and OSD (ELL effect) and ΔR^2 for CSD. Significant values are in **bold** and marked with * ($p < .05$).

also favored non-ELLs (.80 vs. .69). Gaps between ELL and non-ELL for human–human agreement also varied somewhat across traits, with smaller subgroup differences for organization, word choice, and sentence fluency. Notably, prompting did not eliminate subgroup gaps, and both LLMs and PEG continued to score ELLs less consistently than non-ELLs. These comparisons with human–human agreements should be interpreted cautiously, however, given the limited size of the double-scored sample (10%) by a second human rater.

Table 1 presents results from the fairness evaluation. Based on RQ1 findings, only LLM scores generated with context-enhanced + few-shot CoT prompting—the most accurate overall—were examined further. Specifically, PEG showed significant ELL-based differences in both OSA and OSD for most traits, but these differences were not significant regarding CSD. In contrast, the three LLMs with context-enhanced + few-shot prompting displayed few significant results for OSA and OSD, implying more balanced performance across groups at the overall level. Yet, CSD revealed persistent disparities: GPT-4o showed effects for total score, development of ideas, style, word choice, sentence fluency, and conventions; Gemini 2.5 Flash for total score, organization, style, word choice, and conventions; and Claude Sonnet 4 for nearly all traits except development of ideas. These findings suggest that while LLMs reduced overt subgroup bias relative to PEG, subtler inequities remained once proficiency was considered.

6 Conclusions and Implications

This study provides early empirical evidence that large language models (LLMs), specifically generative LLMs such as GPT-family models, when combined with carefully designed prompting strategies, can approach the performance of feature-based AES systems such as PEG. This study compared not only multiple LLMs but also different prompting strategies, offering valuable insights and practical guidance for future research on prompt design. Context-enhanced + few-shot chain-of-thought prompting consistently outperformed baseline approaches, highlighting the central role of prompt engineering in optimizing LLM-based scoring for both accuracy and consistency.

At the same time, fairness analyses revealed that neither PEG nor LLMs fully eliminated subgroup disparities. PEG exhibited larger discrepancies for ELLs in overall accuracy and error magnitude, whereas LLMs appeared more balanced at the surface level. However, conditional score difference analyses showed that subtle, proficiency-adjusted disparities persisted across traits, suggesting that fairness concerns remain in LLMs. Importantly, this study examined both holistic scores and rubric-based analytical scores, contributing evidence on how LLMs perform across different scoring dimensions. Furthermore, it provides some of the most up-to-date findings on subgroup fairness in LLM-based scoring, adding important validity evidence to ongoing debates about their educational use. These findings underscore the importance of evaluating LLMs with multiple fairness metrics and designing safeguards that ensure equitable performance across student populations.

Finally, this study focused on students in Grades 3–5, a population often overlooked in AES research, thereby extending the scope of evidence to younger learners who are at a critical stage in writing development. Future work should extend these findings to additional grade levels, writing genres, and more diverse student populations. There is also a need for clearer evaluation frameworks and design guidelines to ensure prompt quality and subgroup fairness in LLM-based scoring. As LLMs gain traction in educational measurement, this study underscores the need to pair advanced modeling with thoughtful design to support scoring accuracy, fairness, and validity.

7 Limitations

Several limitations should be acknowledged. First, only 10% of essays were double-scored, limiting the reliability of human–human benchmarks, particularly for subgroup comparisons. Second, the analysis focused solely on informative writing tasks, leaving other genres such as argumentative or narrative unexamined. Finally, only three prompting strategies were tested, while other approaches, such as extended rubric prompts or fine-tuning, remain unexplored. These constraints suggest caution in interpreting findings and point to directions for future research.

References

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)*, Vancouver, Canada. Association for Computational Linguistics.

Ahmed M. ElMassry, Nazar Zaki, Negmeldin AlSheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, pages 1–1. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring [data competition]. Kaggle. Accessed: 2025-02-15.

Yue Huang, Corey Palermo, Ruitao Liu, and Yong He. 2025. An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation. *Chinese/English Journal of Educational Measurement and Evaluation*, 6(2).

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5:e208.

Vivekanandan S. Kumar and David Boulanger. 2021. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 12748, pages 255–267. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? *arXiv preprint*. ArXiv:2403.06149 [cs].

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A Survey. *arXiv preprint*. ArXiv:2402.06196 [cs].

Christopher Ormerod and Alexander Kwako. 2024. Automated text scoring in the age of generative AI for the GPU-poor. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3).

Ellis B. Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Bernadette Quah, Lei Zheng, Timothy Jie Han Sng, Chee Weng Yong, and Intekhab Islam. 2024. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, 24(1). Publisher: Springer Science and Business Media LLC.

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

Meilia Nur Indah Susanti, Arief Ramadhan, and Harco Leslie Hendric Spit Warnars. 2023. Automatic essay exam scoring system: a systematic literature review. *Procedia Computer Science*, 216:531–538.

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*, Austin, Texas. Association for Computational Linguistics.

Tamara P. Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255. Publisher: Elsevier BV.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Masaki Uto and Masashi Okano. 2020. Robust neural automated essay scoring using item response theory. In *Artificial Intelligence in Education: Proceedings of the 21st International Conference*, pages 549–561, Cham. Springer International Publishing.

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. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Joshua Wilson and Yue Huang. 2024. Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing*, 60:100815.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

A Additional Figures

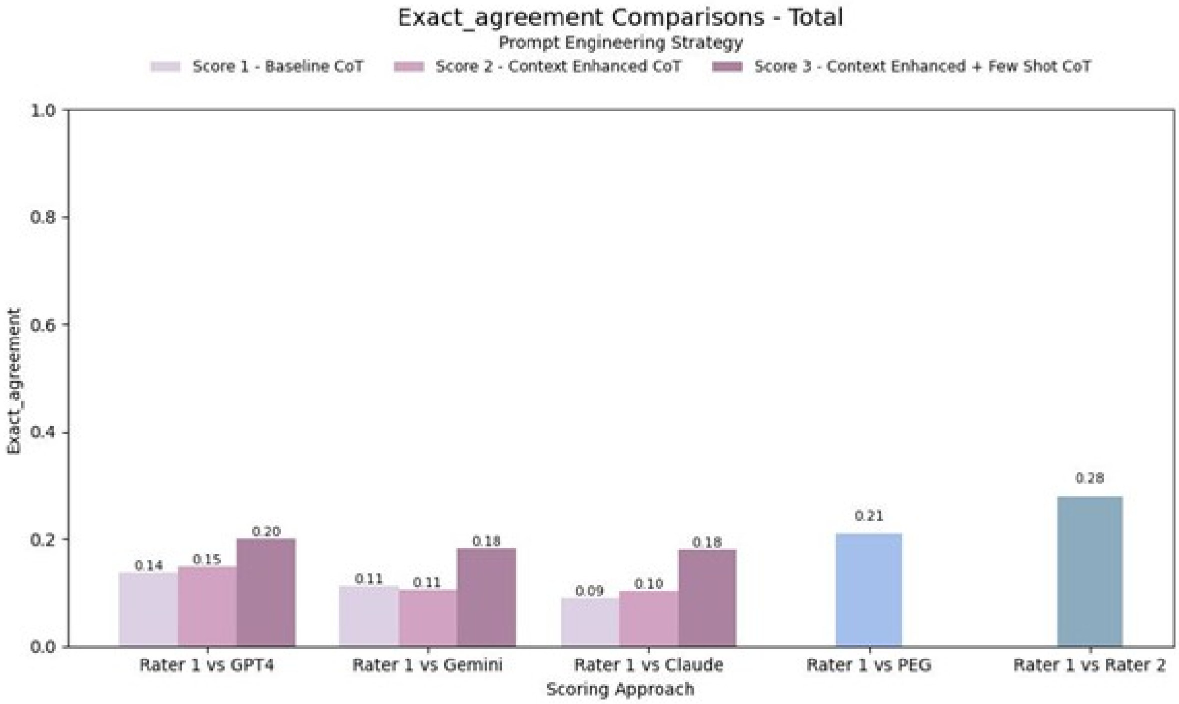


Figure A1: Comparisons of Exact Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

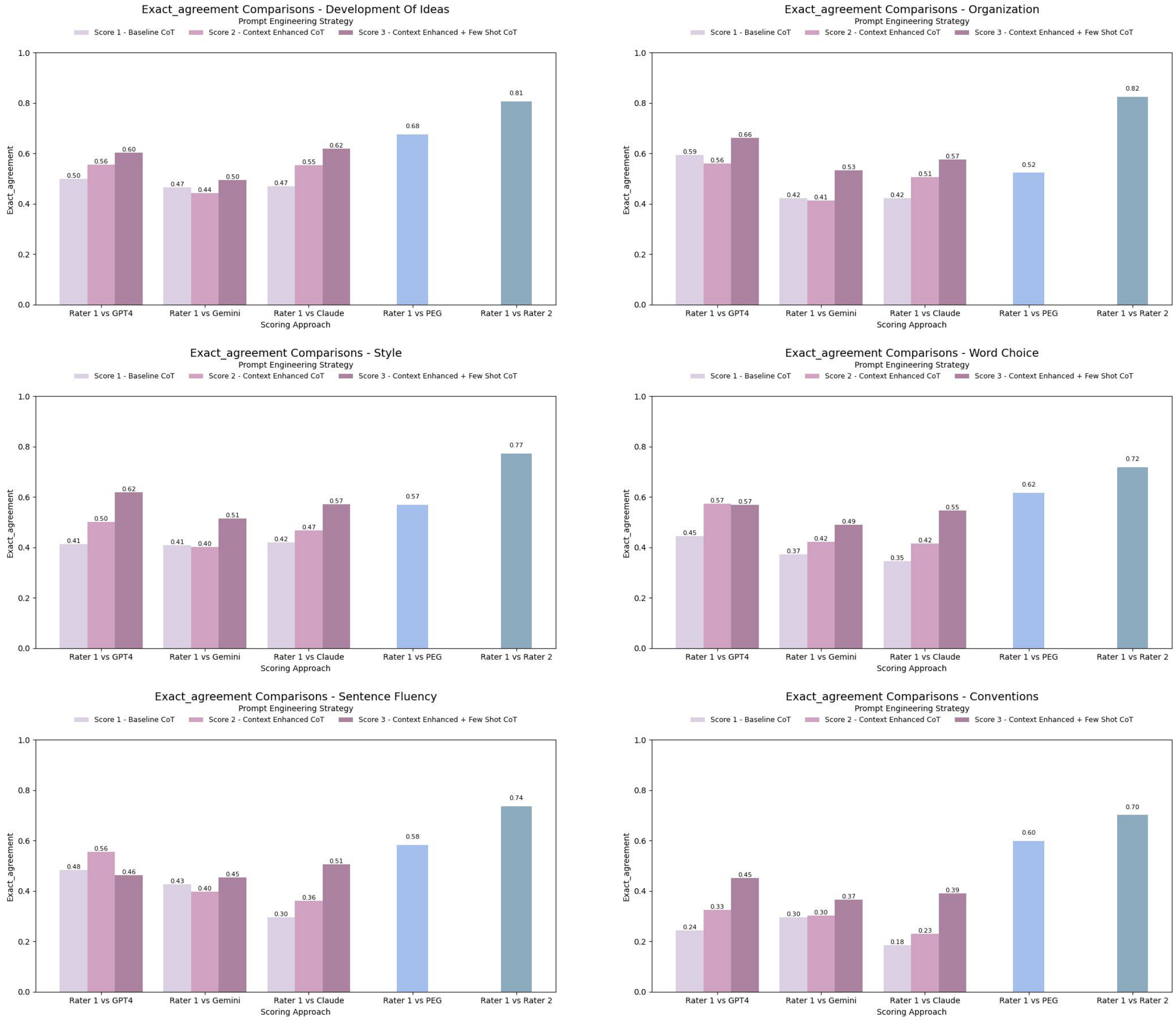


Figure A2: Comparisons of Exact Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait

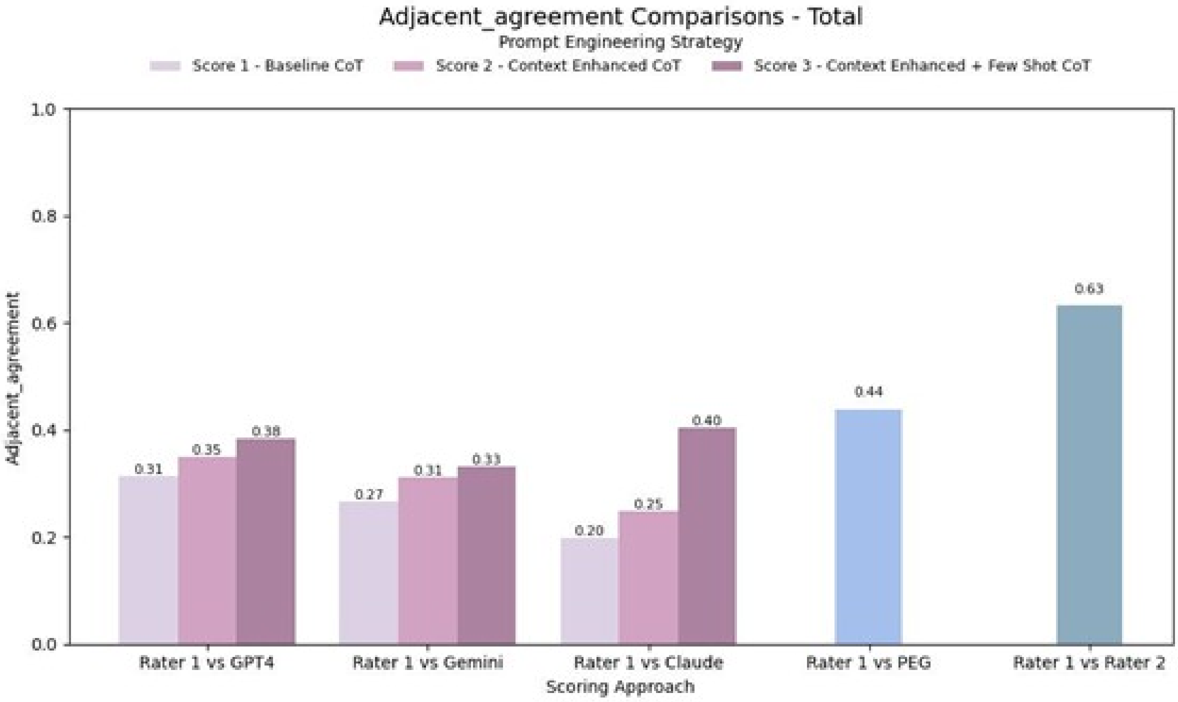


Figure A3: Comparisons of Adjacent Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

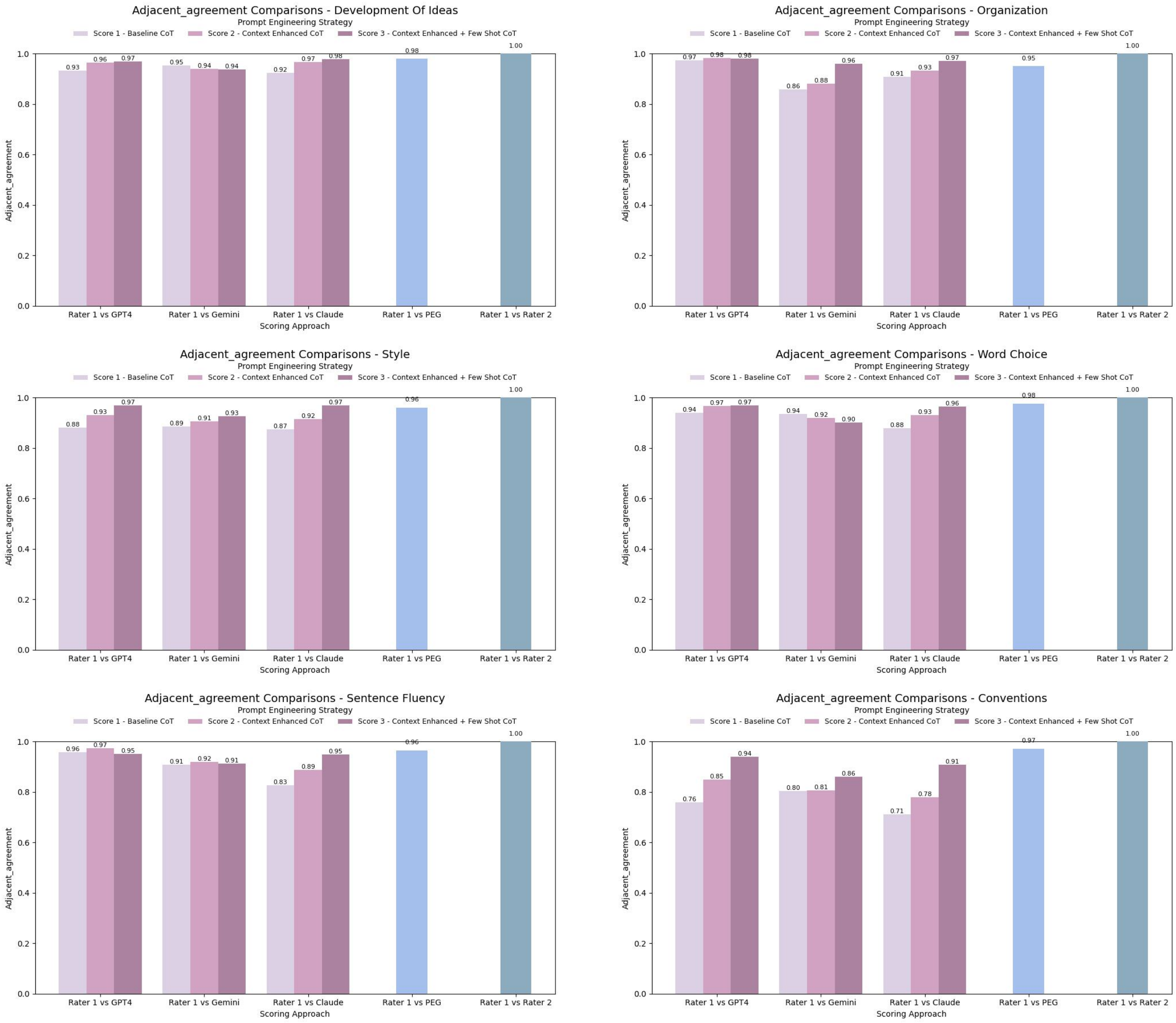


Figure A4: Comparisons of Adjacent Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait

Comparing AI tools and human raters in predicting reading item difficulty

Hongli Li[1], **Roula Aldib**[2], **Chad Marchong**[3], **Kevin Fan**[4]

[1, 3] **Department of Educational Policy Studies**
[2] **Department of Psychology**
Georgia State University
Atlanta, GA.

Abstract

This study examines how well generative AI can predict the difficulty level of reading comprehension items. Four AI tools (i.e., ChatGPT-5, Claude Sonnet 4, Gemini 2.5 Pro, and DeepSeek R1) were evaluated alongside two human raters on 20 items. Claude and Gemini showed the highest agreement with empirical values, in some cases matching or surpassing human raters, while ChatGPT-5 and DeepSeek performed less well. All AI tools and human raters tended to overestimate item ease, though Claude provided the most accurate estimates. These findings suggest that AI tools can complement expert judgment in test development, but empirical validation is necessary for ensuring accurate difficulty estimation.

1 Introduction

In traditional test development, the early stages typically involve field testing to gather pilot data to conduct item analysis. Based on the preliminary results, test items may be revised or discarded as necessary to improve the assessment quality. This process of data collection and analysis is often costly and time-consuming. While generative AI is increasingly recognized for assisting with test development (Bezirhan and von Davier, 2023; Dueñas et al., 2023), its capacity to evaluate item psychometric properties, such as item difficulty, during test development remains unclear.

According to classical test theory (CTT), item difficulty refers to the proportion of test takers who answer an item correctly, known as the p value. Higher p values indicate easier items, while lower p values correspond to more difficult items. A recent study (Li and Marchong, 2024) used ChatGPT to estimate item difficulty for a reading comprehension test and reported moderate correlations ($r > .40$) between ChatGPT's predictions and empirically derived difficulty values. However, that study did not include a direct comparison between AI-generated predictions and human estimates. It remains unclear how AI estimates compare to those made by human experts. Also, it is unknown which AI tools are better suited than others for this task.

To address these gaps, we expand on our previous work by incorporating multiple AI tools and human raters. This study aims to evaluate the effectiveness of AI tools in predicting the difficulty of reading comprehension items in comparison to human raters, with empirical item difficulty (derived from CTT) as the benchmark.

2 Literature Review

Several factors have been systematically identified as influencing reading item difficulty. For instance, Davey (1988) examined a wide range of factors that may contribute to item difficulty, including passage variables (e.g., length, coherence, and syntactic complexity), question types (e.g., response location, inference type), and question format (e.g., stem characteristics and distractor plausibility). Their regression analysis found that stem length and location of response information accounted for a significant amount of variance in item difficulty. Lumley et al. (2012) specifically identified ten factors to predict PISA reading item difficulty, including the number of features and conditions to be comprehended, proximity of pieces of required information, competing information, prominence of necessary textual information, relationship between task and required information, semantic match between task and text, concreteness of information, familiarity of information needed to answer the question, register of the text, and extent to which information from outside of the text is required to answer the question. Overall, studies have shown that both

October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

passage features (e.g., vocabulary sophistication, readability, discourse cohesion) and item features (e.g., stem length, response format, and distractor quality) are associated with reading item difficulty (Choi and Moon, 2020; Davey, 1988; Lumley et al., 2012; Rafatbakhsh and Ahmadi, 2023).

In addition, expert judgment has traditionally been used to estimate item difficulty, often as part of standard-setting or early test development. According to a systematic review conducted by Alkhuzaey et al. (2024), 34% of the included studies compared their systems' predictions with experts' judgement, and on average three experts were recruited per study to judge item difficulty. However, research indicates that expert ratings are subject to bias and inconsistency. For example, Sayın and Bulut (2024) found that although expert predictions of reading item difficulty improved after feedback, their initial ratings often diverged from empirical results. The procedures for expert judgment also vary considerably across studies. In some cases, training was not provided to experts (e.g., Choi and Moon, 2020), or criteria for evaluating difficulty were not clearly specified (e.g., Davey, 1988; Desai and Moldovan, 2019). Such variability raises concerns about the reliability of expert judgments (Alkhuzaey et al., 2024). In sum, while expert review remains common in test development, judgments of item difficulty are often inconsistent and imprecise.

Recently, researchers have begun to explore whether artificial intelligence can provide more consistent predictions of item difficulty than traditional methods. For example, Li, Jiao, and colleagues (2025) modeled item difficulty in large-scale assessments using both small and large language models with different data augmentation strategies. They reported that GPT-4 did not perform as strongly as expected, likely due to limited training data, and suggested that additional data or more advanced reasoning techniques may be required. Their work was based on data from National Board of Medical Examiners (NBME) and the items were about medical practice. This raises the question of whether similar findings extend to reading assessments. It is also unclear whether AI offers advantages over human judgment or simply mirrors its limitations.

Therefore, in this study, we focus on reading comprehension items to investigate whether AI tools (especially LLMs) can provide accurate and reliable estimates of item difficulty. Specifically, we compare predictions from multiple AI tools with human expert ratings and with empirical values derived from examinee responses.

3 Methods

3.1 Instruments and Participants

The reading comprehension test used in this study was a reading section of an English proficiency test. This test evaluates advanced level English language competence of adult non-native speakers of English who plan to use English for academic purposes in a university setting. This test assesses examinees' understanding of college-level reading texts and includes four passages, each followed by five multiple-choice items, for a total of 20 items. Each item has four options, including one answer key and three distractors. All four passages were adapted from newspaper articles. Empirical response data are available from a sample of 2,019 examinees.

The AI tools tested included ChatGPT-5, Claude Sonnet 4, Gemini 2.5 Pro, and DeepSeek R1. ChatGPT-5 is a multimodal model with improved reasoning compared to earlier versions. Claude Sonnet 4 is a medium-sized model with extended context capacity, developed for reasoning and code-related tasks. Gemini 2.5 Pro is also multimodal, with enhanced long-context processing. DeepSeek R1 is an open-source model trained with reinforcement learning, designed to balance reasoning performance with computational efficiency.

Two human raters were invited to provide ratings as well. One was a non-native English speaker with extensive experience teaching English to ESL learners. The other was a non-native English speaker who held a graduate degree and had some experience in literacy research and ESL instruction.

3.2 Data Collection Procedures

Four AI tools were asked to estimate the difficulty of the 20 items on August 30th, 2025. As shown in Appendix A, the authors provided a rating form based on a thorough review of the literature. Below are the sample prompts used to interact with the AI tools:

Researcher: *I have attached a document "rating direction" where you can see the direction of the* *task. I've also provided the document 'Text" which includes the reading comprehension test. Do you understand the task?*

AI tool: *Yes, I understand the task….*

Researcher: *Now I am going to give you a slightly different task. Instead of giving the 5-category rating, can you provide a more nuanced estimation of item difficulty as p value in the classical test theory. Do you understand the task?*

AI tool: *Yes, I understand the task…*

As a result, each of the four AI tools generated both categorical ratings on a 1–5 scale and continuous ratings on a 0–1 scale for each of the 20 reading comprehension items.

In parallel, two human raters independently evaluated item difficulty on a 1–5 scale using the same provided materials; however, they were not asked to provide continuous ratings. To establish a benchmark, empirical difficulty values (p values) were calculated for each item based on the responses of 2,019 examinees.

4 Results

The predictions from both AI tools and human raters were compared to the empirical values with different approaches. First, we used Spearman correlations to examine the association between the estimated categorical difficulty ratings and the empirical p values. The original categorical ratings were coded as 1 = easiest and 5 = hardest. To align the direction of the scales (since higher p values indicate easier items), categorical ratings were reverse coded before correlation analyses. As shown in Figure 1, among the AI tools, Claude demonstrated the strongest alignment with empirical values (ρ = .66), followed by Gemini (ρ = .52) and ChatGPT (ρ = .41). DeepSeek showed the weakest association (ρ = .19). For the human raters, Rater 2 exhibited moderate alignment with empirical values (ρ = .52), while Rater 1 showed lower consistency (ρ = .29). These results suggest that certain AI tools, particularly Claude and Gemini, can approximate empirical item difficulty as well as or better than human raters.

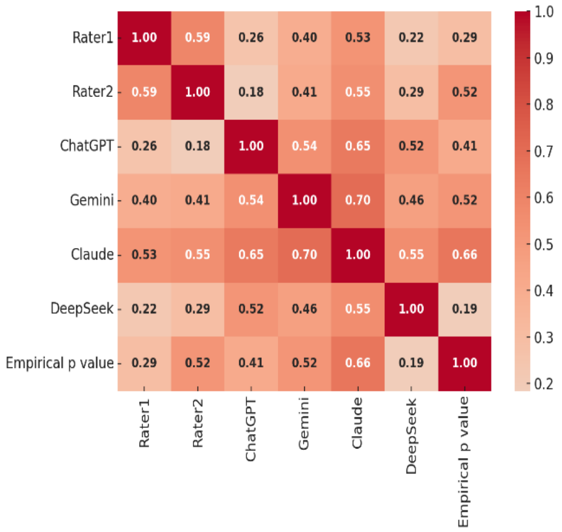


Figure 1: Spearman correlations among categorical ratings by human raters, AI tools, and empirical p value.

We also calculated quadratic weighted kappa (QWK) to evaluate agreement among the two human raters and four AI tools. QWK was selected because it accounts for the ordinal nature of the 1–5 scale and differentially weights disagreements based on their magnitude. As shown in Figure 2, agreement between the two human raters was moderate (κ = .54). Among the AI tools, Claude showed the strongest consistency with both human raters (κ = .44–.47) and other AI tools (κ = .68 with Gemini and κ = .65 with ChatGPT-5). Gemini also demonstrated strong agreement with ChatGPT-5 (κ = .61) and Claude (κ = .68). In contrast, DeepSeek exhibited only moderate agreement with both humans (κ = .24–.32) and the other AI systems (κ = .41–.57). Overall, Claude and Gemini not only aligned most closely with empirical difficulty values but also achieved the highest inter-rater consistency, while DeepSeek showed weaker agreement with others.

Using the AI tools' continuous 0–1 difficulty ratings, we calculated Pearson correlations with the empirical p values. Results mirrored the categorical analysis: Claude showed the strongest association (r = .60), followed by Gemini (r = .57), ChatGPT-5 (r = .43), and DeepSeek (r = .20).

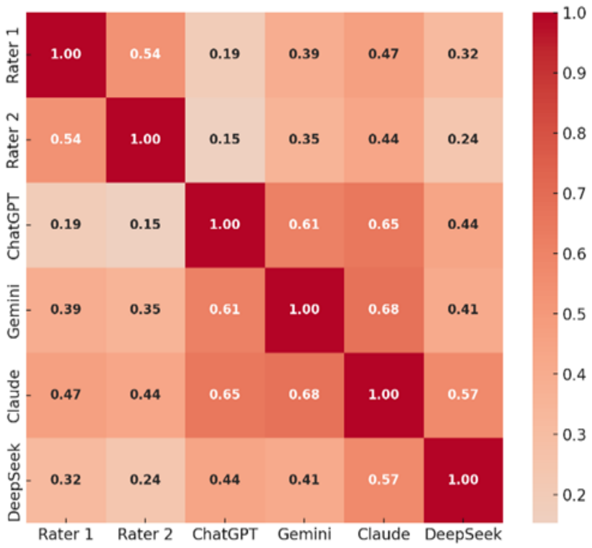


Figure 2: Quadratic weighted kappa agreement between categorical ratings by human raters and AI tools.

Furthermore, as shown in Figure 3, all four AI models systematically overestimated item ease (positive bias). We, therefore, calculated Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between empirical p values and continuous ratings by AI tools. MAE was calculated by averaging the absolute value of the errors, which indicates the average size of the deviations regardless of direction. RMSE was calculated by taking the square root of the averaged squared errors, which is more sensitive to occasional large discrepancies.

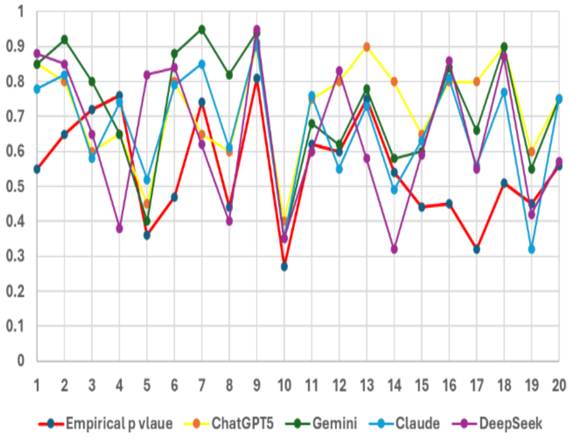


Figure 3: Comparison of continuous ratings by AI tools and empirical p values.

As shown in Table 1, in terms of error magnitude, Claude yielded the lowest mean absolute error (MAE = 0.157) and root mean square error (RMSE = 0.181), indicating the closest alignment with empirical p values. This was followed by Gemini (MAE = 0.186, RMSE = 0.230) and ChatGPT-5 (MAE = 0.204, RMSE = 0.231). DeepSeek R1 showed slightly lower MAE than ChatGPT-5 (0.201 vs 0.204) but the highest RMSE (.244), indicating more large-error outliers. These results suggest that although all AI tools tended to rate items as easier than observed in empirical data, Claude provided the most accurate continuous predictions.

AI Tools	MAE	RMSE
ChatGPT-5	0.204	0.231
Gemini 2.5 Pro	0.186	0.230
Claude Sonnet 4	0.157	0.181
DeepSeek R1	0.201	0.244

Table 1: Error magnitude of AI rating against empirical p values.

5 Discussion

In this study, we found that certain AI tools, particularly Claude and Gemini, can approximate empirical item difficulty as well as, or in some cases better than, human raters. This suggests that AI tools could serve as a valuable supplement to expert ratings in this task. In Li and Marchong (2024), only ChatGPT and the OpenAI API were used to estimate the difficulty of the same 20 reading comprehension items. Their reported Pearson correlations with empirical p values were .48 for ChatGPT-4o and .29 for ChatGPT-4o mini. In the present study, the Pearson correlation between ChatGPT-5 and empirical p values was .43, indicating that ChatGPT has not demonstrated noticeable improvement in predicting reading item difficulty over the past year. By contrast, Claude achieved the highest correlation with empirical p values (r = .60), followed by Gemini (r = .57). These findings suggest that Claude and Gemini currently offer more promising performance than ChatGPT for estimating reading item difficulty.

Our results echo the findings in Li, Jiao and colleagues (2025), who reported that GPT-4 showed limited performance in estimating difficulty of medical practice items, likely due to limited training data. Notably, the RMSE values for GPT-4 in their study (> .35) were higher than those observed in ours, where the RMSE for ChatGPT-5 was .231. This may reflect domain differences. Overall, their findings in medicine and ours in reading comprehension suggest that while LLMs show promise for predicting item difficulty, their effectiveness may depend heavily on model design, training, and the assessment.

Furthermore, in this study, both human raters and AI tools systematically rated items as easier than indicated by the empirical p values. Human experts are prone to underestimating how challenging items are for less proficient examinees, because they are much more proficient than examinees (Nathan and Petrosino, 2003). It seems that AI tools had the same tendency, maybe they are likely to perceive items as easier given their own massive intelligence. In addition, both humans and AI tools may have underweighted the role of distractors in multiple-choice items, which often contribute substantially to empirical difficulty. These findings show the necessity of complementing expert or AI-based predictions with empirical validation.

6 Conclusion

By comparing AI-based predictions to both human expert judgment and empirical values, we aim to understand whether AI models can reliably contribute to the early-stage evaluation of test items. Our findings show that Claude and Gemini achieved the highest agreement with empirical values, outperforming ChatGPT-5 and DeepSeek. Both Claude and Gemini also demonstrated stronger alignment with human raters and were able to predict item difficulty as well as, or in some cases better than, human raters.

However, the four AI tools as well as the two human raters systematically overestimated item ease, though error analyses (MAE, RMSE) suggested Claude provided the most accurate estimates. These findings indicate the potential of AI tools to supplement human judgment in test development; at the same time, they also show the need to include empirical evidence to cross-validate AI-based difficulty estimation.

7 Limitations

While this study provides important insights on the potential of generative AI to support reading assessment development, it also has several limitations. First, the analysis was based on only 20 reading comprehension items drawn from four expository passages. Thus, the findings may not generalize to other item types, genres, or reading assessments. Second, only two human raters were included. A larger pool of experts, possibly with varied backgrounds (e.g., item writers, teachers, researchers), could provide a more reliable benchmark of human judgment. Third, while we used the default outputs of four AI tools, future research could examine how different prompting strategies or fine-tuning approaches influence prediction accuracy.

References

Alkhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914. https://doi.org/10.1007/s40593-023-00362-1

Bezirhan, U., and von Davier, M. 2023. Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5:100161. https://doi.org/10.1016/j.caeai.2023.100161.

Choi, I. C., and Moon, Y. 2020. Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, 17(1):18–42. https://doi.org/10.1080/15434303.2019.1674315.

Davey, B. 1988. Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *The Journal of Experimental Education*, 56(2):67–76. https://www.jstor.org/stable/20151717.

Desai, T., and Moldovan, D. I. 2019. Towards predicting difficulty of reading comprehension questions. In *Proceedings of the Thirty-Second Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pages 8–13. https://cdn.aaai.org/ocs/18267/18267-78886-1-PB.pdf

Dueñas, G., Jimenez, S., and Ferro, G. M. 2023. You've got a friend in... a language model? a comparison of explanations of multiple-choice items of reading comprehension between ChatGPT and humans. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 372–381, Toronto, Canada. Association for Computational

Linguistics. https://aclanthology.org/2023.bea-1.30/

Li, H., and Marchong, C. 2024. Evaluating item difficulty of a reading comprehension test using ChatGPT. Paper presented at the GSU Adult Literacy Research Center (ALRC) Mini-Conference, Atlanta, GA.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., and Lissitz, R. W. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644251344973

Lumley, T., Routitsky, A., Mendelovits, J., and Ramalingam, D. 2012. *A framework for predicting item difficulty in reading tests*. Australian Council for Educational Research (ACER). https://research.acer.edu.au/pisa/5/.

Nathan, M. J., and Petrosino, A. 2003. Expert blind spot among preservice teachers. *American Educational Research Journal*, 40(4):905–928. https://doi.org/10.3102/00028312040004905

Rafatbakhsh, E., and Ahmadi, A. 2023. Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):41. https://doi.org/10.1186/s40862-023-00214-4

Sayın, A., and 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. https://doi.org/10.21449/ijate.1376160

Appendix A: Rating Form

Purpose of the Task

We are conducting a study to estimate the difficulty of 20 reading comprehension items. The test is designed for advanced adult nonnative English speakers preparing for academic study, and measures their ability to understand college-level texts. It consists of four passages adapted from newspaper articles, each followed by five multiple-choice items, for a total of 20 items.

Item difficulty here refers to how likely it is that an average member of the target group (advanced adult non-native English speakers) will answer the item correctly.

Factors You May Consider

1. Linguistic Features

These are characteristics of the reading text itself:

- Vocabulary – Rare words, technical terms, or high lexical density (lots of content words) make comprehension harder.
- Sentence Structure – Longer sentences, more clauses, and complex syntax increase difficulty.

2. Item Features

These are aspects of the test item itself:

- Question Type –
  - *Literal* (answer explicitly in the passage) = easier.
  - *Inference* (paraphrase, bridging, gist, or prior knowledge) = harder.
- Question Phrasing – Negatively worded or confusing stems add difficulty.
- Item Length – Long stems or long answer options increase processing load.
- Answer Options –Longer or more complex distractors make the question harder.

3. Cognitive Demands

These relate to the mental processes required:

- Locating Explicit Information – Easier (requires simple scanning).
- Integrating Across Sentences – Moderate difficulty (requires synthesis of information).
- Higher-Level Inference or Reasoning – Hardest (requires abstraction, generalization, or drawing on prior knowledge).

Please use the provided features holistically, and rely on your expert judgment, experience, and instinct. The goal is to provide your expert impression of relative difficulty.

Rating Scale (5-Point)

Please assign one rating (1–5) for each item:

1. Very Easy – Almost all test-takers are expected to answer correctly.
2. Easy – Most test-takers are expected to answer correctly.
3. Moderate – About half of test-takers are expected to answer correctly.
4. Difficult – Fewer than half of test-takers are expected to answer correctly.
5. Very Difficult – Only a small proportion of test-takers are expected to answer correctly.

Note. Please bear in mind that the target test-takers are advanced non-native English speakers who are seeking to study at English speaking institutions.

When Machines Mislead: Human Review of Erroneous AI Cheating Signals

William Belzak **Chenhao Niu** **Angel Ortmann Lee**
Duolingo, Inc.
`{wbelzak, chenhao, angel.ortmannlee}@duolingo.com`

Abstract

Artificial intelligence (AI) systems are increasingly used to monitor high-stakes online exams, but false positives raise concerns about fairness and validity. To study how human reviewers handle erroneous AI alerts, we intentionally faked "copy-typing" signals and embedded them into authentic exam sessions without proctors' awareness. In two experiments, proctors evaluated these fake signals as part of their normal review process. Study 1 established baseline rejection rates, while Study 2 tested revised guidelines emphasizing corroborating evidence of misconduct. Proctors rejected many fake signals (50–71%), though a notable percentage were still accepted. Rejection rates varied somewhat across test-taker nationalities, and the revised guidelines were associated with more consistent decisions across groups. Guideline updates significantly increased rejections of fake signals but also modestly increased rejections of genuine ones, reflecting a tradeoff between reducing false positives and avoiding false negatives. These findings demonstrate the importance of clear guidance and structured oversight in supporting effective human–AI collaboration in exam security.

1 Introduction

For high-stakes exams, test security involves the deterrence, prevention, and detection of cheating and other forms of misconduct that may artificially inflate a test taker's performance beyond their true proficiency. Breaches in security can undermine the validity of test results and carry serious consequences for examinees and other stakeholders, such as deportation or imprisonment (Main and Watson, 2022; McCray, 2019). As such, test security is essential to ensure that stakeholders can accurately interpret and use test scores (AERA et al., 2014).

In recent years, high-stakes exams have increasingly moved online (Weiner and Hurtz, 2017), reducing costs and broadening access for test takers. This shift, however, introduces new security challenges for providers, from controlling the digital test environment to countering sophisticated technological threats. At the same time, AI creates new opportunities for exam security by providing consistent monitoring at scale, enabling the detection of potential misconduct across large numbers of test sessions more efficiently than human proctors alone.

AI is used in many ways to secure high-stakes remote exams (Dawson, 2020; Zenisky and Sireci, 2021), from verifying identities through facial recognition, keystroke tracking, and voice analysis (Nigam et al., 2021) to proctoring tasks such as flagging when a test taker looks away from the screen (Shih et al., 2024) or when another person is detected in the room. It can also monitor for unauthorized devices, unusual movements, or suspicious sounds, and is increasingly applied to detect plagiarism (Liao et al., 2023), AI-generated answers (Niu et al., 2024), or copy-typing behaviors (Niu et al., 2025).

While AI tools can detect many forms of misconduct, they are not perfect: studies show they can misinterpret benign behaviors as misconduct, raising concerns about accuracy (Tweissi et al., 2022), fairness (Yoder-Himes et al., 2022), and student privacy (Balash et al., 2021). Incorporating human oversight, such as having trained proctors review AI-generated signals, can help reduce false positives (Tweissi et al., 2022) and ensure that decisions about potential misconduct are made with appropriate context. Consistent with Responsible AI standards (Burstein, 2025), we highlight the need to balance technical reliability with human oversight.

2 Background

The rapid expansion of remote testing has made security protocols a central concern. The Duolingo English Test (DET) provides a valuable case study

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 90–97
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

because it integrates AI-based monitoring with human review, offering a real-world setting to evaluate how human decision-makers interact with AI signals in a high-stakes assessment.

2.1 Application

The DET is a remotely administered, high-stakes assessment of English proficiency (Naismith et al., 2025). To protect score integrity, DET employs multiple security measures (Belzak et al., 2025a), including a lock-down digital environment, multi-layered ID verification, an adaptive test design, and standardized administration procedures. Human proctors review test-taker behavior through audio-visual recordings and validate AI-generated signals, ensuring that DET scores remain both reliable and credible. Ultimately, proctors retain final authority in determining whether testing rules were violated or misconduct occurred.

2.2 Copy-Typing Detection

Within the DET security framework, one important safeguard is the detection of potential copy-typing behavior. The DET employs an AI model that analyzes keystroke dynamics, treating a test taker's typing as a sequence of keystroke events and extracting features such as key press durations and intervals between keystrokes. The model architecture captures both local rhythmic patterns and global characteristics across an entire response, enabling it to distinguish between organic composition and transcription from an external source. With the decision threshold selected for this experiment, the estimated false positive rate among test takers who were not copy-typing is about 1%, which represents the upper bound of false positives in the absence of human validation. A full description of the model architecture, feature engineering, and performance is provided in Niu et al. (2025). Although the model has demonstrated strong performance overall, this study did not directly evaluate its construct validity or confidence intervals; readers are referred to prior work for detailed validation (Niu et al., 2025).

When the model flags a test session, the alert is not acted upon automatically. Instead, it is routed to a trained human proctor for review. Proctors are instructed to treat the AI signal as a preliminary alert rather than proof of misconduct. Their primary responsibility is to independently review the audio-visual recording of the session and look for corroborating evidence of cheating as outlined in the proctoring guidelines (see Table 1). Proctors may accept the signal if independent evidence is found, or reject it if no such evidence exists. This ensures a human-in-the-loop validation process.

Even with strong model performance, a small false positive rate can have serious implications in high-stakes testing if unverified alerts are upheld. It is therefore essential to assess whether human proctors can reliably identify and reject false positives, a key safeguard for fairness and test taker protection.

2.3 Research Questions

In this paper, we have three research questions:

1. What percentage of fake copy-typing signals are correctly rejected by proctors?
2. Do rejection rates for fake copy-typing signals differ across test-taker nationalities?
3. Does revising the proctoring guidelines change the likelihood that proctors reject fake copy-typing signals?

We now report on an experiment that aims to answer these questions.

3 Experiment

For this experiment, proctors were asked to accept or reject AI-generated signals indicating potential copy-typing behavior. All signals were intentionally faked, meaning that no test takers were actually flagged for misconduct. These test takers had already received certified DET scores following the standard security review process, ensuring that their results were unaffected.

Fake copy-typing signals were interspersed with real sessions and presented to proctors without their knowledge that any alerts had been faked. Because DET proctoring occurs only after a test is completed, this design was feasible: proctors approached these sessions as part of their normal review process, unaware that the experiment was underway. This allowed us to capture authentic decision-making behaviors under realistic operational conditions.

We ran this experiment twice, first to establish a baseline of rejection rates (Study 1), and second to evaluate how revised proctoring guidelines might have changed those rates (Study 2). Study 1 was conducted from January 30 to February 13, 2025, using the original guidelines. On March 28, 2025, the guidelines were revised to remove "irregular typing patterns" as a criterion, add "presence of an external resource," and to instruct proctors not to apply a copy-typing flag unless suspicious behaviors were observed in the video evidence. These changes were intended to reinforce that proctors should only uphold a copy-typing flag when independent evidence of copy-typing behaviors was present. Study 2 was conducted from July 15 to July 31, 2025, with proctors applying the revised guidelines. Both versions of the guidelines are presented in Table 1.

3.1 Data

For each study, we randomly selected $N = 170$ test sessions that met three conditions: no copy-typing signal was triggered, no misconduct was identified by proctors (all test takers had received certified scores), and the sessions had not been escalated to secondary review for borderline or complex cases. Different sessions were sampled for the two studies because the test content had changed over time and, importantly, to avoid alerting proctors that they might be reviewing the same sessions twice, which could have undermined the realism of the task. Despite being drawn from different time periods, the two sets of sessions showed highly similar distributions of copy-typing detection logits (Figure 1), and the mean values did not differ significantly ($p = 0.433$).

To examine group differences, our analysis focused on three nationality groups—Western (American, Canadian, and French), Chinese, and Indian—as they represented the largest test-taker populations in both our dataset (see Table 2) and the DET (Michalowski et al., 2024). Nationality was selected as a key variable for evaluating group differences in copy-typing decisions, as prior research shows that proctoring outcomes are especially sensitive to this factor. For example, Belzak et al. (2025b) found that both proctor and test-taker nationality influenced the likelihood of being flagged for rule violations, whereas other demographic characteristics such as gender and age did not. Table 3 presents the broader distribution of proctor nationalities aggregated by continent. These data reveal that the majority of proctors are based in the Americas (46%) and Europe (24%), a distribution that contrasts with the larger populations of Chinese and Indian test takers.

In addition to the fake signals, we also collected operational sessions that had been flagged with genuine copy-typing signals during the two periods before and after the guideline revision. These data allowed us to examine whether changes in proctoring instructions influenced how proctors handled authentic AI alerts, providing a real-world complement to the experimental results based on fake signals.

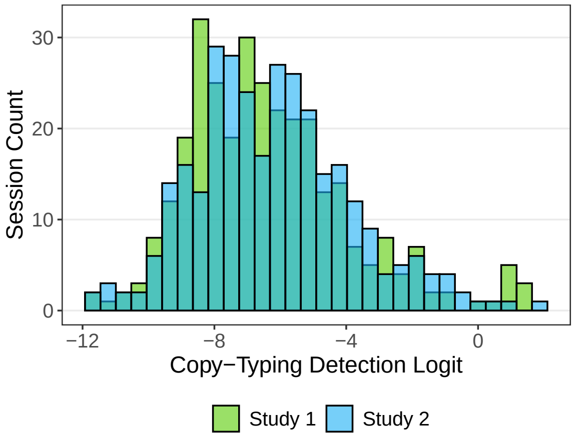


Figure 1: Distribution of copy-typing detection model predictions for sessions used in two studies, with the average values and standard deviations being -6.20 ± 2.36 for Study 1 and -6.36 ± 2.46 for Study 2.

3.2 Methods

We first estimate the probability of proctors rejecting fake copy-typing signals by fitting a logistic mixed-effects model (Raudenbush and Bryk, 2002) to the combined data from both studies:

$$\text{logit}\big(Pr(S_{ij} = 0)\big) = \beta_0 + \beta_1 X_i + u_j, \quad (1)$$

where $Pr(S_{ij} = 0)$ is the probability that proctor j rejects signal i, $X_i = 0$ under the original guidelines and $X_i = 1$ under the revised guidelines, and $u_j \sim N(0, \tau^2)$ is a random effect for proctor j, which accounts for non-independence because each proctor evaluated multiple signals. In this model, β_0 represents the baseline log-odds of rejection under the original guidelines, while β_1 captures the log-odds change after the revision.

Next, we examine nationality effects in two stages. In the first stage, we fit within-study models to test for differences among Chinese, Indian, and Western test takers:

$$\text{logit}\big(Pr(S_{ij} = 0)\big) = \beta_0 + \beta_1 C_i + \beta_2 I_i + u_j, \quad (2)$$

where $C_i = 1$ for Chinese test takers and 0 otherwise, $I_i = 1$ for Indian test takers and 0 otherwise, with Western test takers (American, Canadian, French) as the reference group. Here, β_0

Original Guidelines	**Revised Guidelines**
Review corresponding video segments for suspicious behaviors that indicate copy-typing, including: • Irregular Typing Patterns • Unusual Body Movements • Irregular Eye Movements	Review corresponding video segments for suspicious behaviors that indicate copy-typing, including: • ~~Irregular Typing Patterns~~ • Unusual Body Movements • Irregular Eye Movements • *Presence of an external resource* ***Do not apply** this signal flag if no suspicious behaviors are observed.*

Table 1: Proctoring guidelines for reviewing copy-typing signals. Revisions are marked as ~~deleted~~ or *added*.

Nationality Group	**Study 1**	**Study 2**
Western*	25	30
Chinese	28	17
Indian	15	19

Table 2: Number of test takers by nationality group in Study 1 and Study 2. The Western nationality group includes American, Canadian, and French test takers.

Continent	**Percentage**
Americas	46%
Europe	24%
Asia	13%
Africa	10%
Oceania	6%

Table 3: Percentage of proctor nationalities aggregated by continent.

gives the baseline log-odds of rejection for Western test takers, while β_1 and β_2 capture contrasts for Chinese and Indian test takers, respectively. In the second stage, we use the same model specification as Eq. 1, but fit it separately within each nationality group to assess between-study differences. This allows us to test whether rejection rates changed significantly from Study 1 (original guidelines) to Study 2 (revised guidelines) within each nationality.

Finally, we estimate the probability of rejecting *genuine* copy-typing signals observed during operational proctoring. The model includes the guideline condition (original vs. revised) as a predictor, specified in the same way as Eq. 1. Unlike Eq. 1, however, we use a logistic regression model rather than a mixed-effects model, since each genuine signal was reviewed by only one proctor.

For all models, model-implied probabilities and percentages are obtained by applying the inverse-logit transformation to the estimated log-odds coefficients.

4 Results

Figure 2 shows the model-implied percentages of proctors rejecting fake copy-typing signals under the original and revised guidelines. Results from the logistic mixed-effects model indicate that rejection rates were significantly higher after the revision, with an estimated effect of $\hat{\beta}_1 = 0.880$ ($p = .001$) on the log-odds scale.

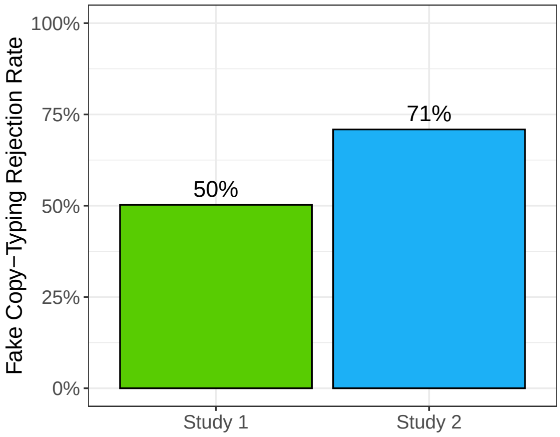


Figure 2: Model-implied percentages of rejecting fake copy-typing signals before (Study 1) and after (Study 2) the revision of proctoring guidelines.

Figure 3 shows the model-implied percentages of proctors rejecting fake copy-typing signals by test-taker nationality and study. The within-study logistic mixed-effects models revealed systematic nationality differences. In Study 1, Chinese test takers were significantly less likely than Western test takers to have fake signals rejected ($\hat{\beta}_1 = -2.230$, $p < .001$). Indian test takers also showed lower rejection rates than Western test takers, though this difference was marginally statistically significant ($\hat{\beta}_2 = -1.159$, $p = .088$). In Study 2, Chinese test takers again exhibited lower rejection rates than Western test takers, but the difference was not significant ($\hat{\beta}_1 = -0.996$, $p = .187$). By contrast, Indian test takers were significantly less likely than Western test takers to have fake signals rejected ($\hat{\beta}_2 = -1.553$, $p = .029$).

The between-study logistic mixed-effects models also revealed systematic differences by nationality. Chinese test takers showed significantly higher rejection rates in Study 2 compared to Study 1 ($p = .001$). Rejection rates for Western and Indian test takers also increased across studies, but these effects did not reach statistical significance ($p = .104$ and $p = .303$, respectively).

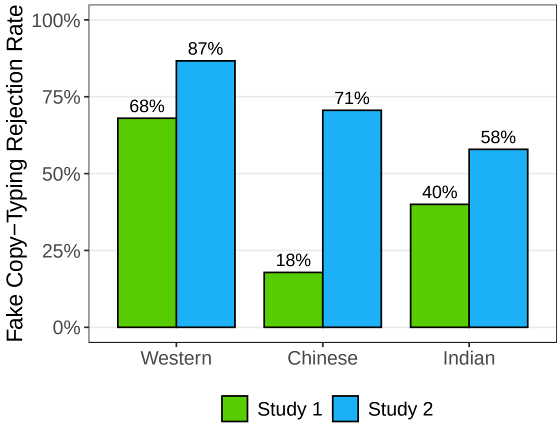


Figure 3: Model-implied percentages of rejecting fake copy-typing signals by test-taker nationality and study.

Figure 4 shows the model-implied percentages of proctors rejecting both fake and genuine copy-typing signals under the original and revised guidelines. Rejection rates for fake signals were substantially higher (50–71%) than for genuine signals (9–13%). The logistic model also indicated a small but statistically significant increase in the rejection of genuine signals after the revision ($p = .003$). We discuss the implications of these findings in the next section.

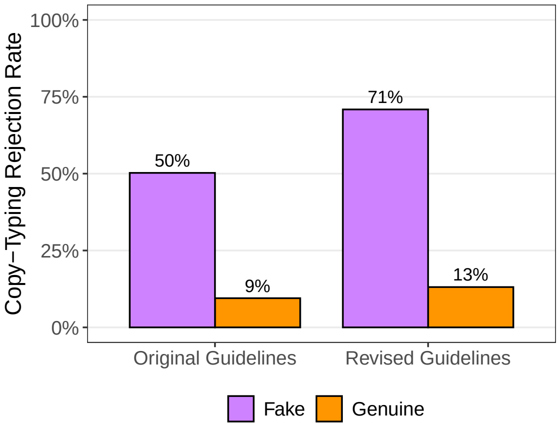


Figure 4: Model-implied percentages of rejecting fake and real copy-typing signals under the original and revised guidelines.

5 Discussion

This experiment examined how trained human proctors interact with AI-generated copy-typing signals in a remotely administered, high-stakes English language assessment. By intentionally fabricating signals and asking proctors to accept or reject them under specific guidelines, we evaluated three questions: (1) whether proctors could correctly reject fake AI signals, (2) whether rejection rates varied by test-taker nationality, and (3) whether revised guidelines changed proctoring decisions.

5.1 Research Question 1: Percentage of Fake Signal Rejections

The first research question asked: *What percentage of fake copy-typing signals are correctly rejected by proctors?* Across both studies, proctors rejected a substantial percentage of fake signals (50–71%). This suggests that proctors can often identify when an AI alert is not supported by independent evidence of misconduct. However, the fact that roughly one in three fake alerts was accepted underscores the risks of overreliance on AI signals in high-stakes contexts (Skitka et al., 1999; Poursabzi-Sangdeh et al., 2021). These findings highlight the value of human review for maintaining fairness and accuracy, while also pointing to the need for additional safeguards to minimize the consequences of false positives (Bansal et al., 2021). As AI detection models improve, false positive rates, and the role of human reviewers in rejecting them, may shift. More research will be critical to ensure systems remain both reliable and fair.

5.2 Research Question 2: Nationality Differences in Rejection Rates

The second research question asked: *Do rejection rates for faake copy-typing signals differ across test-taker nationalities?* The within-study analyses revealed systematic differences: in Study 1, Chinese test takers were significantly less likely to have fake signals rejected than Western test takers, and Indian test takers showed a similar trend. In Study 2, rejection rates for Chinese test takers improved and were no longer significantly different from Western test takers, while Indian test takers were significantly less likely to have fake signals rejected. The between-study analyses confirmed that rejection rates increased significantly for Chinese and Western test takers in Study 2, but not for Indian test takers. These findings suggest that proctor decision-making can vary by nationality group, perhaps due to a structural mismatch between the distribution of test takers and proctors (Belzak et al., 2025b), and that revised guidelines may reduce some differences while leaving others unaddressed. Strategies such as targeted proctor training, bias monitoring dashboards, and regular fairness audits could help ensure that future revisions to proctoring guidelines improve accuracy while also addressing inequities across groups.

5.3 Research Question 3: Effect of Revised Guidelines

The third research question asked: *Does revising the proctoring guidelines change the likelihood that proctors reject fake copy-typing signals?* The evidence indicates that they do. After the guidelines were updated to emphasize the need for corroborating evidence of misconduct, rejection rates of fake signals increased significantly. This suggests that proctoring practices are sensitive to instructional framing and that targeted revisions can improve decision quality (Association of Test Publishers and National College Testing Association, 2024; Buçinca et al., 2021). However, the revised guidelines were also associated with a small but statistically significant increase in the rejection of genuine signals, suggesting that proctors became more cautious about accepting AI alerts but also more likely to dismiss valid cases (Almog et al., 2024). This tradeoff between reducing false positives and increasing false negatives highlights the complexity of calibrating human–AI collaboration in high-stakes testing. More research is needed to refine this balance and identify guidelines that reduce risks without undermining security.

5.4 Implications

Taken together, the findings underscore both the value and limitations of human-in-the-loop AI systems in exam security. Proctors are capable of rejecting false positive copy-typing signals, but not always uniformly across nationalities, and their decisions are shaped by the guidance they receive. Ongoing training, carefully designed guidelines, and continuous monitoring of decision patterns are therefore essential to ensure fairness and validity (Burstein et al., 2025).

5.5 Limitations and Future Work

Several limitations should be considered when interpreting these findings. First, the two studies were conducted on different sets of test sessions and several months apart. This was necessary because the test's visual design had changed, and reusing the same sessions could have signaled to proctors that they were artificial. However, this design also means that unobserved differences in session characteristics or other contextual changes over time may have contributed to the observed effects, making it difficult to attribute differences solely to the revised guidelines.

Second, for genuine copy-typing signals observed during operational proctoring, we cannot determine whether higher rejection rates reflect proctors dismissing false positives or overlooking true positives. As such, estimates of genuine-signal rejection rates should be interpreted with caution. Establishing verified ground truth through simulated or confirmed cases of misconduct would strengthen future studies.

Third, the scope of this experiment was limited in terms of sample size, signal type, and use of fake copy-typing signals. The relatively small samples constrained analyses of nationality differences, and fake signals—while useful for preserving realism—may not capture the full complexity of genuine AI alerts. Moreover, we focused on copy-typing signals only; other alerts, such as those for unusual movements, unauthorized devices, or suspicious sounds, may pose different challenges for human validation. Future work should expand to larger, more diverse datasets and a broader range of signal types to better understand the dynamics of human–AI collaboration in exam security.

Addressing these limitations will be essential for improving both the accuracy and fairness of AI-assisted proctoring. Larger datasets, verified ground truth, and broader signal coverage will help test providers calibrate human–AI decision-making and safeguard the integrity of high-stakes assessments.

6 Conclusion

This study examined how trained human proctors interact with AI-generated copy-typing signals in a high-stakes, remotely administered English language assessment. To ensure operational outcomes were unaffected, fake signals were embedded only into completed sessions where test takers had already received certified scores. This design allowed us to evaluate proctor decision accuracy in rejecting fake AI signals, explore differences across nationality groups, and assess the impact of revised guidelines without altering test results.

Proctors generally identified and rejected fake copy-typing signals, but acceptance of some signals highlights the risks of overreliance on AI. Rejection rates varied by nationality, with differences reduced but not eliminated under revised guidelines. The guidelines also increased rejections of fake signals while slightly raising rejections of genuine ones, underscoring the tradeoff between false positives and false negatives. More research is needed to examine these dynamics in larger datasets, across different AI signals, and in varied testing contexts.

Overall, the findings illustrate both the promise and limits of human-in-the-loop AI for exam security. Clearer guidelines, regular training, and monitoring are essential to support fairness and validity. Practically, testing organizations can refine proctor training, track nationality-related outcomes, and calibrate AI–human collaboration to balance accuracy and fairness. Because the reliability of human–AI systems depends not only on technical performance but also on governance, transparent processes, and oversight, exam security frameworks should be aligned with Responsible AI standards (Burstein et al., 2025). More broadly, stakeholder trust in high-stakes assessments also rests on adherence to the principles of fairness and validity articulated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Acknowledgments

Generative AI was used to review and refine text originally drafted by the authors, improving readability and flow. All content was verified for accuracy. We thank Alina von Davier and Jill Burstein for their thoughtful feedback on the draft.

References

AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

David Almog, Romain Gauriot, Lionel Page, and Daniel Martin. 2024. AI oversight and human mistakes: evidence from centre court. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 103–105.

Association of Test Publishers and National College Testing Association. 2024. Assessment industry standards and best practices for the online observation of tests. `https://www.testpublishers.org/assets/Online%20Observation%20of%20Tests%20Standards%20for%20Public%20Comment%202024.3.23.1207%20numbered.pdf`. Association of Test Publishers & National College Testing Association.

David G Balash, Dongkun Kim, Darika Shaibekova, Rahel A Fainchtein, Micah Sherr, and Adam J Aviv. 2021. Examining the examiners: Students' privacy and security perceptions of online proctoring services. In *Seventeenth symposium on usable privacy and security (SOUPS 2021)*, pages 633–652.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

William Belzak, Basim Baig, Ramsey Cardwell, Rose Hastings, Andre Horie, Geoff LaFlair, Manqian Liao, Chenhao Niu, and Yong-Siang Shih. 2025a. Duolingo english test: Security and score integrity. Duolingo Research Report DRR-25-04, Duolingo English Test.

William Belzak, Jill Burstein, and Alina A. von Davier. 2025b. Evaluating fairness in AI-assisted remote proctoring. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 125–132.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Jill Burstein. 2025. Duolingo english test: Responsible ai standards. Duolingo Research Report DRR-25-05, Duolingo.

Jill Burstein, Geoffrey T. LaFlair, Kathleen Yancey, Alina A. von Davier, and Rotem Dotan. 2025. Responsible ai for test equity and quality: The duolingo english test as a case study. In Earl M. Tucker, Eleanor Armour-Thomas, and Edmund W. Gordon, editors, *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning*. University of Massachusetts Amherst Library Press.

Phillip Dawson. 2020. *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.

Manqian Liao, Sinon Tan, and Baig Basim. 2023. Plagiarism detection using human-in-the-loop AI. *Paper presented at the annual meeting of the National Council on Measurement in Education*.

Ed Main and Richard Watson. 2022. The english test that ruined thousands of lives. `https://www.bbc.com/news/uk-60264106`. Accessed: 26 August 2025.

Vanessa McCray. 2019. Judge in APS cheating trial to remain on case as six seek retrial. *The Atlanta Journal-Constitution*. Accessed via The Atlanta Journal-Constitution.

Allison Michalowski, Ramsey Cardwell, Steven Nydick, and Ben Naismith. 2024. Duolingo english test: Demographic and score properties of test takers. Technical report, Duolingo Research Report). Duolingo. https://go. duolingo. com/demographic-score.

Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Masha Kostromitina. 2025. *Duolingo English Test: Technical Manual*. Duolingo, Inc. Last updated July 2025.

Aditya Nigam, Rhitvik Pasricha, Tarishi Singh, and Prathamesh Churi. 2021. A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26(5):6421–6445.

Chenhao Niu, Yong-Siang Shih, Manqian Liao, Ruidong Liu, and Angel Ortmann Lee. 2025. Keystroke analysis in digital test security: AI approaches for copy-typing detection and cheating ring identification. *Proceedings of Artificial Intelligence in Measurement and Education Conference*.

Chenhao Niu, Kevin P. Yancey, Ruidong Liu, Mirza Basim Baig, André Kenji Horie, and James Sharpnack. 2024. Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 940–953, Miami, Florida, US. Association for Computational Linguistics.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Stephen W Raudenbush and Anthony S Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage.

Yong-Siang Shih, Zach Zhao, Chenhao Niu, Bruce Iberg, James Sharpnack, and Mirza Basim Baig. 2024. AI-assisted gaze detection for proctoring online exams.

Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.

Adiy Tweissi, Wael Al Etaiwi, and Dalia Al Eisawi. 2022. The accuracy of AI-based automatic proctoring in online exams. *Electronic Journal of e-Learning*, 20(4):419–435.

John A Weiner and Gregory M Hurtz. 2017. A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, pages 13–20.

Deborah R Yoder-Himes, Alina Asif, Kaelin Kinney, Tiffany J Brandt, Rhiannon E Cecil, Paul R Himes, Cara Cashon, Rachel MP Hopp, and Edna Ross. 2022. Racial, skin tone, and sex disparities in automated proctoring software. In *Frontiers in Education*, volume 7, page 881449. Frontiers Media SA.

April L. Zenisky and Stephen G. Sireci. 2021. The impact of technology on test security. In Jeffrey K. Smith, editor, *The Oxford Handbook of Educational Measurement*, pages 386–402. Oxford University Press.

Fairness in Formative AI: Cognitive Complexity in Chatbot Questions Across Research Topics

Alexandra Barry Colbert
College Board
acolbert@collegeboard.org

Karen D. Wang
San Jose State University
karen.wang02@sjsu.edu

Abstract

This study evaluates whether questions generated from a Socratic-style AI chatbot designed to support project-based AP courses maintain cognitive complexity parity when inputted with research topics of sensitive and non-sensitive nature. We present empirical findings indicating no significant conversational complexity differences, highlighting implications for equitable AI use in formative assessment.

1 Introduction

AI systems are increasingly deployed in educational settings for instructional and assessment support (U.S. Department of Education, 2023; Samala et al., 2025). These systems offer unprecedented opportunities for providing students with timely, individualized support at scale. Recent work demonstrates that AI systems can generate discussion prompts, assessment items, and lesson plans across various academic subjects (Scaria et al., 2024; Zheng et al., 2024; Maity et al., 2025). One promising direction is to adapt large language models (LLMs) as Socratic questioning chatbots. Specifically, AI-generated questions can prompt students to elaborate on their thinking, make connections between different topics, and develop critical thinking and metacognitive skills (Danry et al., 2023; Al-Hossami et al., 2024; Hung et al., 2024; Favero et al., 2024). This approach may help address concerns about AI tools diminishing critical thinking abilities through cognitive offloading (Gerlich, 2025), as Socratic questioning actively engages students in reflective reasoning rather than providing ready-made solutions. At the same time, successful implementations of AI systems in educational settings require careful examination of how these systems operate across varied content areas and student populations (Garcia Ramos and Wilson-Kennedy, 2024; Kaldaras et al., 2024). In the case of AI systems designed for Socratic dialogue, concerns persist regarding potential algorithmic bias in the way LLM generates questions based on the varying content and perspectives found in student writing.

Existing research on LLM question generation in educational settings has focused on creating assessment items from structured course materials such as textbooks, lesson plans, and existing question banks (Tran et al., 2023; Chan et al., 2025; Shu et al., 2025). While these studies show promising results for automated question generation for STEM subjects, there remains limited investigation into how LLMs generate questions in response to student-produced writing, particularly essays addressing potentially controversial or sensitive topics. This research gap is significant because LLMs may have learned implicit associations during pretraining that influence their question generation patterns when encountering politically charged or socially sensitive content compared to neutral academic materials (Westwood et al., 2025; Bai et al., 2025). The absence of systematic investigation into how LLMs perform when generating questions based on student essays creates uncertainty about their reliability and fairness in real-world educational contexts when students write about various topics.

Our investigation in this study is guided by the following research question: to what extent does the nature of student essay topics influence the characteristics of LLM-generated, student-facing questions? Specifically, we are interested in whether students choosing potentially controversial or sensitive research topics may receive LLM-generated questions that differ in cognitive complexity from those posed to students with more neutral research topics. This study is part of a broader project at College Board's GenAI Studio to develop AI-powered educational tools in collaboration with AP students and educators, learning scientists, learning engineers, psychometricians, and ELA content experts.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 98–106
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

An LLM-based, student-facing chatbot ("SAM") is prompt-engineered to generate Socratic questions and support students' deeper learning of the research process and critical thinking skills. Given the variety of interactions students might take with SAM, it is crucial to explore, prior to classroom deployment, equitable performance across different student inputs.

2 Methodology

2.1 Synthetic Student Work Samples

To ensure breadth of topic coverage and a balanced distribution of sensitive and non-sensitive essay topics, we first used an LLM to simulate student work samples based on 47 suggested themes in AP Seminar research tasks. In the AP Seminar course, to demonstrate research progress, students are required to submit stepwise evidence of their research process, including (1) an Annotated Bibliography or source chart including citations and summaries of source content, (2) an argument Outline which organizes and synthesizes sources, and (3) a completed research paper. For this study, samples were generated for the first two distinct work phases (Annotated Bibliography and Outline) and aim to represent progressing completion statuses of each phase (Severely Underdeveloped, First Draft, Final Draft). We did not generate completed papers as the chatbot is focused on providing formative support.

We used AWS Bedrock Flows to automate synthetic work sample generation through a set of chained prompts against Claude 3.5 Sonnet (see Figure 1). Given a sample theme from AP Seminar Course and Exam Description (CED) (e.g., Utopia) and topic sensitivity ("is controversial or sensitive in nature" or "is not controversial or sensitive in nature"), a topic was generated. Topic sensitivity was loosely defined in the prompt and left to the model's interpretation. The generated topics were reviewed by a human reviewer to ensure the sensitivity distinctions were reasonable. Table 1 presents sample sensitive and non-sensitive topics based on five theme inputs.

Once a topic was generated, the flow branched to generate five subsequent work samples: Annotated Bibliography - Severly Underdeveloped, Annotated Bibliography - First Draft, Annotated Bibliography - Final Draft, Outline - First Draft, and Outline - Final Draft. The zero-shot prompts requested content "at the level of a 10th grader". This process generated 235 synthetic student work samples for sensitive topics and 235 for non-sensitive topics. Table 2 summarizes the distribution of synthetic work samples across work phase, development status, and topic sensitivity.

2.2 Chatbot Configuration and Question Generation

SAM is designed to help students reflect on their research process and provide teachers with evidence of student engagement with their research topics. The version of SAM tested in this study harnesses in-context learning capabilities of Claude 3.5 Sonnet. This foundational model was chosen for its Constitutional AI framework, which uses reinforcement learning to promote safe and fair chat experiences (Bai et al., 2022). SAM's system prompt defines its role as a Socratic mentor and includes instructions related to conversation management, a rubric against which to support question generation, and safety measures to enforce AP-appropriate interactions. For example, in the following snippet of the Annotated Bibliography prompt, SAM is instructed to plan out its conversation with students by generating a starter question of specific types in its chain of thought.

```
Create a starter question for each of the
following, in order. If elements of the
work are not present in the work draft,
do not skip questions, but adapt the
questioning to let the student advocate
for work that has not materialized yet
or how they plan on approaching it.
Output the questions in
<conversation_plan/> tags.
1. A summary question
2. A decision question
3. An evidence question
4. A speculative question
5. A reflection question
```

SAM's line of questioning includes question types tailored to the work phase and span the breadth of the research experience (summary, decision, evidence, speculative, and reflection). The system prompt includes sample questions developed by learning scientists and learning engineering specialists, which serve as few-shot examples to guide SAM's question generation (see Table 5). In order to balance personalization with standardization, SAM was encouraged to ground questions in the provided student work, but also was config-

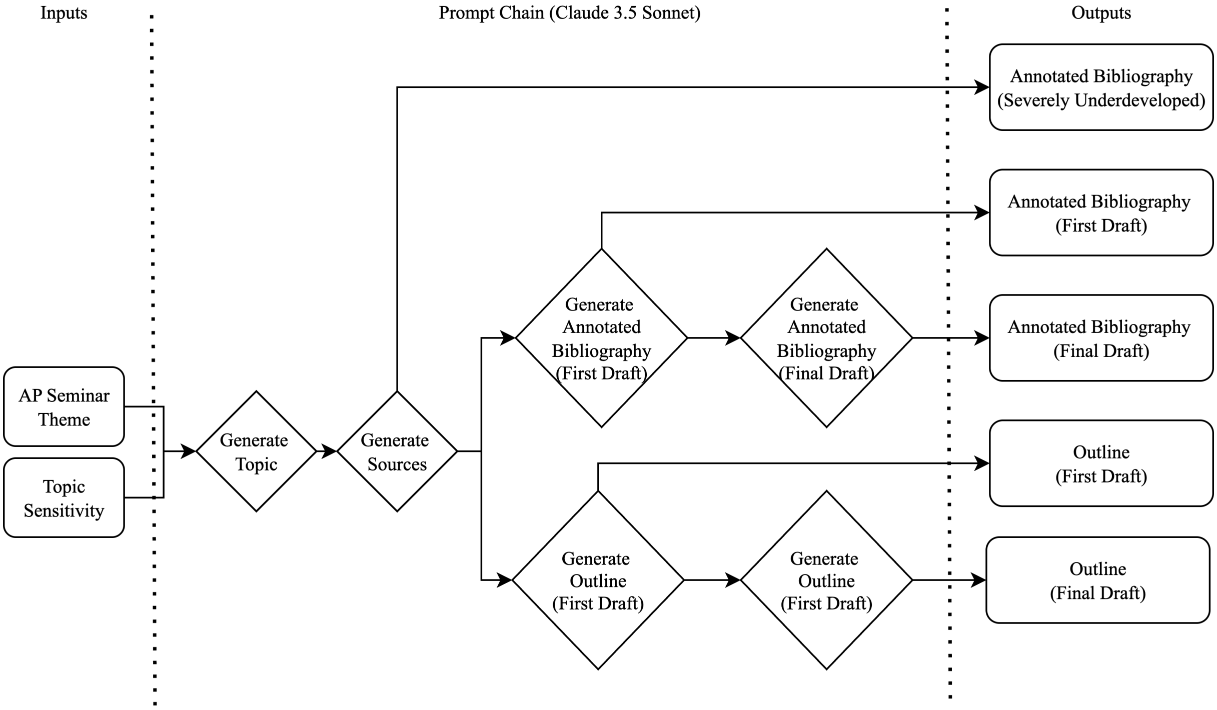

Inputs
Prompt Chain (Claude 3.5 Sonnet)
Outputs
AP Seminar Theme
Topic Sensitivity
Generate Topic
Generate Sources
Generate Annotated Bibliography (First Draft)
Generate Annotated Bibliography (Final Draft)
Generate Outline (First Draft)
Generate Outline (First Draft)
Annotated Bibliography (Severely Underdeveloped)
Annotated Bibliography (First Draft)
Annotated Bibliography (Final Draft)
Outline (First Draft)
Outline (Final Draft)

Figure 1: AWS Bedrock Flows automated generation of synthetic student work samples from input themes and topic sensitivities

Theme	**Sensitive Topic**	**Non-Sensitive Topic**
Evolution	*The role of eugenics in modern evolutionary theory*	*The evolutionary adaptations of migratory bird species*
Home	*The impact of home ownership on wealth inequality*	*The impact of smart home technology on energy efficiency*
Language	*The impact of gendered language on societal perceptions and biases*	*The impact of bilingual education on cognitive development in children*
Place	*The impact of gentrification on marginalized communities*	*The impact of urban green spaces on community well-being*
Rights and responsibilities	*The right to die*	*The role of individual rights and civic responsibilities in a democratic society*

Table 1: Sample Synthetic Topics by Theme and Topic Sensitivity

Work Phase	**Development Status**	**Topic Sensitivity**	
		Sensistive	Non-Sensitive
Annotated Bibiography	Severely Underdeveloped	47	47
	First Draft	47	47
	Final Draft	47	47
Outline	First Draft	47	47
	Final Draft	47	47
Total		235	235

Table 2: Synthetic Work Input Counts across Work Phase, Development Status, and Topic Sensitivity (n=470)

ured with a temperature of 0 for reproducibility. As a result, some of the questions generated by SAM closely matched the few-shot examples, while others referenced specific topics, sources, and arguments from the student work input.

Each of the 470 synthetic student work samples was submitted to SAM, which generated a conversation plan containing five starter questions for each sample. In total, this process produced 2,350 Socratic-style questions for subsequent coding and analysis.

2.3 Depth of Knowledge (DOK) Coding

Depth of Knowledge (DOK) was identified as a suitable framework for measuring cognitive complexity of SAM's questions. DOK measures the complexity of thinking required to complete a task and focuses on context and depth of the task (Christopherson and Webb, 2024), making it appropriate for comparison of rigor across questions. An expert human rater labeled the DOK level of each unique SAM question, blinded to each output's topic sensitivity condition. Table 3 presents sample SAM outputs labeled by their DOK level.

2.4 Statistical Analysis of Conversation DOK Levels

For samples that had four or more classifiable questions, Conversation DOK was calculated as the mean Depth of Knowledge value across questions in each conversation. To compare cognitive complexity between groups, an independent two-sample t-test was performed on mean Conversation DOK levels by topic sensitivity. To control for multiple comparisons, a Bonferroni correction was applied.

3 Results

Of the 2,350 Socratic-style questions produced by SAM, 1,062 were unique, as the same questions can be generated for multiple work samples. SAM's most frequent questions across conversations are reported in Table 6.

3.1 Question DOK by Work Phase and Topic Sensitivity

Figure 2 compares the distribution of DOK labels across work phases for SAM's questions. 5.9% of Annotated Bibliography questions and 33.4% of Outline questions were unclassifiable. One reason for the high volume of unclassifiable questions in the Outline phase was that many of SAM's questions followed a sample question in the prompt, which elicited "yes"/"no" responses that prevented accurate cognitive complexity classification. Given that one third of Outline questions were unclassifiable, including multiple unclassifiable questions per conversation, conversation-level DOK could not be reliably calculated for Outline conversations. For this reason, we focus the DOK comparison for the questions generated for Annotated Bibliography work samples only.

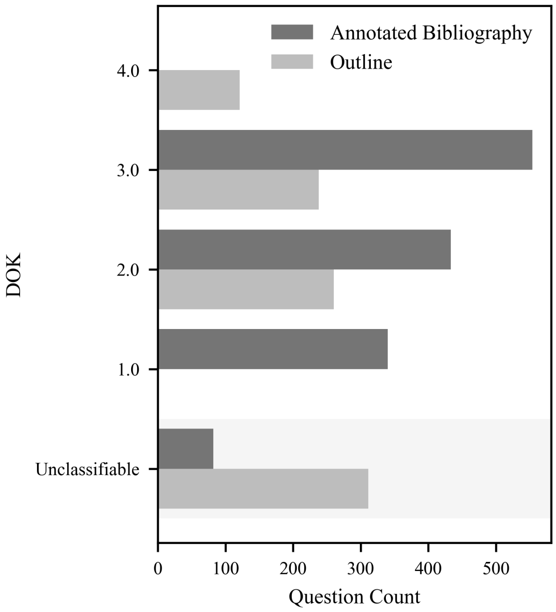


Figure 2: Question DOK (n=2,350) by Work Phase

Figure 3 visualizes question DOK distribution across the 282 Annotated Bibliography conversations, comparing those generated from (a) sensitive and (b) non-sensitive topics. Read vertically, each subplot illustrates the progression of conversational DOK over time. The conversations typically start with a question of low cognitive complexity (DOK of 1) and end with a question of high cognitive complexity (DOK of 3). Visual comparison of these conversation trajectories reveals no substantial differences in DOK patterns between sensitive and non-sensitive topics across all five questions.

3.2 Conversation DOK by Topic Sensitivity

We calculated conversation-level DOK scores for all 282 Annotated Bibliography conversation plans to quantitatively assess whether topic sensitivity influenced the cognitive complexity level of LLM-generated questions. Figure 4 displays the distribution of conversation-level DOK scores across sen-

DOK Level	SAM Question	Question Type	Work Phase
Level 1 (Recall)	*What is the main focus of your research based on these sources?*	Summary	Annotated Bibliography
Level 2 (Skills and Concepts)	*What is your current research topic, and how did it evolve as you developed the outline?*	Decision	Outline
Level 3 (Strategic Thinking)	*Which of these sources do you find most influential in shaping your understanding of entrepreneurial opportunities and small business success?*	Evidence	Annotated Bibliography
Level 4 (Extended Thinking)	*Reflecting on your research and outline development process, what have you learned about conducting research on such a complex topic?*	Reflection	Outline
Unclassifiable	*Based on your research so far, can you think of any potential alternatives to the death penalty that might address the concerns raised by both supporters and opponents?*	Speculative	Annotated Bibliography

Table 3: Sample SAM Questions by DOK Level

	Sensitive	Non-sensitive
Conversation Count	141	141
Mean	2.152	2.153
Standard Deviation	0.303	0.299

Table 4: Descriptive Statistics of Conversation DOK by Topic Sensitivity (n=282)

sitive and non-sensitive topic groups, with descriptive statistics presented in Table 4. Levene's test confirmed homogeneity of variances, $F(1, 280) = 0.010$, $p = 0.919$, satisfying the assumptions for parametric testing. An independent samples t-test comparing group means revealed no significant difference in conversation-level DOK between sensitive and non-sensitive topic, $t(280) = -0.020$, $p = 0.984$. This result suggests that SAM generated questions with equivalent levels of cognitive challenge, regardless of topic sensitivity.

3.3 A Note about Potential Bias in Questions

Even though the expert labeler was not specifically tasked with identifying bias in the questions generated by SAM, they did surface five outputs that exhibited nuanced bias. These five questions followed a consistent pattern: *How might your research change if you were to include more sources from [an ethnic/religious/political group] perspectives?*

All five instances were speculative questions, likely generated because AP instructional materials on synthesizing perspectives were incorporated into SAM's system prompt. These outputs present a dual interpretation challenge. On the one hand, such questions could be interpreted as problematically inviting students to presume that all members of a particular group share the same view of an issue. On the other hand, such questions might have value in encouraging students to include perspectives from communities that they had not previously considered.

While these five outputs were generated from sensitive topics, the data is not sufficient to establish a systematic relationship between topic sensitivity and problematic question generation. Fisher's exact test revealed no significant association after multiple comparison correction ($p = 0.030 > 0.00625$, OR $= 11.1$, 95% CI: 0.62–202). Future research conducting targeted bias assessment at the individual question level is needed to determine whether topic sensitivity, question type, or other variables introduce biased questioning patterns.

4 Discussion

This study examined whether cognitive complexity of LLM-generated questions differs based on the sensitivity of student research topics. Our analysis

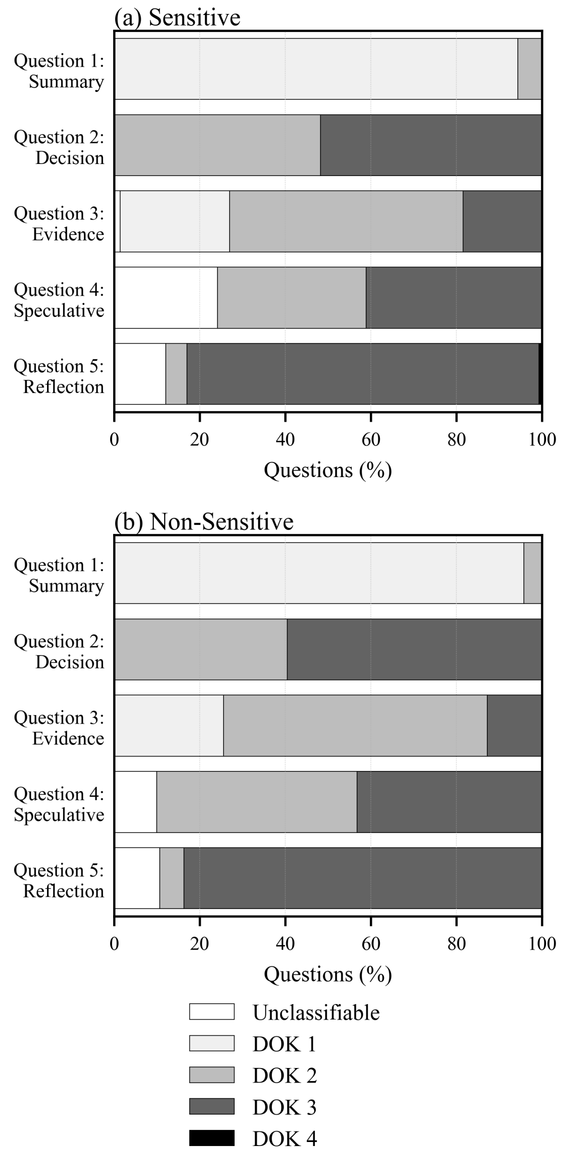


Figure 3: Question DOK (n=2,350) of Annotated Bibliography conversations by Topic Sensitivity

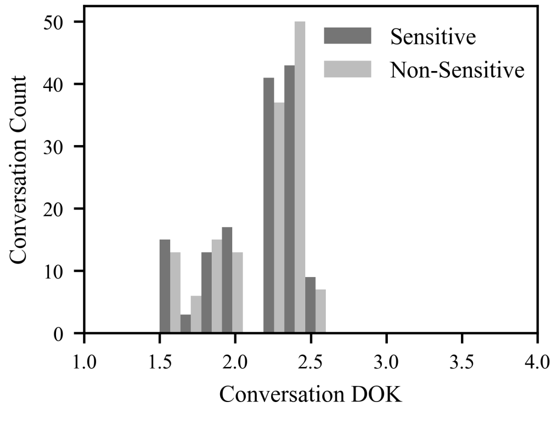


Figure 4: Conversation DOK (n=282) distribution by Topic Sensitivity

of 282 conversation plans, each consisting of questions generated by an LLM-based Socratic chatbot (SAM) in response to synthetic student Annotated Bibliography work samples, revealed comparable cognitive complexity patterns regardless of topic sensitivity. These results suggest that, under the tested configurations, SAM's prompt engineering successfully maintains equitable question generation across both sensitive and non-sensitive topics. This finding is encouraging for developers and educators implementing LLM-based formative feedback, especially in courses where students explore controversial or sensitive topics. At the same time, our results also underscore the importance of systematic psychometric evaluation in verifying LLM performance on student-facing tasks such as question generation, especially prior to classroom deployment.

4.1 Limitations

There are several limitations that affect the generalizability of our findings. First, our analysis relied on synthetic student work samples generated by LLMs rather than authentic student work. While this approach enabled controlled comparison across topic sensitivity categories, synthetic samples may not accurately reflect the style, structure, and linguistic diversity present in real student work. However, this limitation may be less pronounced for Annotated Bibliography work samples, as this assignment follows a relatively standardized format that constrain the range of possible outputs. Second, this study relies heavily on the performance of a single LLM (Claude 3.5 Sonnet) for both synthetic work generation and question generation. As the reasoning capabilities of LLMs continue to improve, it is important to re-evaluate these findings with newer models and configurations. Third, our DOK comparison focused on one stage of the AP Seminar research process (Annotated Bibliography) and examined only "starter" questions. In real classroom implementations, SAM can ask up to three follow-up questions per starter question, designed to draw more information from students based on rubric criteria and personalized to specific student responses. Due to their dynamic, response-dependent nature, follow-up questions are challenging to anticipate without authentic student interactions. However, since follow-up questions are expected to comprise more than 50% of actual conversation content, future conversational DOK measures must incorporate them to provide complete evaluation.

Future research should address these limitations by examining authentic student work interactions with the chatbot across multiple phases of the research process and incorporating follow-up questions analysis. Additionally, investigations should extend beyond topic sensitivity to examine other student-related variables that may influence the rigor and equity of support provided by the LLM-based chatbot, such as language proficiency, socioeconomic status, and cultural background.

Acknowledgments

The authors thank Julie Miles and Garrett Ziegler for their content expertise and contributions to question coding, without which this work would not have been possible. The authors also thank the greater College Board GenAI Studio team for their collaboration on the development of SAM.

References

Erfan Al-Hossami, Razvan Bunescu, Justin Smith, and Ryan Teehan. 2024. Can language models employ the socratic method? experiments with code debugging. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 53–59.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. 2022. *arXiv preprint arXiv:2212.08073*, 8(3).

K. W. Chan, F. Ali, J. Park, K. S. B. Sham, E. Y. T. Tan, F. W. C. Chong, L. Y. Ng, J. H. Lim, P. B. Ooi, L. K. Mun, B. K. S. Gan, T. S. Hew, N. W. Er, and G. K. Sze. 2025. Automatic item generation in various stem subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 100344(8).

S. Christopherson and N. Webb. 2024. Complexity and difficulty in a coherent standards-based education system. `https://www.webbalign.org/difficulty-and-complexity`.

Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a socratic chatbot. In *International Workshop on AI in Education and Educational Research*, pages 17–32. Springer.

Jennifer Garcia Ramos and Zakiya Wilson-Kennedy. 2024. Promoting equity and addressing concerns in teaching and learning with artificial intelligence. In *Frontiers in Education*, volume 9, page 1487882. Frontiers Media SA.

Michael Gerlich. 2025. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.

Jui-Tse Hung, Christopher Cui, Diana M Popescu, Saurabh Chatterjee, and Thad Starner. 2024. Socratic mind: Scalable oral assessment powered by ai. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 340–345.

Leonora Kaldaras, Hope O Akaeze, and Mark D Reckase. 2024. Developing valid assessments in the era of generative artificial intelligence. In *Frontiers in education*, volume 9, page 1399377. Frontiers Media SA.

S. Maity, A. Deroy, and S. Sarkar. 2025. Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, 100370(8).

Agariadne Dwinggo Samala, Soha Rawas, Tianchong Wang, Janet Marie Reed, Jinhee Kim, Natalie-Jane Howard, and Myriam Ertz. 2025. Unveiling the landscape of generative artificial intelligence in education: a comprehensive taxonomy of applications, challenges, and future prospects. *Education and Information Technologies*, 30(3):3239–3278.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *International Conference on Artificial Intelligence in Education*, pages 165–179. Springer.

C. Shu, N. Yao, Y. Chen, V. Wijeratne, L. Ma, J. Loo, S. Khan, M. Tariq, B. G. Zapirain, and A. Abuelmaatti. 2025. Ai-assisted multiple-choice questions generation with multimodal large language models in engineering higher education. In *2025 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE.

A. Tran, K. Angelikas, E. Rama, C. Okechukwu, D. H. Smith, and S. MacNeil. 2023. Generating multiple choice questions for computing courses using large language models. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Office of Educational Technology U.S. Department of Education. 2023. Artificial intelligence and the future of teaching and learning: Insights and recommendations. Technical report, U.S. Department of Education, Washington, DC. “Examples Are Not Endorsements”, public domain; Prepared by Miguel A. Cardona; Roberto J. Rodríguez; Kristina Ishmael.

Sean J Westwood, Justin Grimmer, and Andrew B Hall. 2025. Measuring perceived slant in large language models through user evaluations. Technical Report 4262, Stanford Graduate School of Business Faculty Working Papers.

Ying Zheng, Xueyi Li, Yaying Huang, Qianru Liang, Teng Guo, Mingliang Hou, Boyu Gao, Mi Tian, Zitao Liu, and Weiqi Luo. 2024. Automatic lesson plan generation via large language models with self-critique prompting. In *International Conference on Artificial Intelligence in Education*, pages 163–178. Springer.

A Sample Synthetic Work Input and SAM Outputs

Presented below is a sample synthetic Annotated Bibliography (First Draft) for a sensitive topic, accompanied by SAM’s generated conversation plan.

A.1 Sample Sensitive Annotated Bibliography

Annotated Bibliography

1. Harrington, J. (2019). The Moral Case for Capital Punishment. Journal of Criminal Justice Ethics, 18(2), 45-67.

Annotation: This article presents a philosophical argument in favor of the death penalty, contending that it is a morally justified form of punishment for the most heinous crimes. The author examines various ethical frameworks, including retributivism and utilitarianism, to build a comprehensive case for capital punishment.

2. Wilkins, E. (2019). Executing the Innocent: A Systemic Failure of the Death Penalty. Harvard Law Review, 132(4), 789-832.

Annotation: This comprehensive review of the death penalty system in the United States highlights the issue of wrongful convictions and executions. The author analyzes various factors, such as eyewitness misidentification, false confessions, and ineffective legal representation, that contribute to the execution of innocent individuals.

3. Chambers, R. (2020). The Deterrent Effect of the Death Penalty: A Review of the Evidence. Criminology and Public Policy, 15(1), 171-204.

Annotation: This article examines the empirical evidence on the deterrent effect of the death penalty. The author critically evaluates the existing research and concludes that the deterrent effect of capital punishment is inconclusive, challenging the common argument that the death penalty deters crime.

4. Linden, S. (2018). Cruel and Unusual: The Psychological Toll of Executions on Prison Staff. Journal of Correctional Health Care, 22(3), 256-271.

Annotation: This study explores the psychological impact of carrying out executions on prison staff. The author presents findings from interviews and surveys, highlighting the emotional and mental health challenges faced by those involved in the execution process, and the need for better support and counseling for these individuals.

5. Malik, A. (2021). The Death Penalty and Human Rights: A Global Perspective. International Journal of Human Rights, 27(2), 345-378.

Annotation: This article provides a global analysis of the death penalty from a human rights perspective. The author examines the international legal framework, the practices of various countries, and the ongoing debates surrounding the compatibility of capital punishment with fundamental human rights.

A.2 SAM Output

<conversation_plan>

1. Summary question: What is your research topic or question regarding capital punishment?

2. Decision question: How did you go about selecting these specific sources for your annotated bibliography?

3. Evidence question: Which of these sources do you find most compelling or influential in shaping your understanding of the death penalty, and why?

4. Speculative question: Based on your research so far, can you think of any potential alternatives to the death penalty that might address the concerns raised by both supporters and opponents?

5. Reflection question: How has your understanding of the death penalty evolved as you’ve researched and created this annotated bibliography?

</conversation_plan>

B Additional LLM Inputs and Outputs

Presented below are additional tables referenced in the text, including sample few-shot example questions and the most frequent SAM questions.

Question Type	Question Type Description	Work Phase	Few-Shot Example
Summary	Asks students to summarize their research topic	Annotated Bibliography	*What is your research topic?*
Decision	Asks students to explain their decision rationale for choosing a specific topic or including specific arguments or evidence	Annotated Bibliography	*How did you go about choosing your topic and narrowing down your focus?*
		Outline	*Describe for me the perspective in X source. How are you thinking of using that in section Y?*
Evidence	Asks students to recall sources used to supporting their conclusion/claims	Annotated Bibliography	*Which of the sources in your annotated bibliography were most influential in shaping your conclusion about [your research topic]?*
Speculative	Asks students to provide reasoning for observations or further explore related concepts	Annotated Bibliography	*Can you think of any evidence or data that would change your conclusion about [your research topic]?*
Reflection	Asks students to reflect on their research learning experience	Annotated Bibliography	*How did your understanding of [the topic] evolve as you researched and created your annotated bibliography?*
		Outline	*If you could improve one section, which would it be and what would you do?*

Table 5: Sample Few-Shot Example Questions in SAM's Prompt by Question Type

Rank	Question	Work Phase	Question Type	DOK	Count
1	*Are there any sections or elements of your outline that you would like to improve or expand on?*	Outline	Reflection	Unclassifiable	125
2	*Reflecting on your research and outline development process, what have you learned about conducting research?*	Outline	Reflection	4	100
3	*How did you decide which perspectives or arguments to include in your outline?*	Outline	Decision	3	91
4	*What is the main focus of your research based on these sources?*	Annotated Bibliography	Summary	1	86
5	*What is your current research topic, and how did it evolve as you developed the outline?*	Outline	Reflection	2	79

Table 6: Most Frequent SAM Questions

Keystroke Analysis in Digital Test Security: AI Approaches for Copy-Typing Detection and Cheating Ring Identification

Chenhao Niu[*], **Yong-Siang Shih**[*], **Manqian Liao, Ruidong Liu, Angel Ortmann Lee**

Duolingo, Inc.
`{chenhao,yongsiang,mancy,ruidong,angel.ortmannlee}@duolingo.com`

Abstract

In high-stakes remote exams like the Duolingo English Test, sophisticated cheating methods such as *copy-typing* from external sources and the operation of *organized cheating rings* pose significant security threats. We introduce a two-pronged, content-agnostic approach that analyzes behavioral data to address these issues. For **copy-typing detection**, we present a sequence model that combines 1D convolutional layers with a Transformer encoder, which captures both local and long-range dependencies in keystroke dynamics to distinguish between organic and transcribed typing, outperforming tabular and other sequential baselines. For **cheating ring identification**, we propose a deep metric learning framework that generates a unique behavioral embedding for each test session using both keystroke and mouse dynamics. The model outperforms a traditional statistical baseline in linking test sessions completed by the same individual. Together, these AI-driven methods provide a powerful and scalable toolkit for safeguarding the integrity of remote assessments.

1 Introduction

Duolingo English Test (DET) (Naismith et al., 2025) is a remotely administered, high-stakes English proficiency exam. Ensuring the integrity of online assessments like the DET presents persistent challenges, especially as new forms of digital test fraud emerge (Belzak et al., 2025a). Two key threats are *copy-typing* and *organized cheating rings*.

Copy-typing refers to the act of transcribing text from an external source, rather than composing it organically. In the context of high-stakes testing, this form of misconduct often involves test takers using pre-written answers, receiving live assistance, or employing large language models (LLMs) to generate responses. Because security measures like disabling copy-paste functionality are applied, these illicitly obtained answers have to be manually typed into the response box. This behavior creates subtle but detectable deviations from natural typing patterns. Our work introduces a content-agnostic model that analyzes these behavioral signals in keystroke dynamics to effectively detect copy-typing.

Organized cheating rings are groups or commercial services that assist multiple test takers in cheating. They pose a significant threat to the integrity of the test due to their large-scale operations. Although such services may use various methods, we focus on a prevalent cheating ring scenario in which a human helper effectively completes tasks for multiple test takers using remote-control software or external peripherals unobservable in webcam footage. Because the same helper tends to assist many test takers, an effective strategy is to *link* test sessions completed by the same helper. We propose to fingerprint behavior using keystrokes and mouse dynamics, then compare test sessions through learned embeddings. Identified clusters of linked test sessions can be surfaced for human review (Shih et al., 2024)[1]. To this end, we employ a deep metric learning framework that embeds each test session into a high-dimensional representation, enabling efficient retrieval of similar test sessions via approximate nearest neighbor search.

Contributions.

- A content-agnostic keystroke sequence model for **copy-typing detection** that combines 1D convolutional networks with a Transformer encoder to effectively capture both local and long-range dependencies in keystroke dynamics.

---

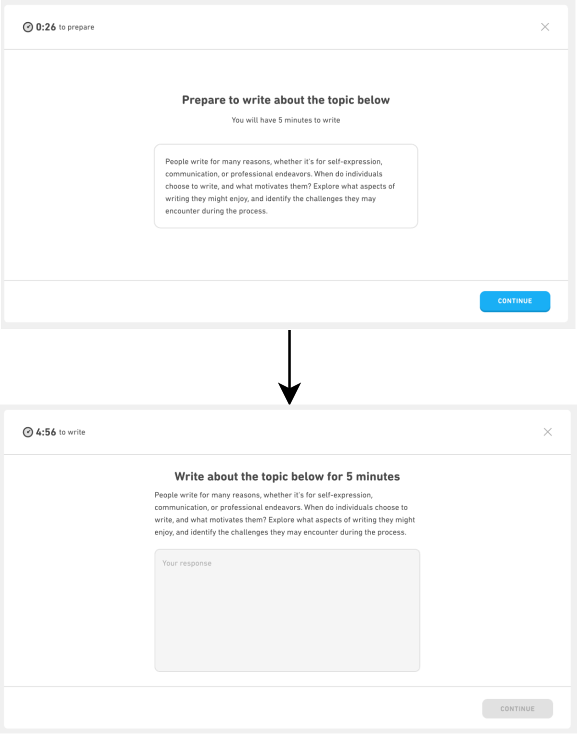
[*] Equal contribution.

[1] An initial high-level description of the cheating ring identification system appeared in Shih et al. (2024). Here we contribute full technical details and new experiments.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 107–116
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

- A deep metric learning framework for **cheating ring identification**, including a modified n-pair training objective with hard negatives controlling for hardware/region confounds.

2 Data Source and Context



Prepare to write about the topic below

You will have 5 minutes to write

People write for many reasons, whether it's for self-expression, communication, or professional endeavors. When do individuals choose to write, and what motivates them? Explore what aspects of writing they might enjoy, and identify the challenges they may encounter during the process.

Write about the topic below for 5 minutes

People write for many reasons, whether it's for self-expression, communication, or professional endeavors. When do individuals choose to write, and what motivates them? Explore what aspects of writing they might enjoy, and identify the challenges they may encounter during the process.

Your response

Figure 1: An example of the writing item in DET. The test taker has 30 seconds to prepare for the topic, and 5 minutes to write about it by typing on their keyboard.

Context. Our data originates from the Duolingo English Test (DET), a high-stakes, remotely proctored assessment of English proficiency (Naismith et al., 2025). The test's security protocol involves multiple layers, such as video recording, screen sharing, and input monitoring. Upon completion, every test session undergoes a rigorous review process that combines AI-driven analysis with human proctor oversight to identify any integrity violations (Belzak et al., 2025a). In our copy-typing detection research, we focus on an open-ended writing task in the DET, where test takers have 30 seconds to read a question given by text, and 5 minutes to type their response on a computer (See Figure 1 for an example). Comparatively, in our cheating ring identification research, we utilize the mouse movement patterns from the entire test session in addition to the keystroke patterns from writing tasks to identify the individual who completed the test.

Raw data collection. All DET test sessions record keystroke and mouse activity, including the timestamp of each key press and release, mouse movements and clicks, cursor position, and other contextual information (See Table 1 and Table 2 for details). We extract features from the raw log files with application-specific feature engineering methods.

Labels. To train and evaluate the models, the collected data is labeled based on human and AI-assisted proctoring decisions. The specific labeling criteria for both experiments are detailed in §4.

3 Methodology

Our methodology leverages machine learning to analyze behavioral data for two security tasks, based on the same data source of raw keystroke data.

3.1 Copy-Typing Detection

In this section, we introduce our method for copy-typing detection. At a high level, we treat the problem as a binary classification problem, and we train a machine learning model with labeled data, using a combination of tabular and sequential features.

3.1.1 Feature Engineering

For feature engineering, we process the raw features (§2) of each keystroke event to extract both sequential features and aggregated tabular features for each sample.

Key code categorization. To ensure the model is content-agnostic, we replace the exact `key-code` with a categorical `action-type`, which takes values among { INPUT, DELETE, MOVE, OTHER }, providing a coarser categorization. In addition, we include a binary `is-punctuation` indicator for punctuation keys.

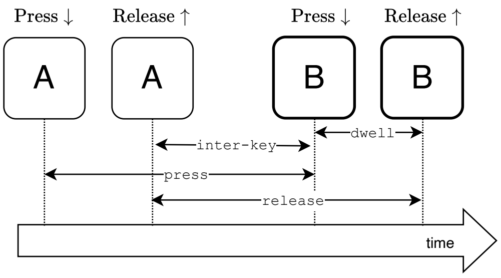


Figure 2: An example of extracting temporal features for the keystroke of the key "B". We consider four types of durations: `dwell`, `inter-key`, `press`, and `release`.

Raw Feature	Description
`key-code`	The exact key being pressed, such as “a”, “shift”, “delete”, etc.
`timestamp`	The timestamp of the key being pressed, in milliseconds.
`dwell-duration`	The duration the key remains pressed, in milliseconds.
`text-length`	The number of characters in the text box at the time of the key being pressed.
`caret-position`	The current position of the caret in the text box, counted by the number of characters from the beginning of the text box.

Table 1: Raw keystroke features and their descriptions.

Raw Feature	Description
`position`	The current mouse cursor position on the screen.
`timestamp`	The timestamp when the data is recorded, in milliseconds.
`is-clicked`	Whether it’s a click event or just a record of the current mouse position.
`dwell-duration`	The duration the mouse button remains pressed, in milliseconds.

Table 2: Raw mouse movement features and their descriptions.

Temporal sequence features. Similar to existing research (Acien et al., 2021; Stragapede et al., 2024), we extract timestamp-related features from `timestamp` and `dwell-duration`, by calculating the following four types of durations for each keystroke. See Table 3 for details, and Figure 2 for a concrete example.

Text length and caret position processing. For `text-length` and `caret-position`, we normalize them by the maximum text length observed within the item, to convert the feature values to the range of $[0, 1]$.

Auxiliary tabular features. In addition to temporal sequence features, we compute several tabular features for each sample. These features enhance the sequential model and enable comparison against tabular baselines. Specifically, the tabular features include: (i) counts by `action-type`, (ii) summary statistics { mean, std, p1, p25, p50, p75, p99, skewness } for each temporal feature, and (iii) pause-related measures: the count, total paused time, and average pause duration for pauses longer than thresholds { 200ms, 500ms, 1,000ms, 2,000ms, 3,000ms }.

All aggregates are normalized using training-set statistics. When combined with the sequence encoder, we include only a minimal, non-duplicative subset.

3.1.2 Model Architecture

Inspired by existing research (Acien et al., 2021; Stragapede et al., 2024), we model copy-typing detection as binary sequence classification over keystroke tokens. Each item is represented as a sequence $\{\mathbf{x}_i\}_{i=1}^{L}$, where $\mathbf{x}_i \in \mathbb{R}^{D_{\text{seq}}+D_{\text{tab}}}$ is the per-keystroke feature vector (§3.1.1), including D_{seq} sequential features concatenated with D_{tab} tabular features. Figure 3 is a diagram of the model architecture.

Overview. The architecture consists of: (i) one-dimensional convolutional networks (1D CNN) (Krizhevsky et al., 2012; Lea et al., 2016) that down-samples the sequence; (ii) a Transformer encoder (Vaswani et al., 2017) that models long-range dependencies; and (iii) a classifier head that produces an item-level logit. The design prioritizes robustness and efficiency while preserving discriminative temporal patterns characteristic of copy-typing.

Down-sampling with 1D CNN. An element-wise multi-layer perceptron (MLP) projects each input embedding from the width $D_{seq} + D_{tab}$ to a hidden width D_h. We then apply 1D CNN blocks with kernel size 3 and stride-2 max pooling to progressively halve the temporal resolution. With N_{CNN}=3, this yields a sequence of length $L/2^3$ with width $D_h = 32$. Intuitively, the CNN captures subword-scale rhythms while reducing the attention burden downstream.

Temporal Feature	**Description**
`dwell`	the duration between the press and release of the current key.
`inter-key`	the duration between the release of the previous key and the press of the current key.
`press`	the duration between the press of the previous key and the press of the current key.
`release`	the duration between the release of the previous key and the release of the current key.

Table 3: Definition of temporal sequence features that are extracted from raw keystroke timestamps.

Transformer encoder. We employ a Transformer encoder with $N_{\text{Transformer}} = 4$ layers and multi-head attention. Following a typical approach for Transformer-based classification (Devlin et al., 2019), we add positional embeddings and a trainable `[CLS]` token embedding at the input of the Transformer encoder, where the `[CLS]` token is a special token used for the model to aggregate the information for classification. The `[CLS]` embedding of the last layer is the output of the Transformer encoder.

Classification. At the final classification layer, we concatenate the output of the `[CLS]` token embedding from the Transformer encoder with a linear-transformed vector of the tabular features, and use a multi-layer perceptron (MLP) with sigmoid output to get the final classification output.

Training. Following a common approach in binary classification, we use Binary Cross Entropy (BCE) loss as the training objective with the AdamW optimizer (Loshchilov and Hutter, 2017). For a batch with N samples, the BCE loss is defined as

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

Where $y_i \in \{0, 1\}$ is the ground truth label for the i-th sample, and $\hat{y}_i \in (0, 1)$ is the prediction for that sample.

3.2 Cheating Ring Identification

In this section, we describe our cheating ring identification method. At a high level, the problem can be framed as a binary classification task: determining whether a pair of test sessions was completed by the same individual or not. Rather than training a direct classifier, we approach this with a deep metric learning framework (Kaya and Bilge, 2019). The approach makes deployment more practical, as the learned representations can be leveraged to efficiently retrieve similar test sessions using approximate nearest neighbor search (Li et al., 2019).

3.2.1 Feature Engineering

Keystroke features. Following existing research (Young et al., 2019), we compute summary statistics – mean, standard deviation, and sample count – for (i) the `dwell` duration of each key and (ii) the `transition` duration[2] between key pairs within each test session. These summary statistics constitute the keystroke features.

Mouse features. For mouse patterns, we compute several mouse movement metrics adapted from Zheng et al. (2011). The histograms of these metrics are used as mouse features:

- For any three consecutive recorded mouse cursor positions, A, B, and C, where the test taker clicked the mouse at C, we compute the following metrics:
  1. Direction: the angle between the horizontal line and $\overrightarrow{AB}$.
  2. Angle of Curvature: the angle between $\overrightarrow{BA}$ and $\overrightarrow{BC}$.
  3. Curvature Distance Ratio: the ratio between (1) the distance from B to $\overrightarrow{AC}$ and (2) the distance from A to C.
- For any two mouse clicks, we compute the time duration between the two clicks.

3.2.2 Model Architecture

Overview. A visual overview of our metric learning model is shown in Figure 4. We use an MLP to encode the input features to an embedding that represents the test session. The embeddings from two test sessions are used to compute a dissimilarity score, which is used in both training and inference.

Note that the same procedure is used to train two MLPs based on keystroke features and mouse fea-

---

[2] In our implementation, the `inter-key` interval in Figure 2 is used as the `transition` duration if the previous key is released before the next key is pressed; otherwise, the `press` interval is used.

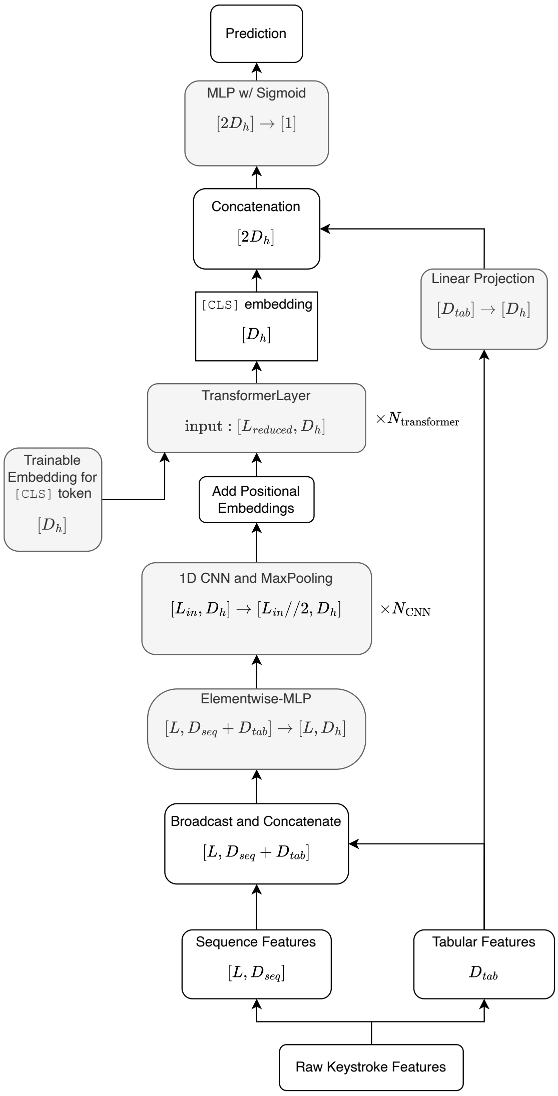


Figure 3: Copy-typing detection model architecture.

tures, and in the following discussions, we will not distinguish whether the input features are keystroke features or mouse features. During inference, we aggregate the dissimilarity scores by computing a weighted sum of the two. The weights are selected by fitting a Logistic Regression model (Cox, 1958) on the training dataset using the dissimilarity scores as inputs to predict whether a pair of test sessions is positive or negative.

Training. We use a modified multi-class n-pair loss (Sohn, 2016) to train the MLP encoder, replacing cosine with L_2 distance $d(\cdot)$ and augmenting batches with **hard negatives** matched on device type or region to reduce confounding. Intuitively, positives are pulled together while negatives are pushed apart.

Let $\{(x_i, x_i^+, x_i^-)\}_{i=1}^N$ be N triplets of input features sampled from the dataset, where each (x_i, x_i^+) is a positive pair corresponding to test sessions completed by the same individual (e.g., the same helper), and each (x_i, x_i^-) is a hard negative pair corresponding to test sessions completed by different individuals but involving the same type of keyboard or mouse, or occurring in the same region. We formulate a modified n-pair loss as shown in Equation 1. [3]

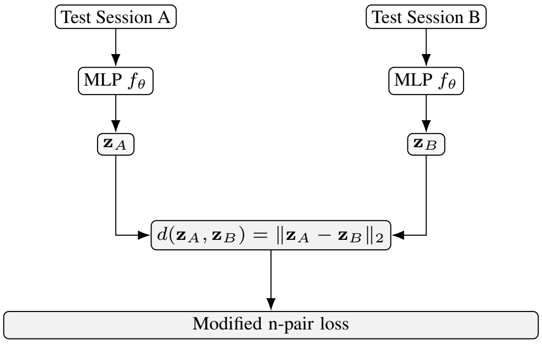


Figure 4: Each test session is encoded by the same MLP to produce an embedding. The embeddings can be compared to compute the dissimilarities between pairs of test sessions. Here $\mathbf{z}_A$ and $\mathbf{z}_B$ represent the output embeddings.

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{\hat{x}\in S}\frac{\delta(x_i,\hat{x})}{c_i}\log\frac{\exp(-d_\theta(x_i,\hat{x}))}{\sum\limits_{x'\in S}\exp(-d_\theta(x_i,x'))}, \tag{1}$$

$$\begin{aligned}
\text{where } c_i &= \sum_{\hat{x}\in S}\delta(x_i,\hat{x}),\\
d_\theta(x_i,\hat{x}) &= d(f_\theta(x_i), f_\theta(\hat{x})),\\
d(\mathbf{z}_A,\mathbf{z}_B) &= \|\mathbf{z}_A-\mathbf{z}_B\|_2,\\
S &= \{x_i^+\}_{i=1}^N \cup \{x_i^-\}_{i=1}^N,\\
\delta(x_i,\hat{x}) &= \begin{cases} 1, & \text{if } (x_i,\hat{x}) \text{ is a positive pair,}\\ 0, & \text{otherwise.}\end{cases}
\end{aligned}$$

Deployment. In our framework, the model learns a distance function between test sessions through their high-dimensional representations. At deployment, these representations could be stored in a vector database, enabling efficient retrieval of similar test sessions using any test session as a query. By surfacing hidden connections between a test

---

[3] In our setup, there is a small probability that multiple positive pairs exist in Equation 1. The normalization term c_i is introduced to ensure a valid probability distribution for the cross-entropy loss. This is similar to the implementation of `tfa.losses.npairs_loss` in TensorFlow Addons.

session under review and its most similar test sessions, the system helps human proctors more effectively identify organized cheating rings. To combine keystroke-based and mouse-pattern-based dissimilarity scores, we scale each embedding by the inverse square root of its corresponding weight. This normalization allows us to directly leverage approximate nearest neighbor search in standard vector databases.

4 Experiment

4.1 Copy-Typing Detection Experiment

Dataset. As described in §2, we collect samples from DET test sessions. For copy-typing detection specifically, we filter out samples with short responses (less than 100 characters) or with non-standard `key-code` values, and collect 12,000 positive samples and 126,000 negative samples from January 12, 2024 to August 1, 2024, and randomly split them by a ratio of 6:2:2 for training, validation, and testing. We use the training set to train the copy-typing detection model, the validation set for early stopping and hyperparameter selection, and the testing set for evaluation.

Labels. To support robust model development and evaluation, labels are derived from both human and AI-assisted proctoring decisions. Specifically, positive examples are test sessions where proctors confirmed the usage of external resources, such as large language model (LLM) generated responses, with both AI-based detection (Niu et al., 2024) and video-based evidence. Negative examples are drawn from clean certified test sessions, where proctors find no violations in the current test session, and the user has no previous violations.

Settings. To validate the effectiveness of the proposed approach against other settings, in addition to the model architecture described in §3.1.2 (i.e., CNN + Transformer in Table 4), we select the following variations:

- **Tabular models**: To verify the effectiveness of adding sequential features, we adopt two baselines that use only the tabular features, including Logistic Regression (Cox, 1958) and LightGBM (Ke et al., 2017). They are noted as "Logistic Regression" and "LightGBM" in Table 4.
- **Other sequential models**: To verify the advantage of the Transformer-based encoder, we replace the Transformer blocks with other sequential models such as 1D convolutional networks (CNN) and Long Short-Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997). They are noted as "CNN only" and "CNN + LSTM" in Table 4.

Metrics. Given that the positive and negative samples are naturally imbalanced, we use the Area Under the ROC Curve (AUROC) as a metric. Besides, since the practical application typically requires a low False Positive Rate (FPR), we also report the True Positive Rate (TPR, a.k.a., recall) at a 1% FPR.

4.2 Cheating Ring Identification Experiment

Dataset. As detailed in §2, our dataset is built from DET test sessions. Specifically, we sample certified test sessions from Q1 2025, excluding any with insufficient keystroke data. The resulting dataset covers approximately 102,000 test takers. We partition these into training, validation, and test sets with a 6:2:2 split. For validation and testing, pairs of test sessions were sampled and the labels are defined below. For training, we sample triplets of test sessions and construct pairs following the methodology described in §3.2.2. We use the validation set for early stopping, while the test set is used for evaluation.

Labels. For this task, we aim to detect when two test sessions were taken by the same human helper. However, the ground-truth data on cheating rings is scarce. Therefore, we construct our dataset using certified test sessions rather than confirmed cheating ring test sessions. The key intuition is that test sessions completed by the same individual can serve as positive pairs, since their behavior patterns originate from the same person, and test sessions completed by different individuals can serve as negative pairs. To reduce potential confounding factors, we require that test sessions in negative pairs match on keyboard/mouse type or region.

Evaluation scenarios. We evaluated our models under two scenarios. In the first scenario, 3051 positive pairs and 3051 negative pairs were sampled for each split, and the models were evaluated based on binary classification. We report the AUROC and the true positive rate at a 1% FPR. In the second scenario, we compare one test session against K different test sessions from other test takers. The K comparisons are treated as a single false positive if any of the K pairs are predicted as positive by the model. Otherwise, they are treated as a single true negative. The test session is additionally compared with one other test session from the same test taker. The comparison is treated as true positive if the model predicts positive. Otherwise, it is treated as a false negative. 3051 test sessions were sampled from the test split to construct $3051(K+1)$ pairs as described above, and we report the AUROC of different models with different K.[4]

Baseline. We compare our proposed method with an in-house t-test based method that is built upon the work of Young et al. (2019), which utilizes the summary statistics of `dwell` durations and `transition` durations to determine if the two tests are completed by the same individual.

5 Results

5.1 Copy-Typing Detection

The evaluation results for our copy-typing detection models are summarized in Table 4. These results suggest two critical points. First, the superior performance of sequential models over the tabular baselines (Logistic Regression and LightGBM) confirms that the temporal dynamics of keystrokes contain essential signals for detecting copy-typing. Second, the choice of sequence architecture is crucial. While a simple "CNN only" model offers little advantage over a strong tabular baseline, incorporating a sophisticated encoder like a Transformer or LSTM to model long-range dependencies unlocks substantial performance gains. This highlights the necessity of using powerful sequence models to fully leverage the predictive patterns in keystroke data.

5.2 Cheating Ring Identification

The results based on the first evaluation scenario for cheating ring identification are shown in Table 5. As shown in the table, the proposed method outperforms the baseline. We additionally report the performance of our method using different subsets of features. The superior performance of the deep-full model demonstrates that keystroke and mouse dynamics provide complementary signals, and that combining them creates a more robust and accurate behavioral fingerprint.

The results based on the second evaluation scenario with K comparisons are shown in Figure 5. In the plot, we can see that our method remains competitive against the baseline as K increases, demonstrating the robustness of our methods.

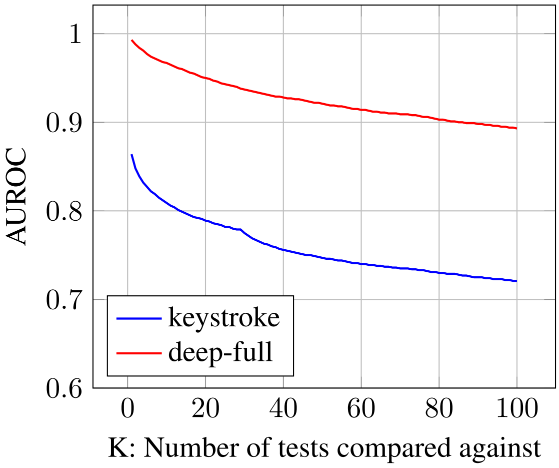


Figure 5: Performance comparison between our deep-full method and the t-test based keystroke baseline.

5.3 Fairness Analysis

Fairness is an important aspect of a Responsible AI system (Burstein et al., 2025). With the notion of *equality of opportunity* (Hardt et al., 2016), we evaluate the True Negative Rate (TNR) across demographic groups for both AI models. Intuitively, a similar TNR across groups means that innocent test takers in each group have a similar possibility of not being falsely flagged by the AI models. For our study, we focus on major geographical sub-regions according to the United Nations geoscheme[5], and evaluate the TNR of the models on clean[6] certified test sessions (for copy-typing detection) and negative pairs (for cheating ring identification) across groups. Table 6 and Table 7 present the proportion of test takers and pairs from each major sub-regions in the dataset, and the evaluation results of group-wise TNR. Note that the demographic distributions are slightly different for the two experiments because the datasets are sampled from different time periods. For both experiments, the results verify that the TNRs are within a small difference across groups, ensuring the equality of honest test takers.

---

[4] The second scenario simulates the real-world use case: a test session is compared with multiple test sessions to detect links to the same human helper, and even a single predicted match among them may require proctors' review.

[5] https://unstats.un.org/unsd/methodology/m49/

[6] Filtered to reduce data contamination.

Features	Model	AUROC	TPR@1%FPR
Tabular	Logistic Regression	84.91%	11.12%
Tabular	LightGBM	86.56%	12.64%
Sequential	CNN only	84.96%	15.18%
Sequential	CNN + LSTM	94.76%	39.05%
Sequential	CNN + Transformer	**95.22%**	**41.41%**

Table 4: Copy-typing detection model performance. The best results are in bold.

Method	AUROC	TPR@1%FPR
keystroke	86.44%	69.42%
deep-keystroke	98.56%	72.93%
deep-mouse	93.63%	39.86%
deep-full	**99.28%**	**89.28%**

Table 5: Performance for cheating ring identification methods. **keystroke** is the t-test based method, and **deep-*** are our proposed methods with different input features, where **deep-full** uses both keystrokes and mouse features.

6 Conclusion

We presented a two-pronged, content-agnostic framework for enhancing test security using behavioral data. For **copy-typing detection**, our CNN-Transformer model effectively learns sequential patterns in keystroke dynamics, outperforming both tabular and simpler sequential baselines in identifying copy-typing behavior. For **cheating ring identification**, our deep metric learning system produces robust embeddings from keystroke and mouse features, outperforming a statistical baseline and enabling efficient, large-scale deployment via approximate nearest neighbor search. Together, these AI-driven approaches provide a powerful, scalable toolkit for safeguarding the integrity of remote assessments. However, the deployment of such powerful tools necessitates a commitment to Responsible AI standards (Burstein et al., 2025). To that end, these models are best implemented not as automated judges, but as essential components in a human-in-the-loop proctoring framework that surfaces evidence for human review. This approach ensures that the AI models are governed with human oversight, providing both a secure and accountable system for assessment.

Limitations

For **copy-typing detection**, although the proposed model achieves a meaningful TPR at a low FPR of 1%, there still exist practical challenges when adopting such AI-generated signals in test proctoring process. For instance, since the output from the deep learning model is not interpretable, the proctoring guidelines need to be carefully designed for human proctors to accurately confirm the copy-typing detection result. Belzak et al. (2025b) discuss the findings in the practical usage of copy-typing detection in detail. Besides, although our approach is content-agnostic, prior work (Liang et al., 2023) has shown that some AI-text detectors can disproportionately flag non-native English writing as AI-generated, raising fairness concerns in educational settings. We therefore audit our models for group-wise error differences (§5.3) and maintain human-in-the-loop confirmation before adverse actions.

For **cheating ring identification**, one limitation is that our proxy positives (same-person test sessions) are not equivalent to true cheating ring labels, and may not fully capture the operational complexity of organized cheating rings. For example, the helper may deliberately vary typing/mouse behaviors to evade detection. In addition, the method is designed to be used as a retrieval aid rather than a fully automated proctor; its outputs must be combined with other safeguards – such as additional behavioral rules, corroborating signals, or human proctor judgment – to avoid unfair penalization of innocent test takers.

References

Alejandro Acien, Aythami Morales, John V Monaco, Ruben Vera-Rodriguez, and Julian Fierrez. 2021. Typenet: Deep learning keystroke biometrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):57–70.

William Belzak, Basim Baig, Ramsey Cardwell, Rose

Sub-region	**Proportion**	**TNR@1%FPR (Copy-Typing Detection)**
Southern Asia	25.46%	98.51%
Eastern Asia	12.99%	97.21%
Latin America and the Caribbean	12.89%	99.68%
Northern America	9.58%	99.77%
Western Asia	7.10%	99.50%
Sub-Saharan Africa	6.75%	99.51%
South-eastern Asia	6.29%	99.69%

Table 6: Proportion of test takers from major sub-regions (> 5%), and the True Negative Rate (TNR) in each group.

Sub-region	**Proportion**	**TNR@1%FPR (Cheating Ring Identification)**
Eastern Asia	35.66%	99.22%
Southern Asia	15.63%	99.05%
Northern America	13.42%	99.40%
Western Asia	12.24%	99.64%
Latin America and the Caribbean	11.34%	98.94%
South-eastern Asia	6.18%	99.06%
Western Europe	6.01%	99.51%

Table 7: Proportion of pairs from major sub-regions (> 5%), and the True Negative Rate (TNR) in each group. Note that since each pair involves two test sessions, which may come from different groups, a single pair can be counted more than once. As a result, the sum of the proportions may exceed 100%.

Hastings, André Kenji Horie, Geoff LaFlair, Manqian Liao, Chenhao Niu, and Yong-Siang Shih. 2025a. Duolingo English Test: Security and score integrity. Duolingo research report, Duolingo.

William Belzak, Chenhao Niu, and Angel Ortmann Lee. 2025b. When machines mislead: Human review of erroneous AI cheating signals. *Artificial intelligence in measurement and education conference*.

Jill Burstein, Geoffrey T. LaFlair, Kathleen Yancey, Alina A. von Davier, and Rotem Dotan. 2025. Responsible ai for test equity and quality: The duolingo english test as a case study. In Earl M. Tucker, Eleanor Armour-Thomas, and Edmund W. Gordon, editors, *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning*. University of Massachusetts Amherst Library Press.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

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.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems 30*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer.

Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. 2019. Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Masha Kostromitina. 2025. Duolingo English Test: Technical manual. Duolingo research report, Duolingo.

Chenhao Niu, Kevin P. Yancey, Ruidong Liu, Mirza Basim Baig, André Kenji Horie, and James Sharpnack. 2024. Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 940–953, Miami, Florida, US. Association for Computational Linguistics.

Yong-Siang Shih, Manqian Liao, Ruidong Liu, and Mirza Basim Baig. 2024. Human-in-the-loop AI for cheating ring detection. *arXiv preprint arXiv:2403.14711*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Giuseppe Stragapede, Paula Delgado-Santos, Ruben Tolosana, Ruben Vera-Rodriguez, Richard Guest, and Aythami Morales. 2024. Typeformer: Transformers for mobile keystroke biometrics. *Neural Computing and Applications*, 36(29):18531–18545.

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.

Jay R Young, Randall S Davies, Jeffrey L Jenkins, and Isaac Pfleger. 2019. Keystroke dynamics: establishing keyprints to verify users in online courses. *Computers in the Schools*, 36(1):48–68.

Nan Zheng, Aaron Paloski, and Haining Wang. 2011. An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 139–150.

Talking to Learn: A SoTL Study of Generative AI-Facilitated Feynman Reviews

Madeline Mattox[1] **Natalie Hutchins**[1] **Jamie Jirout**[1]
[1]University of Virginia
[1]Jirout@virginia.edu

Abstract

Structured Generative AI interactions have potential for scaffolding learning. This Scholarship of Teaching and Learning study analyzes 16 undergraduate students' Feynman-style AI interactions (N=154) across a semester. Qualitative coding of the interactions shows mostly low-level student responses, but some evidence that prompt structure may can promote higher-level cognitive engagement. Results show GAI provides metacognitive support, and suggest the potential of GAI-supported Feynman reviews to provide interactive, personalized learning experiences that align with theories of cognitive engagement and metacognitive support for learning.

1 Introduction

Generative Artificial Intelligence (GAI) technologies, such as ChatGPT, are becoming increasingly prevalent, including in higher education (HE) (Zhu et al., 2025; Kim et al., 2025; Sun & Zhou et al., 2024). GAI has the potential to support learning through on-demand explanation, feedback, and clarification, with research showing support for deeper understanding and more efficiency when used effectively (Zhu et al., 2025; Chan & Hu, 2023; Dong et al., 2025). Students report finding GAI interesting to use and that it makes learning more enjoyable, but both students and faculty have concerns around whether GAI may inhibit some types of learning and negatively impact social interactions during learning (Kim et al., 2025; Chan & Hu, 2023). This study explores the potential of GAI Feynman-Style review activities by assessing the quality of interactions and content experienced in an college course.

There is a need for research to assess GAI impacts on learning and to identify empirically supported practices and principles for its use in education. A broadly held perspective is that GAI is not going away and will become part of typical experiences (Batista et al, 2024). Although some researchers provide justifications for limiting or banning GAI use in HE (e.g., de Fine Licht, 2024), its use is quickly increasing across and reshaping the landscape of labor markets, and, as a result, changing the skillset and knowledge needed in near-future jobs (Resh et al., 2025). Thus, HE has a responsibility to teach GAI skills to prepare students for the future (UNESCO), and should find ways to implement GAI in some contexts, while teaching students and faculty how to best use it ethically and effectively (Zhu et al., 2025; Yang et al; Lee et al). This Scholarship of Teaching and Learning (SoTL) study describes one attempt to use GAI: GAI-facilitated Feynman-style reviews.

The Feynman method involves a student explaining a topic as if to a novice, and then responding to subsequent probing questions to clarify, elaborate on, or deepen the student's thinking and understanding (Reyes et al., 2021). This approach aligns with research on the benefits of self-explanation and teaching for learning, which show that generating simplified explanations promotes deeper processing and transfer of knowledge (e.g., Chi et al., 1994; Fiorella & Mayer, 2013). Most effective are interactive approaches that include self-explanation and teaching along with a knowledgeable partner, but these are resource-intensive; GAI offers potential to simulate the interactive Feynman technique in ways that are scalable and effective (Rajesh & Khan, 2024). For introductory subjects, GAI likely has sufficiently accurate models to assess and explain content, and the capacity to personalize questions and feedback based on students' demonstrated knowledge in real-time, similar to what an expert human teacher or tutor would do in this method. We tested this hypothesis by analyzing interactions between

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 117–124
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

students and GAI across a semester in an introductory infant and child development course.

To understand the potential learning experiences of these activities, this work was grounded in the ICAP framework (Chi & Wylie, 2014), which differentiates levels of cognitive engagement, from most shallow to deepest: passive, active, constructive, and interactive. GAI-guided Feynman interactions can be active or constructive, depending on whether students are simply recalling facts vs. explaining, constructing summaries, or creating examples. If students use GAI to co-construct ideas the engagement is considered interactive, in which GAI contributions shape and extend the student's thinking, resulting in new understanding that wouldn't emerge from the student alone. While true interactive engagement would involve reciprocal co-construction (Chi & Wylie, 2014), which is not possible in that the GAI does not experience conceptual change, it might simulate an interactive experience, and the student may have the benefit of that level of cognitive engagement. This type of interaction can also support metacognition by making gaps in understanding visible (Flavell, 1979; Schraw & Moshman, 1995), leading students to further develop their understanding of a topic. In this study, we analyze how GAI can scaffold cognitive engagement and metacognition through its questions and feedback as aligned with ICAP framework and theories of metacognitive support. Together, these frameworks provide the foundation for interpreting the quality of student–GAI interactions during Feynman-style reviews.

Aims

Prior research shows the Feynman approach is effective, but it is unknown whether the effectiveness would be similar with interactions facilitated using GAI. This study addresses this question using data from a small (N=29) undergraduate child development course taught at a mid-sized, highly selective public research university located in the southeastern United States. Students in the course used a GAI of their choice to complete assignments including Feynman-style dialogues (Reyes et al., 2021). In these assignments, students were challenged to explain course concepts in simple terms and then were asked questions by the GAI to assess and refine their understanding while given feedback after each response. The current study explored the questions and feedback provided by the GAI, and the student responses elicited through the interactions[1] to understand the efficacy of this method for providing effective personalized learning experiences,. Our specific aims included:

1. Identify patterns of student engagement with GAI tools through these dialogues.
2. Explore how GAI can support metacognitive processes, such as explanation and clarification.
3. Assess the consistency and quality of these GAI responses.

2 Methods

2.1 Sample

Participants included students who opted in from the target course, of which 16 students consented to have their course materials included. Participants were 94% female, majoring in education (N=8; including Youth and Social Innovations, Speech and Communications Disorder, Kinesiology), arts and sciences (N=6; including undeclared, behavioral neuroscience, and psychology), and engineering (N=2; including undeclared and computer science), and were in their 1st (N=1), 2nd (N=5), 3rd (N=9), and 4th (N=1) year of college.

2.2 Study Design

This study used a qualitative research design to explore undergraduate student engagement with AI tools. Data were collected using a course management system, on which students completed bi-weekly assignments including Feynman-style interactions with a GAI of their choice. For each assignment, students were given a GAI prompt for a Feynman style review for each of two topics covered since their last review (see Figure 1 for an example prompt).

For each review question, a prompt was provided for students to copy into any GAI platform of their choice, which initiated a Feynman style review beginning with the student summarizing the topic and then asking six follow-

---

[1] The course was taught by the senior author. Data of consenting students were deidentified and analyzed only after the course ended, and the study was conducted with IRB approval.

up questions, providing feedback after each response. Students copied/pasted their full interaction into an assignment, which were then exported, de-identified, and compiled for coding and analysis at the utterance level. Three types of data were coded: GAI question prompts, GAI feedback, and student responses.

> Instructions: put the prompts below into your preferred GenAI platform to test your understanding of this topic. (Remember, using CoPilot behind [University Licensed program] allows you to not need to create a new account, but you may use any platform you prefer). Copy your full interaction with the GenAI platform as your responses in the two text boxes provided.
>
> “I want to test my understanding of children's language development using the Feynman method. Consider that I have read short articles that provided examples of children's language learning from experiences and learning processes like referential ambiguity, fast-mapping, and the vocabulary spurt. I've also had about two hours of time learning about these topics, including discussion of receptive and productive language. Use this information about readings and time learning to estimate the level of my knowledge. I'll explain the concept as if I'm teaching it to a beginner. Please ask me one question at a time with about 6 questions total, challenge unclear points, and identify areas where I need more depth or could simplify. After the dialogue, provide a summary of my understanding, highlighting strengths and areas for improvement.”

Figure 1: Instructions and example prompt text provided to students for the generative AI (GAI)–based Feynman-style review activity.

2.3 Coding

A subset of student-GAI interactions was read and discussed among the project team to identify codes that would be useful in answering the research questions, and that were possible to observe when reviewing the interactions. Coding categories were informed by the ICAP framework and models of metacognitive regulation to reflect higher and lower levels of cognitive engagement and examples of metacognitive support (Chi & Wylie, 2014; Flavell, 1979; Schraw & Moshman, 1995). This led to the decision to code GAI questions, student responses, and GAI feedback to responses as separate types of data, each with specific categories aligned with these frameworks and the research aims. The dataset was further divided into utterances to be coded, identified as the smallest segment of text conveying a single, complete idea or meaning. Utterances were all coded descriptively as GAI questions/prompts, student responses, or GAI feedback. Multiple codes could be assigned to an utterance, and if none were assigned there was an ‘other’ code. The following coding categories of each type were used:

GAI Question Codes

Asking For Examples. Prompt to illustrate a concept using a specific, real-life, or hypothetical example of a concept. Example: “Can you explain this concept through a real-life example?”

Asking For Comparisons. Prompt to analyze the relationships between two or more ideas, particularly their similarities or differences. Example: “Can you briefly compare their Piaget’s and Vygotsky’s views?”

Simple Question. Prompt for recall or evidence of comprehension, such as to define, describe, or explain a concept or multiple concepts without asking for comparison, contrast, application, revision, analysis, etc. Example: “What is Bronfenbrenner’s Ecological Model?”

Asking for further information: Prompt requests clarification. Example: “Can you explain what you mean by critical periods in this context?”

Elaboration of Ideas: Prompt eliciting further elaboration related to what a student responded. Example: “Now let’s push further: Can you think of an example of how sensory exploration leads to a deeper understanding of an object?”

Opportunity to revise a response: Prompt for student to change their response in a specific way. Example: “How might you reword that to better distinguish between the microsystem and the mesosystem?”

Explicit reference to response: While several prompts built on students’ responses, some explicitly referenced what students said before giving one of the above prompts. Example: “you mentioned the microsystem...” (followed by a question or prompt).

Prefacing question: Prompt provided information that was not in the form of a request or question, but more often giving background or information to contextualize the coming question.

Example: “You’ve read about how attachment might differ across racial backgrounds.”

Student Response Codes

Admitting uncertainty: Student explicitly acknowledges a lack of knowledge, confidence, or clarity about how to answer all or part of the AI’s question. Example: “I understand Piaget’s theory, but I don’t know how to apply it here.”

Giving examples: Student gives an example (spontaneously or prompted) to illustrate, apply, or demonstrate understanding of a concept, or to clarify or apply a concept. Example: "Sure, an example of a mesosystem could be the relationship between a child’s parents and their teacher."

Making comparisons: Student discusses talking about the differences or similarities between aspects of a concept (spontaneously or prompted); responses reflect relational understanding, such as highlighting how two concepts differ, explaining what two theories have in common, showing contrast in function, impact or structure. Example: “Unlike the mesosystem, the Microsystem involves direct interaction with the child”.

Giving a simple response: Student responds with a fact explanation requested, without including examples, comparisons, or other deeper-level information that goes beyond what was asked. Example: “The layers include the micro, meso, exo, macro, and chronosystem.”

GAI Feedback Codes

Examples: The GAI provides an example to help illustrate or clarify a concept. Example: "For instance, when a child takes deep breaths to calm down after being frustrated, that’s a form of self-regulation."

Correcting: GAI explicitly identifies part or all of what the student has said is incorrect and provides a corrected explanation. Example: "Not quite—the macrosystem isn’t a physical place. It refers to cultural values and norms.”

Affirming: GAI provides supportive or validating praise or encouragement, indicating the student is correct or on the right track. Example: “Nice work explaining that.”

Expanding: GAI builds on a student’s response by offering additional information, elaboration, or nuance, pushing the idea further. Example: "You’re right that secure attachment helps with emotional regulation. In fact, research shows it can influence stress response systems as well."

3 Results

To identify patterns of student engagement with GAI tools (Aim 1), and specifically for metacognitive support and related responses (Aim 2), we provide descriptive patterns of individual codes for each aspect of the student-GAI interaction, showing the frequency of occurrence of each code and the relative occurrence. To assess consistency and quality of these GAI responses (Aim 3) we use unconditional multilevel models to estimate intraclass correlation coefficients for each aspect of the GAI interaction (interactions nested with students).

The content of the 154 student-GAI full interactions was segmented into codable utterances, resulting in 2686 utterances coded within GAI questions (text written by the GAI). Of these, 2315 received one or more of our target codes (non-target codes included utterances not related to the content, such as, “I’m excited to hear what you have to say.”) In student responses, a total of 2994 utterances were coded, of which 2979 were coded with a target code (e.g., a student beginning their response with, “yes, I can do that.” before responding) and in feedback to responses, a total of 4934 utterances were coded, of which 4221 were coded with a target code (e.g., simply restating the student’s response).

At the student level, for each student-GAI interaction, students, on average, received 17.4 GAI question utterances (SD=3.23), responded in 19.8 utterances (SD=4.57), and received 32.2 GAI feedback utterances (SD=10.7). As a reminder, individual utterances could include multiple content codes, but only one descriptive code identifying the utterance as a question, response, or feedback (i.e., a question asked that built off a student’s response would not also be assigned feedback codes even if feedback was implied, such as asking for clarification).

3.1 Aim 1: Types of GAI Questions, Student Responses, and GAI Feedback

The utterances segmented for GAI questions, student responses, and GAI feedback were coded for one or more codes within each of the respective types. The frequencies of each are presented in Tables 1-3, respectively, including the number of observations receiving each code (Total Obs.), average observation per interaction (Mean Obs.), and proportion of all coded utterances (% of

GAIQ) are presented by utterance types (GAI questions - GAIQ, student responses - SR, GAI feedback - GAIF). Items with asterisks are those considered supportive of metacognition.

GAI Question Type	Total Obs.	Mean Obs.	% of GAIQ
Asking for examples*	311	2.0	13%
Asking for comparisons*	208	1.3	9%
Simple question	931	5.9	38%
Asking for further info. *	108	0.7	4%
Elaboration of ideas*	605	3.9	25%
Opportunity to revise*	24	0.2	1%
Explicit reference to response	196	1.2	8%
Prefacing question	61	0.4	2%

*indicates category supportive of metacognition

Table 2: Descriptive information for the types of GAI Questions observed.

Student Response Type	Total Obs.	Mean Obs.	% of GAIQ
Admitting uncertainty	12	0.1	0.4%
Examples (prompted) *	348	2.2	11%
Examples(unprompted) *	154	1.0	5%
Comparisons (prompted) *	111	0.7	4%
Comparisons (unprompted) *	25	0.2	1%
Simple correct response	2401	15.3	78%
Simple incorrect response	23	0.1	1%

*indicates category supportive of metacognition

Table 1: Descriptive information for the types of student responses observed.

GAI Feedback Type	Total Obs.	Mean Obs.	% of GAIQ
Examples*	299	1.9	7%
Correcting	347	2.2	8%
Affirming	2324	14.8	54%
Expanding*	1356	8.6	31%

*indicates category supportive of metacognition

Table 3: Descriptive information for the types of GAI Feedback observed.

3.2 Aim. 2: Metacognitive Support

Codes in each category identified as having potential to support metacognition are those that elicit reflection and deeper cognitive engagement in alignment with the ICAP theory and models of metacognitive reflection. We created composite scores for how frequently these high-support utterances appeared in the interactions. For the GAI questions, higher support codes included the prompts related to clarifying, elaborating or applying, revising, or comparing and contrasting, or giving examples. For the student responses, high support codes included giving examples and comparisons, and we looked separately at when these were given in response to a specific prompt or were unprompted. For feedback, high support included giving examples and expanding.

Utterances receiving codes identified as potentially highly supportive for metacognition included 945 GAI questions, 638 student responses, and 1655 feedback utterances. Proportionally, this was 39% of the GAI question codes, 13% of student responses, and 38% of GAI feedback. There were, of course, sources of variability across students and across the different topics reviewed in the assignments, so we explore the amount of variability both between students and across questions in these high-support composites as well as low composites to explore stability in the interaction quality.

3.3 Aim 3: Consistency and Quality

To explore interaction consistency, we compare the intraclass correlation coefficients (ICC) to understand whether the amount of variability at the level of students (i.e., individual differences explaining differences in the patterns of codes, P-ICC) compared to the level of variability at the topic level, nested within student (i.e., variability explained by differences in the prompts students pasted in for each topic).

At a basic level, we first estimated ICCs for the number of questions students received and the overall number of utterances observed in each of the three data types. Intraclass correlation coefficients (ICCs) indicated that a low proportion of variance in the number of GAI question utterances was explained by either the question/topic across semester or individual differences across students (person ICC = .099; question ICC= .027). In contrast, higher person ICCs for student responses and GAI feedback at the student level suggest some consistency in the number of utterances students made in their responses across activities (person ICC = .428; question ICC= .074), and moderate consistency in the number of feedback utterances they were given across activities in response to those utterances (person ICC = .214; question ICC= .065) ,.

To explore the *quality* of interactions, the composites described above for metacognitive support were assessed similarly to estimate person and question ICCs. We estimated ICCs (both low and high) when predicting the composite scores of GAIQ quality, SR quality (including low quality and prompted vs. unprompted high-quality responses), and GAIF quality (low and high). Results for the analyses are presented in Table 4.

Composite Score Type	P-ICC	Q-ICC
GAIQ – Low	.026	.295
GAIQ – High	.133	.037
SR – Low	.390	.072
SR – High, prompted	.037	.296
SR – High, unprompted	.120	.077
GAI Feedback – Low	.136	.050
GAI Feedback – High	.192	.104

Table 4: Intraclass correlation coefficients (ICCs) at the person level (P-ICC) and question level (Q-ICC) for composite coding categories.

ICCs indicated that reliability varied across coding categories, with some dimensions showing greater variance attributable to students (higher P-ICCs; e.g., SR–Low = .390) and others showing greater variance attributable to questions nested within students (higher Q-ICCs; e.g., GAIQ–Low = .295, SR–High, prompted = .296). These patterns suggest that some aspects of the coding (e.g., low-level, simple student responses) are more consistent across individual students, whereas others (e.g., types of GAI questions or student high-level responses when prompted) vary more at the topic/prompt level.

4 Discussion

Students' GAI interactions during a Feynman-style review show clear patterns of engagement between students and the GAI, in which students are providing many responses to provided questions. Students receive feedback both about their accuracy and understanding, and also feedback to expand their current knowledge. While there were clear individual differences explaining some variability in students responses (e.g., 39% of variability in simple responses provided), there was also indication that the prompts to initiate the activity also explain variability in responses (e.g., 30% of the variability of deeper-level student responses).

Our first two aims were to identify patterns of student engagement with GAI tools and explore how GAI can support metacognitive processes. We observed that the most common GAI Question type was simple questions (38%), but 39% of the GAI questions were considered higher-level, such as encouraging elaboration of ideas and prompting for examples or application of information. These align with ways of encouraging deeper cognitive engagement, which support greater learning (Chi & Wylie, 2014). Yet, students overwhelmingly gave simple correct responses (78%) with relatively few examples (16%) or comparisons (5%), regardless of prompting. One possibility is that the platform design of exchanging back-and-forth text with GAI might encourage short responses that can't convey as much depth (Torricelli et al., 2024).

The GAI feedback students received predominantly pushed them to expand their ideas (54%), which could help in revising knowledge (Chi & Wylie, 2014). Feedback affirming the student (34%) could promote self-efficacy and motivation (Zimmerman, 2000). The high rate of affirming feedback presents a valuable opportunity for students to receive positive reinforcement, much like they may receive from a supportive tutor or peer. Receiving real-time, responsive feedback can scaffold metacognition and reflection.

Of note, very few students admitted uncertainty (<1%) or gave incorrect responses (1%) suggesting that students either felt confident and actively engaged in looking up more information before responding, were not encouraged to express doubt or explore alternatives, or perhaps didn't experience uncertainty. It is also possible that the design of the activity, beginning with the students' current level of understanding, did not sufficiently challenge students. Future research should further explore whether students were guided to recognize their knowledge gaps and address them before or during their response, or whether prompts might be useful to create clear gaps in knowledge for students to experience and work through to support their learning (Loibl & Rummel, 2014).

Our third aim was to assess the consistency and quality of GAI interactions, and we observed somewhat variable patterns. Simple student responses were more consistent across individual students, with similar patterns within student in how frequently (or rarely) they give simple, low-level responses across topics of reviews. Further support that this is due to individual differences in students was that there was variability in the GAI low-level questions attributable to the prompts, yet this was not reflected in the student responses. On the other hand, a much higher portion of the variability in *high-level* student responses was attributed to the prompt used to initiate the activity (30%) than that explained at the student-level (4%), though only for high-level responses that were prompted. This suggests the potential importance of prompts to promote deeper cognitive engagement from students, and the opportunity for prompt engineering to elicit this depth.

This work demonstrates how interactive dialogue with GAI can provide meaningful and personally responsive questioning and feedback to students, supporting self-assessment in a low-stakes, formative approach. At a broader level, these assignments can help to support students' understanding of the ways GAI can be used to support learning, and to get exposed to and practice with prompt engineering, supporting the need for developing skills with AI in productive ways. It can also provide a more personalized and engaging way to review material outside of the classroom, and this specific activity is one model for incorporating new technology to encourage learning and critical thinking.

5 Limitations and Future Research

Limitations to this work include in its small sample size representing limited educational experiences. The students participating only represented about 55% of the class studied, and the class itself was a small, interactive class (29 students total). Future research could study similar patterns of interaction in other types of courses to explore whether the patterns observed here generalize. This work is also limited in its exploratory and correlational design, and in not including learning assessments separate from the activity studied. It will be important for future work to assess learning and link it to the interaction experiences. This work can be used to inform further work using control groups and outcome measures to assess the specific influences of the GAI in facilitating the Feynman experiences, including experimental tests of how to best promote high-quality interactions and learning.

The results presented here are a first step in exploring GAI-facilitated Feynman-style review interactions for learning. In addition to collecting more data to increase the students and content represented, the data presented here can be further analyzed to explore reciprocal dynamics in student–AI interactions, allowing us to predict what types of utterances lead to higher cognitive engagement. The dynamic sequence of reflection and revision in reciprocal dialogue is what matters for learning (Chi & Wylie, 2014), so looking at sequences of utterances will be meaningful to understand how these activities can provide meaningful learning experiences. We will also do further analysis of the characteristics of the prompts that students copied into the AI platform, coding for features such as length, specificity, and thematic focus, informing future design of prompts that can elicit higher quality interactions.

6 Conclusion

Feynman-style reviews using GAI can provide an interactive, personalized learning activity. Typical experiences with traditional quizzes used for review elicit low-level, simple responses, which these activities also showed. However, there was also evidence of metacognitive support during the interactions between GAI and students. Importantly, the results suggest that higher-level constructive or interactive engagement, which is conducive of greater learning, was more dependent on the contextual scaffolding provided by GAI prompts, indicating the potential for prompt engineering to support high-level cognitive engagement and learning in a personalized, scalable modality.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 2145478. 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 National Science Foundation. Implementation of the Feynman activity was part of course innovation supported by a Thrive grant from the UVA Center for Teaching Excellence. MM received summer support for part of this research from a UVA TYDE grant. Training support for NH, including contributions to the research reported here, is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200005 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We are grateful to Abigail Wilson and Maddie Zipper for their help with reliability coding and formatting the datasets for coding, along with Genesis Cartegena and Taylor Goree. We are especially thankful for Abby Krissinger for support with data management and preparing the data for coding and analysis, along with providing input across stages of the project.

References

Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review. *Information,* 15(11), 676.

Chan, C.K.Y., Hu, W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *Int J Educ Technol High Educ* 20, 43 (2023). https://doi.org/10.1186/s41239-023-00411-8

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science, 18*(3), 439-477.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist, 49*(4), 219-243.

de Fine Licht, K. (2024). Generative artificial intelligence in higher education: Why the 'banning approach' to student use is sometimes morally justified. *Philosophy & Technology,* 37(3), 113.

Dong, L., Tang, X., & Wang, X. (2025). Examining the Effect of Artificial Intelligence in Relation to Students' Academic Achievement in Classroom: A Meta-Analysis. *Computers and Education: Artificial Intelligence,* 100400.

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38*(4), 281-288.

Kim, J., Klopfer, M., Grohs, J. R., Eldardiry, H., Weichert, J., Cox, L. A., & Pike, D. (2025). Examining Faculty and Student Perceptions of Generative AI in University Courses. *Innovative Higher Education*, 1-33.

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., ... & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6, 100221.

Loibl, K., & Rummel, N. (2014). Knowing what you don't know makes failure productive. *Learning and Instruction*, 34, 74-85.

Rajesh, A., & Khan, S. (2024, December). Learn Like Feynman: Developing and Testing an AI-Driven Feynman Bot. In 2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 1-8). IEEE.

Resh, W. G., Ming, Y., Xia, X., Overton, M., Gürbüz, G. N., & De Breuhl, B. (2025). Complementarity, augmentation, or substitutivity? The impact of generative artificial intelligence on the US federal workforce. *arXiv preprint arXiv:2503.09637.*

Reyes, E. P., Blanco, R. M. F. L., Doroon, D. R. L., Limana, J. L. B., & Torcende, A. M. A. (2021). Feynman technique as a heutagogical learning strategy for independent and remote learning. *Recoletos Multidisciplinary Research Journal*, 9(2), 1-13.

Sun, L., & Zhou, L. (2024). Does generative artificial intelligence improve the academic achievement of college students? A meta-analysis. *Journal of Educational Computing Research,* 62(7), 1896-1933.

Torricelli, M., Martino, M., Baronchelli, A., & Aiello, L. M. (2024, May). The role of interface design on prompt-mediated creativity in Generative AI. In *Proceedings of the 16th ACM Web Science Conference* (pp. 235-240).

UNESCO. "Guidance For Generative AI In Education and Research." https://unesdoc.unesco.org/ark:/48223/pf0000386693 (2023).

Yang, J. W., Choe, K., Lim, J., & Park, J. (2025). Asking questions matters: comparing the effect of learning through dialogue with peers, ChatGPT, and intelligent tutoring system. *Interactive Learning Environments,* 1–17.

Zhu, Y., Liu, Q., & Zhao, L. (2025). Exploring the impact of generative artificial intelligence on students' learning outcomes: a meta-analysis. *Education and Information Technologies*, 1-29.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology,* 25(1), 82-91.

AI-Powered Coding of Elementary Students' Small-Group Discussions about Text

Carla M. Firetto[1], P. Karen Murphy[2], Lin Yan[1], & Yue Tang[2]

[1] Arizona State University; Mary Lou Fulton College for Teaching and Learning Innovation (MLFC); Tempe, AZ, USA
[2] The Pennsylvania State University; College of Education; University Park, PA, USA

Abstract

We present a novel application of an AI-powered approach for coding elementary students' small-group discussions about text. We used AILYZE to identify instances of individual and collective argumentation within a set of 371 transcripts. We gathered evidence of reliability (i.e., via comparability checks with human-produced codes) and criterion validity (i.e., via ground truth checks). There was sufficient agreement between AI-generated and human-produced codes, and initial validity evidence exceeded the established threshold of near-perfect agreement on a small ground truth check. Findings provide evidence that AI may serve to accurately code discussion transcripts in ways that were not previously feasible with only human-produced coding.

1 Introduction

Until recently, educational research examining the use of small-group discussions in preK-20 classrooms has been a resource-demanding area of study. Historically, quantitative analyses have required hand coding by research team members, which comes at significant time and cost expense (Longo, 2019; Murphy et al., 2018; Siiman et al., 2023). Consequently, troves of data often go un- or under-analyzed, yielding the potential loss of innumerable scientific advancements.

Recent developments in artificial intelligence (AI) now provide seemingly unlimited potential regarding automated AI-based discussion coding (Tran et al., 2024; Wang et al., 2024). In the present study, we build on the rapidly advancing work leveraging AI as a tool to code student discussions. Specifically, we present a novel application of an AI tool used to code small-group discussions about text along with the associated evidence of reliability and validity as part of a recent secondary analysis of small-group discussions (Firetto et al., 2025).

1.1 Value of Coding Discussion

There is a large body of research investigating the impact of small-group discussions in preK-20 classrooms and the myriad benefits on various outcome measures (e.g., comprehension, reasoning, transfer, motivation; Bae et al., 2021; Bennett et al., 2010; Murphy et al., 2009). Some of this research examines "after the discussion" or distal outcomes (e.g., class grades or test performance). For example, in our prior research, we found that elementary students engaging in small-group discussions evidenced increases in their written argumentation after discussions about what they read in their language arts class (Firetto et al., 2019; Murphy et al., 2022).

In contrast, other research examines learning based on what occurs "in" or "during" the discussions. For example, in our prior research, we identified indicators of high-level comprehension and tracked their frequency over time (Murphy et al., 2018). Coding and analyzing the discourse directly is particularly beneficial as it may allow for more accurate proximal measures without having to rely on transfer or delayed posttest measures. Researchers have explored a wide variety of coding schemes (Tao & Chen, 2023) and identified a variety of indicators present within the discussion that are associated with high-level comprehension (Soter et al., 2008) and other indicators of academic performance (Howe et al., 2019; Muhonen et al., 2018).

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 125–134
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

While there is generally a consensus that small-group discussions can benefit students' learning, there are many empirical questions that remain unanswered. For example, little is known about ways to group students: Should groups consist of students with similar or different ability levels? (Murphy et al., 2017); Should students be grouped in single-sex groups or mixed-sex groups? (Bennett et al., 2010). Moreover, there is also much to be learned about whether grade, content, or other factors may serve as moderating variables, ultimately impacting what we know about best practices and the associated recommendations for teachers.

While in-depth, qualitative examinations on smaller samples have contributed important findings toward these ends (e.g., Lobczowski et al., 2020), it is also beneficial for researchers in the field to conduct quantitative examinations derived from large samples (e.g., an experimental study testing multiple different group configurations vs a study with only a treatment and a control). Despite the importance, however, there are massive time and financial costs associated with coding a large corpus of discussion data (Murphy, 2015).

1.2 Leveraging AI to Code Discussion

Given these time and financial costs, researchers have long worked toward finding automated ways to expedite the process of coding. For example, several years ago, we used large language models to derive a series of potential indicators (e.g., complexity, oral expression), which we then compared to comprehension measures (Kosh et al., 2018). This allowed us to identify both word rareness and word diversity as indicators closely associated with students' posttest reading comprehension.

Since then, the ability to leverage AI as a tool to support automated coding processes has grown exponentially (Wang et al., 2024). This shift has meant moving beyond traditional classifiers toward sophisticated, transformer-based systems that track the ebb and flow of classroom talk. For example, in 2021, Song et al. used an artificial neural network-based model to classify the semantic content of classroom dialogue into eight categories. Not surprisingly, their findings indicate that performance (i.e., precision and recall) of the automated coding was better for some categories than others. For example, the prior-known knowledge and analysis categories were high, while other categories, like querying and speculation, were low. The overall F1 score (i.e., a measure of the accuracy of the codes calculated as the harmonic of precision and recall) across all categories was .680.

Advancements in AI are progressing at such a rapid pace that the potential for significant increases in accuracy and speed is growing every day. However, there are three areas, in particular, that need further exploration: (1) additional evidence demonstrating reliability and validity of automated codes along with comparisons to human coding; (2) transparency regarding the coding in ways that ensure model decisions are interpretable to teachers and researchers (i.e., *explainability*); and (3) closing the gap in AI-and-discourse research, whereby studies prioritize model building over *in situ* evaluation (Wang et al., 2024). Together, these issues underscore the need for further study and exploration.

1.3 The Present Study

Over the past year and a half, we conducted a secondary analysis of small-group discussions based on a large set of previously uncoded video-recorded small-group discussions collected as part of a large federally funded grant (R305A130031). We employed an AI-powered coding approach that allowed us to examine changes in students' individual and collective argumentation over time while also investigating the roles of genre and grade-level (Firetto et al., 2025).

Specific to the aims of AIMEcon (i.e., the theme "validity and reliability of AI-driven automated scoring systems"), the present study extends our previous work by examining the comparability of AI- and human-coded outcomes as well as the ways in which AI-powered coding can be rigorously employed. We explored two primary RQs:

RQ1: Are codes produced by AILYZE roughly comparable to those previously produced by humans? Does Cohen's Kappa agreement between AI-generated codes and human-produced codes meet or exceed .60 (i.e., *substantial* agreement)?

RQ2: Are codes produced by AILYZE accurate, based on a ground truth check? Does Cohen's Kappa agreement between AI-generated codes and human verifications meet or exceed .80 (i.e., *near-perfect* agreement)?

2 Method

2.1 Sample

The sample consisted of 371 transcripts of small-group discussions (i.e., typically 4-6 students per group). The discussions were conducted in fourth- and fifth-grade classrooms and collected over an entire school year. 3PlayMedia produced the transcriptions from video recordings using professional human transcribers (see Murphy, 2025). In accordance with our IRB protocol, research team members cleaned the transcripts to remove identifying information before entering the files into AILYZE (see Appendix A for detailed specifications).

2.2 Codes

We focused on identifying instances of two specific discourse indicators of high-level comprehension (i.e., individual and collective argumentation). *Individual argumentation* was intended to capture instances where a specific student produced an extended response that included multiple pieces of argumentation (e.g., a claim supported by reasoning and evidence). For example, a student explained, “I'd feel brave because, if I were Sahar, I would be going past the limit where I was supposed to be swimming. And I would be kind of a hero for saving that tiger.” This example illustrates individual argumentation as it included a specific claim about how the student believed they would feel if they were the main character (i.e., brave), along with two pieces of support for that claim (i.e., risk-taking by going beyond their swimming boundary and rescuing a tiger). Individual argumentation codes are informed by the notion of elaborated explanation (Chinn et al., 2000; Webb, 1991), which is a well-established discourse indicator of high-level comprehension.

Collective argumentation, on the other hand, represented episodes of talk where two or more students co-constructed understanding together. Importantly, our coding definition required the inclusion of an element of disagreement (e.g., a challenge or counterargument). For example, the discussion excerpt presented in Table 1 represents collective argumentation.

This example illustrates collective argumentation as it included multiple turns of students exploring the idea about whether the story was realistic or not, specifically the notion of whether “saving a polar bear” is something that one could realistically do, particularly given the massive size of adult polar bears. Collective argumentation codes are informed by the concept of exploratory talk (Mercer, 1995, 2000), another well-established discourse indicator of high-level comprehension.

Student A	Anything’s possible, especially something that is realistic. [referring to a previous statement about whether the story they read was something that could happen in real-life]
Student B	Except trying to save a polar bear. Might just be scared.
Student A	You could save polar bears.
Student B	Well, yeah, you could. Except, by the way, it is heavy because its really heavy. They almost weigh, like, thousand millions of pounds. [inaudible/interposing voices] Still, it's heavy.
Student A	They actually weigh, like 1,000 pounds.
Student B	Still, its heavy and bigger. You can get crushed.
Student C	Not a baby one.

Table 1: Talk Excerpt of Collective Argumentation

2.2.1 Previous Approach to Coding: Human-Produced Codes

Before new research assistants on our team code independently, they begin with an orientation to coding (e.g., reading the coding manual and related standard operating procedures), learn about the video recording software, and receive extensive mentoring with an experienced coach. Over the past decade, we have documented that it takes new human coders approximately 40 hours of coding training and practice to become relatively proficient at coding the recordings of small-group classroom discussion using our coding manual. Moreover, even after they have demonstrated proficiency, research assistants continue to engage in regular fidelity checks. Thus, as coders engage in the coding over time, 20% of the recordings are independently coded by a second research assistant, the codes are compared, and the two research assistants justify to each other why they coded or didn't code a specific event where there was a point of disagreement and then come to an understanding about which is the best fit. While time consuming, this procedure helps to maintain fidelity to the codebook. This is also due, in part, because one of the guiding principles of the codebook is to maintain low levels of inference (e.g., not to assume a student's intention or meaning), thus these fidelity checks also serve as an accountability check toward this standard.

2.2.2 Novel Approach to Coding: AI-generated Codes

In line with existing Human-in-the-Loop approaches to discourse analysis (e.g., Cohn et al., 2024), we leveraged AILYZE to produce AI-generated codes in a way that augmented, not replaced, expert judgment. In our case, we employed an approach in which the research team defined the codebook, designed the prompts, and decided the acceptance criteria prior to large-scale transcript coding for individual and collective argumentation.

Our initial plan involved deriving training examples from a sample of previously human-coded discussions to fine-tune the AI model to increase the coding accuracy. To do this, we transferred the human-produced codes from the video recording coding software onto the transcribed text documents for a sample of discussions that had been coded by two research assistants. However, the codes derived from the video recordings did not always translate directly and accurately to the transcripts. For example, in some cases, during the video recording, it was clear to observers that a single student articulated individual argumentation within a given turn, yet on the transcription, it might appear that this turn was interrupted (i.e., a student speaking over another and cutting one turn into multiple turns). In addition, the human coders in our lab and the humans who transcribed the recordings at 3PlayMedia may have had differences in what they heard and understood during the discussion, influencing how a word or phrase was interpreted or dismissed as inaudible. Ultimately, we decided not to include human-produced codes in the training of the AI model, and instead we used them to conduct a comparability check (i.e., RQ1).

Across multiple iterations, we revised a prompt informed by the definitions and coding criteria established in the Quality Talk coding manual (Murphy et al., 2017). We reviewed the AI-generated codes and corresponding justifications to refine the prompt, adding additional details as needed (e.g., the role of the teacher) and fixing data mapping issues (e.g., rows without dialogue).

Once the first and second authors independently agreed that we had developed a prompt that led to sufficiently accurate AI-generated codes, we conducted: (a) a comparability check, in which AI-generated codes were compared with previous human-produced codes from the video recordings, and (b) a ground truth check, in which the first two authors (each with hundreds of hours of discourse coding experience) collectively manually coded two transcripts to serve as the reference standard. Because both checks exceeded our pre-established thresholds (see results below), the AILYZE model was then applied to the deployment phase, coding the full set of transcripts (see Appendix B and Appendix C).

3 Results

3.1 RQ1

For the first research question, we examined the extent to which AI-generated codes were comparable to those previously produced by humans. We compared the AI-generated codes to the human-produced codes using 37 transcript excerpts (i.e., 10% of the total number) containing 3,249 turns. Due to the aforementioned difference in modality (i.e., transcript vs. video coding), we set our Cohen's Kappa threshold at .60, representing at least *substantial* agreement. Both codes exceeded this threshold: individual (Cohen's Kappa = 0.735, SD = 0.022, 95% CI [0.691, 0.775]) and collective argumentation (Cohen's Kappa = 0.849, SD = 0.014, 95% CI [0.823, 0.875]). Overall, there was sufficient consistency between the AI-generated codes from the transcripts and previously produced human codes from the video recordings.

To better illustrate the impact that modality may have had on coding, we identified an example from one of the discussions where there was disagreement between the AI-generated code and the human-produced code. The discussion excerpt in Table 2 begins after a statement made about there being lots of things to do outside.

Student C	I agree with Student B's idea because I would [INAUDIBLE] my house a lot. So I usually go outside. I have a trampoline so I can jump on that. But I go outside, and I pretend that I'm going to teach her something. I have magical powers and [INAUDIBLE].
Teacher	[Chuckles]
Student C	Because there's like—
Teacher	[Chuckles] I pretend I'm a teacher with magical powers too sometimes.
Student C	Yeah. Because we have a wooden hat. I get a stick and then [INAUDIBLE] or something.

Table 2: Talk Excerpt Illustrating AI/Human Disagreement

The first turn by Student C was identified as an instance of individual argumentation by AILYZE with the justification: *The claim is 'I agree with Student B's idea' supported by the reasons 'I usually go outside' and 'I pretend that I'm going to teach her something'. This provides a claim + multiple reasons + personal experience as evidence.* It is relatively clear from both the transcript and the AILYZE justification that this turn meets the criteria for an individual elaboration code. However, this turn was not coded by the research assistants. While there are many possible reasons why the human coders did not identify this turn as an instance of individual argumentation, the larger transcribed excerpt illustrates two possible explanations: (a) The teacher's chuckles and verbal/non-verbal input may have interrupted or influenced the student's talk as the human coders watched the video, which does not seem to be the case based on the way it was transcribed; (b) The human coders may have understood more or less of the words than the professional transcriber, who already noted "[INAUDIBLE]" in several places. This could have influenced the research assistants' decision to identify this turn as an instance of individual argumentation (e.g., hearing words that may have changed the meaning, aiming for a low inference interpretation of what they actually could hear).

3.2 RQ2

For the second research question, we aimed to gather evidence regarding whether the AI-generated codes were accurate based on a small ground truth check, where we selected two of the transcripts with both AI-generated and human-produced codes and then verified the accuracy of the codes at each turn (i.e., $n = 144$ transcript turns). Because we conducted the ground truth check using the text transcripts, we aimed for *near-perfect* Cohen's Kappa (i.e., at least .80) for the AI-generated codes. Both codes exceeded this threshold: individual (Cohen's Kappa = 1.00, SD = 0, 95% CI [1.00, 1.00]) and collective argumentation (Cohen's Kappa = 0.959, SD = 0.017, 95% CI [0.926, 0.991]). The AI-generated codes exceeded the pre-established threshold.

Given that the ground truth check was performed on transcripts that also had the human-produced codes (i.e., transferred from the video recordings), we also conducted an exploratory calculation of the Cohen's Kappa agreement for the human-produced codes: individual (Cohen's Kappa = 0.773, SD = 0.084, 95% CI [0.608, 0.937]) and collective argumentation (Cohen's Kappa = 0.573, SD = 0.041, 95% CI [0.493, 0.654]).

While it is important to underscore again the differences in modality (i.e., coding the transcripts vs coding the video recordings) as well as the relatively small sample size (i.e., two transcripts; $n = 144$ turns), it is noteworthy that both of the Cohen's Kappa values were higher for the ground truth-to-AI-generated codes than they were for ground truth-to-human-produced codes and that there was no overlap in the confidence intervals.

4 Conclusions

Human coding of qualitative data can be extensively resource-intensive (Longo, 2019; Siiman et al., 2023). AI-powered coding can decrease the resources needed to conduct such research and allow for scientific advancements that may not have been previously feasible (Feuston & Brubaker, 2021; Lixandru, 2024; Siiman et al., 2023; Tran et al., 2024). Our findings suggest that AI can be used to code discourse transcripts consistent with human coders when prompts and rules can be derived from established codes and evidence-based manuals (e.g., Murphy et al., 2017).

Notably, our AI coding was completed in a drastically less time than human coding would permit. As a point of reference, experienced coders require about an hour to code one small-group discussion. For this sample, that would have required roughly 371 hours of coding for the research assistants, a load which is typically split (i.e., about 185 hours each) between two coders. In addition, each coder would spend an additional 37 hours of coding the discussions for the fidelity check (i.e., 20%, as described above) and another 37 hours meeting with each other to discuss the instances of agreement/disagreement. Taken together, each coder would need to devote about 259 hours to coding the discussions, assuming they could code continuously. In our funded projects, we have hired undergraduate and graduate research assistants to code the discussions. Typically, graduate research assistants work 20 hours per week and attend classes. As a result, it would take the research assistants nearly an entire semester (i.e., 13 weeks) to complete the coding; however, in our experience, even expert humans cannot code accurately over long periods of time. As such, a more realistic estimate is that it would take two research assistants the better part of an academic year to code this many hours of video.

In contrast, AILYZE processed all 371 transcripts within ~12 hours (i.e., 2 minutes per transcript), clearly illustrating the potential to save time. Importantly, however, human time is still required to develop the prompts and check the AI over time, just as it is required as part of the process for developing, mentoring, and supporting research assistants when doing human-produced coding. Moreover, beyond these efficiency gains, reallocating research assistants' time to more enriching activities could help to move the field forward via increased productivity and dissemination possibilities (e.g., assisting in writing manuscripts, interpreting the data) and by better preparing them for their future research and career endeavors. Beyond the time costs, however, financial costs, environmental costs, energy costs and other costs need to be carefully considered and weighed.

Leveraging AI also has the potential to enhance coding consistency by reducing sources of variability that are common with human coders. As mentioned, human coders can be affected by limitations such as fatigue, overload, or selective attention when working with lengthy qualitative texts (Miles & Huberman, 1994). AI systems, by contrast, apply the same coding criteria uniformly across large corpora without a decline in performance over time. Although prior work has noted that AI-only coding can yield limited reliability (e.g., Prescott et al., 2024), we found that once the AI prompt was refined with human review, it achieved strong alignment with human codes. These findings highlight that while AI contributes efficiency and consistency, human oversight remains essential for guiding the framework, validating outputs, and ensuring methodological rigor.

A further benefit of using AILYZE was its capacity to generate explanations and justifications for coding decisions. This provided transparency that strengthened our human-in-the-loop process. Both in the validation and final coding review processes, we could monitor the rationales for AI-generated codes, which enabled us to identify points of alignment and divergence with human reasoning. This transparency underscores how AI can complement, rather than replace, expert judgment.

Importantly, within the context of our analysis, it is critical to note that the lower rates of agreement for the ground truth analysis and the human-produced codes do *not* negatively reflect on the quality of the human-produced coding. Rather, they highlight the role of modality differences. The human coders were coding with higher accuracy with video and audio, while AI coded based on the text transcripts. In essence, text "stands still." The reliability check, therefore, was to evaluate whether, despite these modality differences, the overall patterns of coding remained comparable across human and AI coders. Future research must take into consideration the potential impact of modality on the codes (e.g., transcripts vs. videos).

Our approach prioritized student privacy, data governance, and responsible use. In accordance with our IRB protocol, all transcripts were de-identified by the research team before any AI processing. We selected AILYZE in part because of its security features, auditability, and policy not to train on user data. To mitigate automation bias and address opacity, we used the aforementioned human-in-the-loop workflow, comparability, and ground truth checks prior to large-scale deployment as well as careful attention to known modality differences between transcripts and video recordings that can shape meaning. Collectively, these measures aim to reduce risks while ensuring transparent, auditable, and pedagogically responsible use of classroom discourse data.

4.1 Future Research

In the present study, we were able to document an AI-powered coding process that took less time than traditional human coding. However, future research in this area should also consider other important variables beyond time, including costs such as environmental impacts, electricity and/or water usage (Kandemir, 2025), and conduct a true cost–benefit analysis.

During our oversight process, we noted that students' individual and collective argumentation represented a variety of quality (e.g., some students made sophisticated arguments with counter-arguments, rebuttals that extended over a long period of time, while others were more simplistic and succinct). Moving forward, it is critical to understand more about the quality of students' responses, beyond just coding the presence or absence of argumentation within the discussions. While such coding may not be feasibly possible with human coding, we think that it is possible to extend the procedures we employed herein to move beyond binary codes (i.e., presence or absence of a code for a given turn) to develop an AI-generated *quality* score that can capture characteristics such as accuracy, depth, and length.

Finally, now that we have established an automated coding that meets the requisite criteria we established, we can begin to explore how other aspects can be automated. For example, Li et al., 2025 found that AI-generated feedback about classroom discussions was useful for teachers, and thus, coding may also be used to support coaching.

Acknowledgments

We acknowledge the contributions of our collaborators related to the larger project that this study was drawn from: Emily Starrett, Emilee A. Herman, and Jeffrey A. Greene. We also appreciate the support of James Goh for assisting us during the AI-powered coding procedures and for providing detailed descriptions of AILYZE (see Appendix A).

Funding

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130031 to the Pennsylvania State University. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not represent the views of the Institute or the U.S. Department of Education. Additional support was also provided by MLFC internal grant funding at Arizona State University and McMichael Professorship in the School of Education at the University of North Carolina at Chapel Hill. This material is based upon work supported by the National Science Foundation under Award No. (1912415). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest

The authors report no known conflict of interest. The authors purchased use of AILYZE for the purpose of this project and have no financial relationship with AILYZE.

References

AILYZE, Inc. (n.d.). *AILYZE.* https://www.ailyze.com/

Bae, C. L., Mills, D. C., Zhang, F., Sealy, M., Cabrera, L., & Sea, M. (2021). A systematic review of science discourse in K–12 urban classrooms in the United States: Accounting for individual, collective, and contextual factors. *Review of Educational Research*, *91*(6), 831-877. https://doi.org/10.3102/00346543211042415

Bennett, J., Hogarth, S., Lubben, F., Campbell, B., & Robinson, A. (2010). Talking science: The research evidence on the use of small group discussions in science teaching. *International Journal of Science Education*, *32*(1), 69-95. https://doi.org/10.1080/09500690802713507

Chinn, C. A., O'Donnell, A. M., & Jinks, T. S. (2000). The structure of discourse in collaborative learning. *The Journal of Experimental Education*, 69(1), 77–97. https://doi.org/10.1080/00220970009600650

Cohn, C., Snyder, C., Montenegro, J., & Biswas, G. (2024). Towards a human-in-the-loop LLM approach to collaborative discourse analysis. In *International Conference on Artificial Intelligence in Education* (pp. 11–19). Springer Nature Switzerland.

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://dl.acm.org/doi/pdf/10.1145/3479856

Firetto, C. M., Murphy, P. K., Greene, J. A., Li, M., Wei, L., Montalbano, C., Hendrick, B., & Croninger, R. M. V. (2019). Bolstering students' written argumentation by refining an effective discourse intervention: Negotiating the fine line between flexibility and fidelity. *Instructional Science,* 47, 181–214. https://doi.org/10.1007/s11251-018-9477-x

Firetto, C. M., Murphy, P. K., Starrett, E., Herman, E. A., Greene, J. A., Tang, Y., & Yan, L. (2025). Investigating grade-level and text genre effects in Quality Talk discussions: An AI-powered discourse analysis of upper primary students' high-level comprehension. *Learning and Instruction*, 100, 102208. https://doi.org/10.1016/j.learninstruc.2025.102208

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Kandemir, M. (2025, April 8). *Why AI uses so much energy-and what we can do about it*. Institute of Energy and the Environment. https://iee.psu.edu/news/blog/why-ai-uses-so-much-energy-and-what-we-can-do-about-it

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice,* 37(2), 20–34. https://doi.org/10.1111/emip.12174

Li, X., Han, G., Fang, B., & He, J. (2025). Advancing the in-class dialogic quality: Developing an artificial intelligence-supported framework for classroom dialogue analysis. *Asia-Pacific Education Researcher,* 34, 495–509. https://doi.org/10.1007/s40299-024-00872-z

Lixandru, D. (2024). The use of artificial intelligence for qualitative data analysis: ChatGPT. *Informatica Economica*, *28*(1). https://doi.org/10.24818/issn14531305/28.1.2024.05

Lobczowski, N. G., Allen, E. M., Firetto, C. M., Greene, J. A., & Murphy, P. K. (2020). An exploration of social regulation of learning during scientific argumentation discourse. *Contemporary Educational Psychology,* 63, https://doi.org/10.1016/j.cedpsych.2020.101925

Longo, L. (2019). Empowering qualitative research methods in education with artificial intelligence. In *World Conference on Qualitative Research* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-31787-4_1

Mercer, N. (1995). The guided construction of knowledge: Talk amongst teachers and learners. Multilingual Matters.

Mercer, N. (2000). *Words and minds: How we use language to think together*. Routledge. https://doi.org/10.4324/9780203464984

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Muhonen, H., Pakarinen, E., Poikkeus, A. M., Lerkkanen, M. K., & Rasku-Puttonen, H. (2018). Quality of educational dialogue and association with students' academic performance. *Learning and Instruction*, *55*, 67-79. https://doi.org/10.1016/j.learninstruc.2017.09.007

Murphy, P. (2025). Investigating grade-level and genre effects in Quality Talk discussions: An AI-powered discourse analysis of upper primary students' high-level comprehension. *Databrary*. https://databrary.org/volume/1858

Murphy, P. K. (2015). Mooring points and touchstones along the road to school-based interventions—An introduction. *Contemporary Educational Psychology, 40*, 1-4. https://doi.org/10.1016/j.cedpsych.2014.10.003

Murphy, P. K., Firetto, C. M., Greene, J. A., & Butler, A. M. (2017). *Analyzing the talk in Quality Talk discussions: A coding manual.* http://doi.org/10.18113/S1XW64

Murphy, P. K., Greene, J. A., Firetto, C. M., Croninger, R. M., Duke, R. F., Li, M., & Lobczowski, N. G. (2022). Examining the effects of Quality Talk discussions on 4th- and 5th-grade students' high-level comprehension of text. *Contemporary Educational Psychology,* 71, https://doi.org/10.1016/j.cedpsych.2022.102099

Murphy, P. K., Greene, J. A., Firetto, C. M, Li, M., Lobczowski, N. G., Duke, R. F., & Croninger, R. M. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology*, *51*, 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B., Li, M., Montalbano, C., & Wei, L. (2018). Quality Talk: Developing students' discourse to promote high-level comprehension. *American Educational Research Journal,* 55(5), 1113–1160. https://doi.org/10.3102/0002831218771303

Murphy, P. K., Wilkinson, I. A., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students’ comprehension of text: A meta-analysis. *Journal of Educational Psychology*, 101(3), 740. https://doi.org/10.1037/a0015576

Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*, e54482. https://doi.org/10.2196/54482

Siiman, L. A., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023, August). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In *International Conference on Innovative Technologies and Learning* (pp. 87-96). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_9

Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2021). Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research,* 59(3), 496–521. https://doi.org/10.1177/0735633120968554.

Soter A. O., Wilkinson I. A. G., Murphy P. K., Rudge L., Reninger K., Edwards M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47, 372–391.

Tao, Y., & Chen, G. (2023). Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature. *Learning, Culture and Social Interaction*, *39*. https://doi.org/10.1016/j.lcsi.2023.100702

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L.C. (2024). Analyzing large language models for classroom discussion assessment. arXiv. https://doi.org/10.48550/arXiv.2406.08680

Wang, D., Tao, Y., & Chen, G. (2024). Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, *123*. https://doi.org/10.1016/j.ijer.2023.102275

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education*, *22*(5), 366-389.

5 Appendices

5.1 Appendix A

AILYZE Specifications: Reported by AILYZE

AILYZE's LLMs (i.e., a mix of Grok-1, Mistral 8x22B, and Phi-2.5-MoE) are trained on curated, high-quality open corpora commonly used in multilingual and scholarly modeling, such as the UN Parallel Corpus (multilingual proceedings and debates), ParlaMint (TEI-standard legislative proceedings with speaker-linked context), and S2ORC (millions of scholarly articles with citations and structure). Importantly for the context of the present study, it is also trained on education-specific transcripts, such as TalkBank’s ClassBank (curated classroom discourse collections), the NCTE Elementary Math Classroom Transcripts (1,660 lessons from 4th–5th grade), and the Teacher–Student Chatroom Corpus (one-to-one teacher–learner lessons). This mix supports discourse-focused tasks, including segment-level coding with grounded justifications.

In addition, AILYZE complies with measures outlined in the HECVAT (Higher Education Community Vendor Assessment Toolkit), which is a standardized framework developed by higher education organizations to assess data and AI risks associated with technology services. AILYZE does not train on user data and all project data, prompts, and outputs are encrypted, access-controlled, and exportable for archiving. All runs are also versioned so that the same codebook, same engine version and same transcripts yield identical results, ensuring full reproducibility.

AILYZE’s deterministic inference setting was used, which locked the codebook and engine version for the entire run. This ensures that if future researchers re-run the same transcripts with the same project configuration, they will obtain identical labels and justifications, supporting fully reproducible analyses.

5.2 Appendix B

Individual Argumentation Prompt:
“Code the interview transcript to identify all instances of elaborated explanations. Elaborated explanations are instances in which students explain their thinking in fairly coherent form to others. They occur in a single turn where a student explains how he or she arrived at a conclusion or idea by giving a step-by-step description or detailed account of how the conclusion or idea was reached or how a problem might be resolved. They are elaborated descriptions of how things work, why some things are the way they are, or how they should be thought about. They include details of how to think about an issue and justification or rationale for thinking that way. Elaborated explanations relate to the quality of explanations given by an individual student, not a collective of students, and not the teacher. They can take various forms including: claim + 2 or more independent reasons, claim + 2 or more conjunctive reasons, claim + 2 or more causally connected reasons, claim + reason(s) + evidence, claim + reason(s) + warrant, or claim + evidence + evidence. Elaborated explanations must include all components within a single "turn." The components begin at the start of the claim and continue through the end of the speaker's turn, unless the topic shifts away from that claim. A claim may be implied in verbal discourse when it immediately follows a question, but the response must directly respond to or follow from a question within the same question event. Elaborated explanations can only occur within authentic question events; responses to test questions cannot be coded as elaborated explanations.”

5.3 Appendix C

Collective Argumentation Prompt:
“Code the interview transcript to identify all instances of exploratory talk. Exploratory talk occurs when students share, evaluate, and build knowledge over at least three turns. It is talk in which partners engage critically but constructively with each other's ideas where relevant information is offered for joint consideration. Proposals may be challenged and counter-challenged but, if so, reasons are given and alternatives are offered. Agreement is sought as a basis for joint progress, with knowledge made publicly accountable and reasoning visible in the talk. It embodies a kind of 'co-reasoning,' with speakers following ground rules which help them to share knowledge, evaluate evidence, and consider options in a reasonable and equitable way. The key component of exploratory talk is the element of challenge, with only one challenge statement necessary for an episode to be classified as exploratory talk. Exploratory talk episodes consist of instances where students co-construct understanding over at least three consecutive, uninterrupted turns about the same topic. Exploratory talk is characterized by students actively constructing knowledge with students primarily interacting with and talking to each other. Episodes end when the topic shifts, someone asks a different question, a statement is made that deviates from the trajectory, or the students arrive at consensus. A student must initiate the challenge for talk to be considered exploratory. The teacher can be present but is not influencing the discourse or episode of talk. Exploratory talk in essence is a way of using language to think collectively—to 'interthink.'”

Evaluating the Reliability of Human–AI Collaborative Scoring of Written Arguments Using Rational Force Model

Noriko Takahashi
Montclair State University
takahashin1@montclair.edu

Abraham Onuorah
Montclair State University
onuoraha1@montclair.edu

Alina Reznitskaya
Montclair State University
reznitskayaa@montclair.edu

Evgeny Chukharev
Iowa State University
evgeny@iastate.edu

Ariel Sykes
Montclair State University
Sykesa@montclair.edu

Michele Flammia
Independent Researcher
micheledapila@gmail.com

Joe Oyler
Maynooth University
joe.oyler@mu.ie

Abstract

This study aims to improve the reliability of a new AI collaborative scoring system used to assess the quality of students' written arguments. The system draws on the Rational Force Model and focuses on classifying the functional relation of each proposition in terms of support, opposition, acceptability, and relevance.

We evaluated GPT-4o under zero-shot and few-shot prompting. Results show that few-shot prompting improved classification accuracy: Acceptability Support (AS) reached an F1 score of 0.95, Relevance Support (RS) rose from 0.08 to 0.72, and Acceptability Objection (AO) increased from 0.42 to 0.74. Relevance Objection (RO) was rare but false positives decreased. Error analysis revealed that misclassifications often stemmed from overreliance on lexical cues rather than contextual nuance. For instance, GPT-4o tended to treat extreme words like never or any as objections, even when the context indicated support. These findings highlight the potential of RFM-guided prompts to enhance automated essay scoring and provide more reliable, reasoning-focused feedback.

1 Introduction

Research on automated essay scoring (AES) for argumentative writing has advanced significantly over the past decade. Foundational studies established methods for identifying core argumentative elements such as claims, reasons, and evidence (Stab and Gurevych, 2014; Persing and Ng, 2015). Building on this foundation, more recent systems increasingly employ transformer-based large language models (LLMs), including BERT, GPT, and LLaMA, to improve scoring accuracy and robustness. For example, Carlile et al. (2018) created a dataset of student essays labeled for persuasiveness and related qualities, offering early resources for argumentative writing research, while Toledo et al. (2019) leveraged BERT-based architectures to rank arguments. Hicke et al. (2023) introduced a transformer-based method for labeling persuasive segments as "effective" or "ineffective," reaching near-human performance. Similarly, Sun and Wang (2024) developed a multi-dimensional model that assesses vocabulary, grammar, and coherence with high predictive accuracy.

Despite these gains, most AES systems still operate at the level of isolated features or segments and therefore struggle to capture how propositions interconnect to form a coherent line of reasoning. Argumentative writing unfolds through chains of interdependent propositions: some supply direct evidence, others provide conceptual linkage, and still others contest earlier claims. Treating these components independently obscures the discourse-level relationships that determine overall logical quality and persuasiveness. Modeling these relationships remains a central challenge.

This limitation is especially consequential in educational settings. Scholars have argued that emphasizing the mere presence of claims, evidence,

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 135–140
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

and counterarguments can divert attention from the coherence and quality of reasoning (Chinn et al., 2016; Newell et al., 2011; Rapanta et al., 2013). Backman et al. (2023) further contend that such structural checklists can impede both teachers and students from developing a deeper understanding of what distinguishes strong from weak arguments. Accordingly, automated scoring should be aligned with educational perspectives that prioritize the quality of reasoning, not just its components.

To address the limitations of the current approaches, we adopt the Rational Force Model (Naess, 1959; Backman et al., 2012, 2023), a framework that evaluates important but largely overlooked dimensions of argument quality, specifically focusing on the relational role each proposition plays in connection to another, as well as its acceptability and relevance. We discuss this framework next.

2 Rational Force Model (RFM)

The Rational Force Model (RFM), developed by Naess (1959) and extended by other researchers (e.g., Backman et al., 2012, 2023; Björnsson et al., 1994), provides a fine-grained framework for evaluating argumentative quality. Rather than focusing on the mere presence of certain argument elements (claims, reasons), RFM centers on both the proposition's function and its epistemic strength, thus examining how a proposition supports or opposes another proposition within the overall line of reasoning.

According to a more recent version of the RFM, discussed by Backman and colleagues (2023), RFM proceeds in two phases. In the descriptive (reconstruction) phase, a text is segmented into discrete idea units (propositions). Each proposition is mapped to a target (the main claim or another proposition) and classified by intended function: Acceptability Support (AS): A proposition intended to increase another proposition's acceptability, Acceptability Objection (AO): A proposition intended to decrease another proposition's acceptability, Relevance Support (RS): A proposition intended to increase another proposition's relevance, or Relevance Objection (RO): A proposition intended to decrease another proposition's relevance, as shown in Table 1. This reconstruction yields a directed structure of support and opposition.

In the evaluative (scoring) phase, each proposition receives two scores: Acceptability (A), the degree to which there is reason to believe the proposition is true; and Relevance (R), the degree to which, if true, the proposition advances resolution of the issue or supports its target. The proposition's rational force is the product of these values: $\mathrm{RF}_i = A_i R_i$.

In sum, RFM highlights not just the presence of argumentative components, but their functional roles, accuracy, and logical strength. As such, RFM provides a principled basis for analyzing important, but largely overlooked, dimensions of written arguments, thus generating valuable diagnostic information to support meaningful feedback.

	Support	**Objection**
Acceptability	Acceptability Support(AS)	Acceptability Objection(AO)
Relevance	Relevance Support (RS)	Relevance Objection (RO)

Table 1: Four types of propositions in an RFM analysis.

Note. Adapted from Backman et al. (2012, 2023).

3 Aims

The present study is part of a larger project (Reznitskaya et al., 2025) aimed at developing a related AES system. Here, we focus on one key component of that effort: evaluating the ability of AES systems to assess not just individual propositions, but the relationships between them.

Specifically, we frame each proposition in terms of its function—Acceptability Support (AS), Relevance Support (RS), Acceptability Objection (AO), or Relevance Objection (RO)—within the structure of reasoning. By treating function identification as a classification task, we examine the extent to which AI systems can recover the relational architecture of arguments.

Our research questions (RQ) are:

RQ1: Can GPT-4o reliably classify the functional relation between two propositions as AS, RS, AO, or RO, compared to a human label?

RQ2: Does few-shot prompting improve GPT-4o ability to distinguish these roles?

4 Sample

Our study draws on a corpus of 504 argumentative essays written by Grade 5 students (10–11 years old) in public schools at two research sites in the United States (New Jersey and Ohio). The essays were collected as part of a quasi-experimental study aimed at improving students' argumentation skills (Wilkinson et al., 2023; Reznitskaya and Wilkinson, 2020).

In New Jersey (n = 239), students were primarily White (60.7

The writing task was based on a short story, The Pinewood Derby (776 words), in which a boy named Jack faces a moral dilemma of whether to report his classmate Thomas, who cheated by not building his model car himself. After hearing and reading along with the story, students were asked to write a letter to their teacher explaining whether Jack should tell on Thomas, supporting their opinion with reasons and evidence, addressing possible counterarguments, and concluding their response. Students were given 25 minutes to complete the task, which pilot studies confirmed was sufficient time.

From this larger dataset of 504 essays, we randomly selected 25 essays for detailed manual annotation in the current study.

5 Method

Each essay was segmented into idea units. An idea unit "expresses one action or event or state, and generally corresponds to a single verb clause" (Mayer, 1985, p. 71). This segmentation step ensures that long or complex student sentences are broken down into smaller, analytically meaningful parts, each representing a distinct claim or piece of reasoning. For example, the sentence "Thomas should tell on Jack because he cheated" would be divided into two idea units: one expressing the main claim ("Thomas should tell on Jack"), and another the supporting reason ("he cheated").

Within the RFM framework, each idea unit, we called a source, is aimed at one other idea unit, we called a target. Trained annotators labeled each source idea unit specifying its relation to the target, selecting from the four RFM categories: Acceptability Support (AS), Relevance Support (RS), Acceptability Objection (AO), or Relevance Objection (RO) (see Table 1).

To improve clarity and reduce ambiguity, annotators also created a reconstructed, or standardized idea unit, which paraphrased the student's statement into its core meaning. These reconstructions, shown in brackets [ ], helped resolve cases where children's writing was unclear, incomplete, or colloquial. For example, if a student wrote "There is no reason to feel sorry for Thomas," the reconstructed idea unit might be [Mean people don't deserve empathy], ensuring that the intended meaning was explicit. This process was essential for maintaining consistency in annotation and for allowing both human raters and the AI system to work with clearly defined propositions (see Table 2).

We analyzed 200 pairs of annotated idea units and examined the reliability of GPT classification compared to a human label. We used GPT-4o with prompt strategies. The target–source pairs were provided, but without human labels. GPT was prompted to assign an RFM label to the target idea unit. We had two experimental conditions: a zero-shot prompt and a few-shot prompt. In the zero-shot condition, we provided general instructions about RFM labels without examples. In the few-shot condition, the labels were explained in more detail with examples and exceptions to provide clearer classification guidelines.

No.	Target Idea	Source Idea	Label
1	because Thomas didn't make the model car by himself. [Thomas didn't build the car on his own]	I think so because in the story it said "No my brother did it." [Thomas said "No, my brother made it"]	AS
2	This shows that he is very mean. [Thomas was mean]	I think Thomas should be nicer to other people. [Thomas should be nicer to people]	RS
3	because Thomas cheated and he won. [Thomas won by cheating]	And that's not fair to everyone else [Winning by cheating is unfair]	AS
4	Some people might say "no" because he was unliked by many kids [Students didn't like Thomas]	There is no reason to feel sorry for Thomas [Mean people don't deserve empathy]	AO

Table 2: Sample of dataset with annotated labels and reconstructed idea units.

Category	Zero-shot			Few-shot		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
AS	0.83	0.70	0.76	0.96	0.94	0.95
RS	0.06	0.14	0.08	0.81	0.64	0.72
AO	0.89	0.42	0.42	0.63	0.90	0.74
RO	0	0	0	0	0	0

Table 3: F1 scores for each category.

6 Results

In comparing the zero-shot and few-shot prompt results (see Table 3), F1 scores improved overall in the few-shot condition. The AS category showed an increase in F1 score from 0.76 to 0.95. The most substantial change occurred in RS, which increased from 0.08 to 0.72. AO also showed improvement, with F1 rising from 0.42 to 0.74. RO, which rarely appeared in the essays and was not part of human labeling, remained at 0 for both prompts. However, the False Positives for RO decreased from 8 in the zero-shot prompt to 1 in the few-shot prompt, indicating an improvement.

The few-shot prompt (see Table 4) contributed substantially to the improvement in RS and AO by providing clearer definitions for these categories. The model struggled to distinguish between AS and RS, as well as AS and AO in the zero-shot prompt, so differences between them were added in the few-shot prompt. The few-shot prompt also encouraged considering the context of the Pinewood Derby story and the student's likely intent.

Based on these results, the answers to the research questions are:

RQ1: GPT-4o reliably classified the functional relation for AS with a high F1 score. Other categories were also reliably classified with the few-shot prompt.

RQ2: Few-shot prompting improved GPT-4o's performance, particularly for RS and AO. However, RS and AO still show variability in Precision and Recall, indicating areas that require further refinement.

7 Discussion

The results demonstrate that the few-shot prompt improved GPT-4o's ability to classify functional relations, particularly RS and AO. However, further refinement is needed. For example, for the target idea "Thomas never did any of the hard work," the source "Thomas painted and decorated his car" was labeled AS by human annotators because it provides evidence for the claim 'did not do hard work,' as painting and decorating are considered easy tasks in the context of the story. In contrast, GPT-4o labeled it as AO, since the source describes Thomas doing some work, even though it's considered less difficult. This discrepancy likely arises because GPT-4o tends to focus on extreme expressions like 'never' or 'any,' which it interprets as strong markers of absolute negation. As a result, GPT struggles to account for the nuanced difference between what is considered 'hard work' versus 'easy work' in the context of the story. This highlights the need for further improvement in the few-shot prompt.

This study highlights the potential to improve the Human–AI Collaborative Scoring system using the RFM framework. It supports the development of scoring procedures that (1) target theoretically and pedagogically important aspects of argument quality and (2) can be applied reliably to naturally occurring student arguments. Despite some inconsistencies in the scoring system due to the variety of propositions in student essays, the results suggest a path toward refining rules to handle exceptional cases.

Limitations

This study has several limitations. First, the distribution of categories was uneven, with Relevance Objections (RO) almost absent in the student essays. As a result, the model's performance on this category could not be meaningfully evaluated. Second, the study focused on a single model (GPT-4o) under two prompting conditions (zero-shot and few-shot), which limits the scope of the findings. Additional experiments with other models, prompting strategies, and fine-tuning approaches are needed to test the robustness of the results. Finally, annotation according to the Rational Force Model (RFM) involves nuanced judgments of acceptability and relevance, which can be open to interpretation. Disagreements among annotators may influ-

	Zero-shot	**Few-shot (Added)**
AS	The *source* strengthens the truthfulness or plausibility of the *target*	(+) AS if the *source* answers "Why believe?"
RS	The *source* strengthens the relevance or usefulness of the *target*	(+) RS if the *source* answers "Why care?" or adds moral/social importance (+) RS if the *source* explains general moral or social norms, not AS
AO	The *source* challenges truthfulness or plausibility of the *target*	(+) AO if the *source* disagrees with truthfulness or plausibility of the *target*, or pushes back against it (rebuttal)
RO	The *source* challenges the relevance or usefulness of the *target*	(+) RO if the *source* disagrees with relevance or usefulness of the *target*, or pushes back against it (rebuttal)
Others	For each pair, classify the *source* label in relation to the *target*	(+) Use only AS or AO when the *target* is a main claim (+) Reference the story "Pinewood Derby" (+) If a sentence is unclear, use the [bracketed] reconstructed idea unit to understand each idea clearly

Table 4: Zero-shot vs. Few-shot prompting comparison.

ence the gold-standard labels and, in turn, affect the evaluation of model accuracy. Strengthening inter-annotator reliability therefore remains an important direction for future studies.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 2506473 and 2506474.

References

Ylva Backman, Viktor Gardelli, Tobias Gardelli, and Anders Persson. 2012. *Scientific Thinking Tools: A Base for Academic Studies*. Studentlitteratur.

Ylva Backman, Alina Reznitskaya, Viktor Gardelli, and Ian A. G. Wilkinson. 2023. Beyond structure: Using the rational force model to assess argumentative writing. *Written Communication*, 40(2):555–585.

Gunnar Björnsson, Ulrik Kilhbom, Folke Tersman, and Anders Ullholm. 1994. *Argumentationsanalys*. Natur och Kultur.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Clark A. Chinn, Richard G. Duncan, Lung-Chi Hung, and Robert W. Rinehart. 2016. Epistemic criteria and reliable processes as indicators of argument quality in science students' argumentation. In *Proceedings of the Annual Meeting of the American Educational Research Association (AERA 2016)*, Washington, DC, USA.

Yann Hicke, Tonghua Tian, Karan Jha, and Choong Hee Kim. 2023. Automated essay scoring in argumentative writing: Deberteachingassistant. arXiv preprint arXiv:2307.04276.

Richard E. Mayer. 1985. Structural analysis of science prose: Can we increase problem solving performance? In *Understanding of Expository Text*, pages 65–87. Erlbaum, Hillsdale, NJ.

Arne Naess. 1959. *Communication and Argument: Elements of Applied Semantics*. Allen & Unwin.

George E. Newell, Richard Beach, Jennifer Smith, Jennifer VanDerHeide, Deanna Kuhn, and Jeroen Andriessen. 2011. Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly*, 46(3):273–304.

Isaac Persing and Vincent 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)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Chrysi Rapanta, Maria Garcia-Mila, and Sergi Gilabert. 2013. What is meant by argumentative competence?

an integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83(4):483–520.

Alina Reznitskaya, Michele Flammia, Noriko Takahashi, Abraham Onuorah, Ariel Sykes, Joe Oyler, and Evgeny Chukharev. 2025. Enhancing diagnostic and instructional value of assessments designed to evaluate written arguments. In *EARLI Conference: Realising Potentials through Education: Shaping the Minds and Brains for the Future*, Graz, Austria.

Alina Reznitskaya and Ian A. G. Wilkinson. 2020. Measuring production and comprehension of written arguments in upper-elementary grades. *Studia Paedagogica*, 24(Special Issue on Argumentation):63–84.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Kun Sun and Rong Wang. 2024. Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. arXiv preprint arXiv:2406.01198.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment – new datasets and methods. 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 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Ian A. G. Wilkinson, Alina Reznitskaya, and Joseph V. D'Agostino. 2023. Professional development in classroom discussion to improve argumentation: Teacher and student outcomes. *Learning and Instruction*, 85:101732.

Evaluating Deep Learning and Transformer Models on SME and GenAI Items

Joe Betts and William Muntean
National Council of State Boards of Nursing (NCSBN)
Chicago, IL

Correspondence: jbetts@ncsbn.org

Abstract

An important and time-consuming aspect of test development is the metadata coding of items within the item bank to be ready for use within the test blueprint. This study leverages deep learning, transformer models, and generative AI to streamline test development by automating metadata tagging predictions to reduce the necessary review time for subject matter experts (SME). Transformer models outperform simpler approaches and provide a direct method for reducing SME workload.

Keywords: classification models, deep learning, transformer models, Generative AI, automated item generation

1 Introduction

Developing good assessments is a time-consuming and intricate process involving numerous phases, stages, and tasks (Downing, 2006; Lane, Raymond, Haladyna, & Downing, 2016). When developing items for any assessment, subject matter experts (SMEs) are tasked with writing and reviewing items. This has traditionally been a long and expensive set of tasks. However, another aspect of the development process that is necessary but not usually identified as a high priority is related to tagging all items for their metadata content.

Metadata tagging involves classifying items by domain, task, cognitive complexity (e.g., Bloom's Taxonomy or Webb's Depth of Knowledge), or other contextual factors within a test blueprint. For this task, the SME must read each item specifically and either validate that it is representative of the task and/or domain indicated or they must provide the coding for that item. This process is labor-intensive, typically requiring SMEs to tag items or validate tags, often involving secondary reviews. However, of all the metadata content, the most important classification has to do with assigning each item to the correct test blueprint domain to ensure appropriate content to test plan blueprint alignment.

Deep learning (DL) methods and large language models (LLMs) should be useful tools in this venture as they are potentially adept at utilizing textual relationships and making predictions about content classifications. DL models, particularly those using text classification and transformer-based embeddings, can potentially reduce this workload by automating metadata tagging. This research explores how different DL and LLMs could be used to make predictions about metadata classification. Thus, building a strong model and automated pipeline could reduce SME work substantially for other work.

This study investigates DL and transformer models for natural language processing (NLP) to classify test items into test plan domains. It evaluates accessible models including Naïve Bayes (Friedman, Geiger, & Goldszmidt, 1997), XGBoost (Chen & Guestrin, 2016), deep learning models (Goodfellow, Bengio, & Courville, 2016), and some BERT family transformer models to evaluate the best approach to predicting item domain classifications. The research compares basic models to identify the

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 141–146
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

most effective approaches before exploring more complex architectures. Additionally, the best fitting model will be used against basic item generated by two different LLM models (GenAI) to see how well the model built from human curated items generalizes to GenAI.

2 Methods

2.1 Sample and GenAI Items

The study uses a subset of an item bank (N = 6,839), split into 60% training, 20% validation, and 20% testing datasets. Items were randomly selected from the group of items that had passed all statistical pretest criteria. The text data consists of item stems for multiple-choice items, labeled by eight high-level test blueprint domains (NCSBN, 2023). Table 1 provides the name of each domain and the distribution of items from each domain. The average length of the stem was 70 words (sd = 21).

For generative items using GenAI, 149 task statements (NCSBN, 2023, 2025), each tied to one of the eight domains, were used to generate 298 items across two LLMs. The prompt did not use any context about the item domain in the item generation but rather only uses the task/activity statements. Here is the prompt text:

"You are a nurse educator and clinician. Write a multiple-choice item for this entry-level nursing task: {{activity}}. The item must be a challenge to answer for an entry-level nurse. The item content should be related to this specific activity: {{activity}}. Make sure that the item content is relevant to the nursing activity. Have the item incorporate usual situations where a nurse would normally have to perform this nursing task. Provide only the item stem and the options in a json format. Do not include any other text. Do not include any explanations or rationale. Do not include any references. Do not include any other text."

Domain ID	Domain Label	Number of Items
0	Management of Care	1,202
1	Safety & Infection Control	792
2	Health Promotion & Maintenance	691
3	Psychosocial Integrity	633
4	Basic Care & Comfort	596
5	Pharmacological & Parenteral Therapies	1,202
6	Reduction of Risk Potential	721
7	Physiological Adaptation	1,002

Table 1 Domains and Number of Items

Items were created using zero-shot learning with Llama 4 Maverick (Meta AI, 2025, mixture of experts) and Claude 3.7 Sonnet (Anthropic, 2025, hybrid reasoning). While the activity statements are nested within a domain, using only the activity statements eliminates the domain context for the item generation task.

2.2 DL and Transformer Models

The analysis used two baseline models for comparing more complex DL models. These were Naïve Bayesian (NB) and XGBoost (XG). Both models used TF-IDF embeddings. NB was implemented using the Scikit-learn version 1.7.1, and XG was implemented using the XGBoost library version 3.0.4.

For DL models, a dense neural network (DNN), a convolutional neural network (CNN), a gated recurrent neural network (GRU), and a long short-term memory (LSTM) network were constructed for comparison using TensorFlow (Abadi, 2015). For embedding the text data, we used the same DL models but varied the embeddings across four different embeddings: TF-IDF, Word2Vec, GLoVe, and TensorFlow's adaptive (TFa) embeddings. Thus, for each DL model, there were four results providing 16 different conditions (four DL models x four embeddings). This was done to evaluate the extent to which both DL model and token embeddings had any identifiable effect on results.

Data was analyzed using a cross-validation approach using training data for the model update and a validation data set for evaluating results across 100 epochs. We used an early stopping methodology with a look back of 10 epochs when the loss function of the validation data stopped decreasing. The activation function for all models was the rectified linear unit (relu), the Adam optimizer was used with learning set at 0.001, and the softmax function for output. The best model was saved as the final model and then applied to the testing data.

The CNN model used 128 filters with a kernel size of 5 and global max pooling. The GRU and LSTM models used 128 units with dropout being set to .02 and recurrent dropout set at 0.2. The DNN used 128 hidden layers. For the transformer models, we used BERT base uncased, BioBERT, and DeBERTa.

Evaluation metrics used on the classification results were accuracy, precision, recall, and F1 (Dalianis, 2018). To evaluate the similarity of text generated by the GenAI process between the two LLMs, the cosine similarity (Dalianis, 2018) was used.

3 Results

All results were based on the use of the test data and the best model trained on the training and validation data. Of the baseline models, the XGBoost (XG) outperformed the Naïve Bayes (NB) across all metrics, see Table 2.

	Metrics			
Models	***Accuracy***	***Precision***	***Recall***	***F1***
NB	0.60	0.70	0.60	0.57
XG	0.73	0.73	0.73	0.73
CNN	0.29	0.33	0.29	0.28
GRU	0.26	0.27	0.26	0.23
LSTM	0.41	0.35	0.41	0.37
DNN	0.77	0.77	0.77	0.77
BERT	0.83	0.83	0.83	0.83
BioBERT	0.83	0.83	0.83	0.83
DeBERTa	0.83	0.83	0.83	0.83

Table 2 Text Classification Metrics for Models

For the DL models, there was no appreciable difference across all of the embeddings except that the TFa was as good or better. Therefore, the results in Table 2 are reported for all of the models using the TFa embeddings. Of the DL models, only the DNN outperformed XG on all metrics with values around .77. Of interest was that the CNN and GRU models had one thing in common: high training accuracy (>90%) but poor generalization (<30% on validation and testing data). These results appeared to indicate significant overfitting. Future research should look at models with more hidden layers and a dropout regularization method to see if this improves the overfitting.

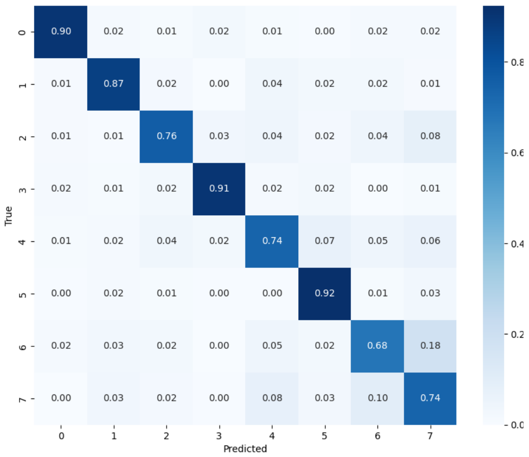


Figure 1 Cross-classification Matrix of True Domain Classification and Predicted Domain from BERT-uncased Model

The BERT family of models showed the best results across all metrics; however, there was not much separation between the models. These models accurately classified around 83% of the items. As the models were so similar, the BERT uncased model results were used for the prediction of the GenAI items.

Results of the cross-classification accuracy are provided in Figure 1 and suggested good predictions for most domains. Results were then dichotomized into correct classification = 1 and incorrect = 0. This resulted in ROC AUC = .81 and Youden's J at or above .83 suggesting an optimal trade-off in classification error (Pepe, 2004).

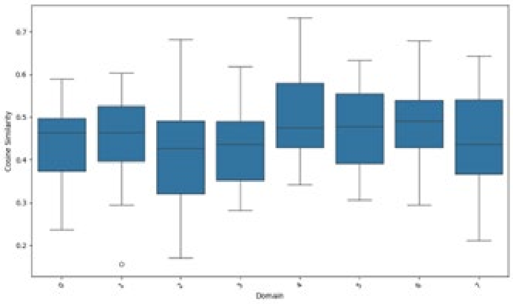


Figure 3 Box Plot of Cosine Similarity across Text Blueprint Domains

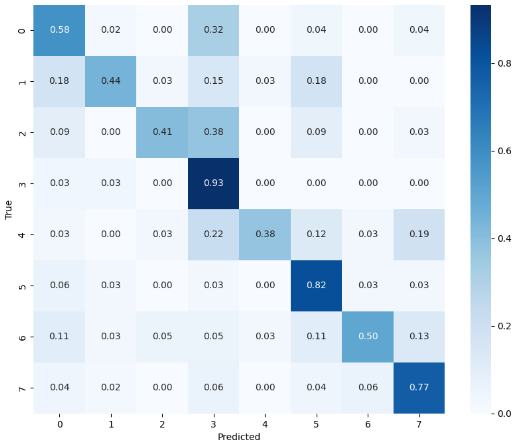


Figure 2 Cross-classification Matrix of True Domain Classification and Predicted Domain on the GenAI items from BERT-uncased Model

Item generation (GenAI) results of the 298 LLM-generated items had an average cosine similarity index between paired task statements between the models of .44 (sd = .11), see Figure 2 for box-and-whisker plot for the values by domain index. Using the best BERT uncased model results suggested a reduction in classification accuracy to around 61% which was significantly lower when compared to the classification using human generated items. Figure 3 provides the cross-classification matrix of 'true' values which was the domain for which the task statement was subsumed and the 'predicted' values from the best performing BERT model.

4 Discussion

This research investigated the extent to which different text-based models could be used for classifying assessment items into content domains. The results were varied with the baseline models having 60–70% accuracy. Of the DL models, the DNN showed the best results with about 77% accuracy. There was no clear difference between text embedding approaches across the DL models. Suggesting that, at least for this current task, any of the embedding methods would be equivalently useful. However, this might not generalize to all assessment program content or scope. It is recommended to investigate which embeddings might be best for each program while researching and developing classification models for one's own data.

The BERT family had the best results with all metrics greater than 80%. However, the results dropped off when the best BERT model was applied to the GenAI items to around 60%. This will certainly be different for each program and for different context engineering approaches to generate items. This research used a very simple prompt with minimal context for generating the items. Future research could evaluate results across various prompts and context engineering strategies to help identify the best ones to use for the programs' distinct needs.

The utility of these results is that it suggests the potential to reduce the work of SMEs by up to 80% when tasked with coding newly written items. Additionally, this process could help with reviewing previously coded items to support on-going quality control of metadata. This type of classification accuracy has the potential to significantly reduce resource utilization on metadata coding for SMEs to focus on content development and reviews that utilize their unique expertise and domain understanding.

This research found that a cut-off of a probability of correct classification of .83 was a reasonable value to balance errors. We would encourage practitioners to utilize their own results to set the relative errors they would be willing to accept. Additionally, to ensure model validation, the SMEs should also systematically review a small percentage of the items in the neighborhood above the threshold. This way, the original model is being continually evaluated in case the model begins drifting.

Future research could expand this approach by evaluating more complex DL models. For the CNN models, applications of dropout regularization could reduce the overfitting.

Combining DL models into more complex models using the strengths of the different methods could be evaluated. Additionally, both the smaller transformer models and extending to proprietary models available like Grok, OpenAI, etc. could be promising. Finally, extension to other metadata would be useful to see if the results generalize to other important categories.

Overall, these results are encouraging. The high rate of classification accuracy has the potential to automate a time consuming and resource intensive aspect of item development. With the automation of these tasks, SMEs can focus on more relevant work to support a program's item development needs.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from https://www.tensorflow.org/

Anthropic. (2025). Claude 3.7 Sonnet [large language model]. Anthropic PBC. https://www.anthropic.com/news/claude-3-7-sonnet

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Dalianis, H. (2018). *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer Open.

Devlin, J., Chang, M-W, Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint: arxiv.org/abs/1810.04805

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M.

Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Lawrence Erlbaum Associates.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(1–3), 131–163. https://doi.org/10.1023/A:1007465528199

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3-18). Routledge.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint: https://arxiv.org/abs/2006.03654

Meta AI. (2025). LLaMA 4 Maverick (17B-128E) [Large language model]. Meta Platforms, Inc. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

National Council of State Boards of Nursing (NCSBN). (2023). *Next Generation NCLEX: NCLEX-RN Test Plan*. National Council of State Boards of Nursing. Chicago, IL.

National Council of State Boards of Nursing (NCSBN). (2025). *2024 RN practice analysis: Linking NCLEX-RN examination to practice*. National Council of State Boards of Nursing. Chicago, IL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Association for Computational Linguistics. https://aclanthology.org/D14-1162/

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Sammut, C., & Webb, G. I. (Eds.). (2011). TF–IDF. In Encyclopedia of Machine Learning (pp. 986–987). Springer. https://doi.org/10.1007/978-0-387-30164-8_832

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097

Comparison of AI and Human Scoring on A Visual Arts Assessment

Ning Jiang, Yue Huang & Jie Chen
Measurement Incorporated
Correspondence: jchen@measinc.com

Abstract

This study investigates the comparability and reliability of scores generated by Artificial Intelligence, specifically a large language model, GPT-4, against scores assigned by trained human raters on a visual arts assessment. Two types of performance tasks, art writing and drawing, were selected from the South Carolina Arts Assessment Program. Responses from 358 fourth-grade students to Task 1 and from 190 students to Task 2 were evaluated independently by both GPT-4 and trained human raters. Both exact and adjacent agreement rates, as well as the Quadratic Weighted Kappa, were examined by task between the two human raters and for GPT-4 versus the first human rater. Additionally, these statistics were compared across tasks to explore whether task characteristics (i.e., text-based vs. drawing-based) contributed to differences in rater agreement. The findings highlight that 1) GPT-4 is more lenient and consistent in grading than human raters for both tasks; 2) the agreement between the human rater and GPT-4 is slightly lower than that between two human raters; and 3) human-GPT-4 scoring agreement remains consistent across visual arts performance tasks. These findings highlight the potential and limitations of using LLMs in arts-based assessment contexts.

1 Introduction

Artificial intelligence (AI) has become increasingly prominent in educational assessment, offering scalable and efficient solutions for scoring student responses. Much of the early work in this area has centered on text-based automated essay scoring (AES), where machine learning and natural language processing techniques have been used to replicate human scoring (Lim et al., 2021). The development of large language models (LLMs), such as OpenAI's GPT-4, Meta's LLaMA, and Google's Gemini, marks a new phase in AI-driven scoring. These models demonstrate advanced language understanding, generative abilities, and even processing both text and image inputs. Consequently, recent studies have explored their application in AES tasks, highlighting their potential for providing consistent and timely scoring. However, a substantial portion of this research remains focused on traditional writing tasks. There is still limited understanding of how LLMs perform in evaluating more complex responses, such as those required in visual and performing arts assessments. Filling this gap is especially important given the subjective and interpretive nature of art-based responses.

In visual arts education, assessing students' abilities to interpret, critique, and create artwork is an essential part of the learning process (Eisner, 1999). Performance tasks in this area often involve written critiques and creative visual outputs, which demand a nuanced understanding of artistic principles, expressive intent, and contextual interpretation. These tasks require subjective judgment and domain-specific expertise, making them challenging to score consistently—even among trained human raters (Perlman, 2003).

While AI technologies have shown promise in cognitive assessments, their application in art assessments remains limited. The rapid development of LLMs introduces new possibilities for scoring complex responses that combine text and imagery. AI scoring systems offer potential benefits such as efficiency, scalability, and reduced rater fatigue (Vetrivel, et

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 147–154
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

al., 2025). However, concerns persist about their reliability and validity in domains that rely on interpretive, aesthetic, or culturally embedded judgments (Clauser et al., 2014). This study tackles this gap by evaluating the use of GPT-4 to score fourth-grade students' visual arts performance tasks and comparing its performance to that of trained human raters.

By comparing AI-generated scores with those assigned by trained human raters, the study seeks to evaluate the reliability of AI-assisted scores in the context of art education. This research is guided by the following questions:

1. How do scores given by GPT-4 compare to those given by trained human raters on a text-based visual arts performance task?
2. How do scores given by GPT-4 compare to those given by trained human raters on a drawing-based visual arts performance task?
3. How does scoring agreement between GPT-4 and trained human raters differ across tasks (i.e., text- vs. drawing-based tasks)?

2 Data

2.1 Instrument

The South Carolina Arts Assessment Program (SCAAP) was used to evaluate fourth-grade students' visual arts achievement during the 2015–2016 school year. Designed to align with the South Carolina Academic Standards for the Visual and Performing Arts, SCAAP delivers technically sound assessments to students enrolled in schools funded through the Distinguished Arts Program. The assessment consists of 45 multiple-choice items and two performance tasks aimed at measuring students’ understanding and application of visual arts concepts.

This study focuses on the two visual arts performance tasks. Each student received a test booklet containing written instructions and designated spaces for completing the tasks. The tasks were administered in group settings by trained test administrators following a standardized manual. Although untimed, each task typically took students at least 30 minutes to complete.

Performance Task 1 (hereafter referred to as Task 1) was designed to assess two key standards:

- Standard 2 – Using structures and functions in visual arts
- Standard 5 – Interpreting works of visual arts

In this task, students were asked to select four vocabulary terms from a provided word bank and write a paragraph for each term, explaining how it applied to the given artwork. Each paragraph was expected to consist of several descriptive sentences that demonstrated the student's understanding of the selected art concepts.

Performance Task 2 (hereafter referred to as Task 2) required students to complete a drawing based on a given prompt and was designed to assess two key standards:

- Standard 1 – Creating art
- Standard 2 – Using structures and functions in visual arts

2.2 Scoring Rubric

Responses to both tasks were scored holistically using a rubric with five proficiency levels, ranging from 0 to 4. Each level clearly described the degree of proficiency demonstrated in students’ responses. Raters could assign augmentation scores using “+” or “–” to adjust the base score by 0.33 points. To compute quadratic weighted kappa (QWK), augmentation scores were rounded to the nearest integers, as QWK analysis requires categorical or ordinal-level data.

2.3 Participants

Three student samples were used to address the three research questions, with demographic details provided in Table 1.

- Sample 1 comprised 358 students whose Task 1 responses were double-scored and was used to address Research Question 1 (RQ 1).
- Sample 2 comprised 190 students whose Task 2 responses were double-scored and was used to address Research Question 2 (RQ 2).
- Sample 3 comprised 166 students whose responses to both tasks were double-scored and was used to address Research Question 3 (RQ 3).

3 Methods

	Gender			Race			
Sample	Male (%)	Female (%)	Missing (%)	Black (%)	White (%)	Other (%)	Missing (%)
1	150 (42)	188 (53)	20 (6)	123 (34)	182 (51)	9 (3)	44 (12)
2	73 (38)	101 (53)	16 (8)	71 (37)	85 (45)	18 (9)	16 (8)
3	67 (40)	86 (52)	13 (8)	62 (37)	74 (45)	17 (10)	13 (8)

Note. Percentages may not sum to 100% due to rounding.

Table 1: Student Demographic Summary by Sample.

3.1 Human Scoring

Student responses were scanned and saved as JPEG files and uploaded to the SCAAP web-based rating system for remote scoring. All raters were trained and required to pass a qualifying test before scoring the student work. In 2016, five trained raters were employed to score responses to Tasks 1 and 2. Considering the change of Task 1 during the 2015-2016 school year, all responses to Task 1 were double-rated (i.e., scored by two raters), and about 50% of responses to Task 2 were double-rated. In instances of non-adjacent scores, an expert rater was brought in to provide a third score. For double-rated responses, the final score was calculated as the average of the two raters' scores. If an expert rater's score was needed due to disagreement, that score would be used as the final score instead. However, in the current study, all samples consisted only of double-scored responses, without applying score averaging adjustments.

3.2 GPT-4 Scoring

GPT-4 was used to assess the same set of students' responses to performance tasks. GPT-4 is OpenAI's latest multimodal model that can process and generate text, images, and audio. It offers fast performance, improved reasoning, and seamless handling of multiple input types. The version adopted for scoring in this study is gpt-4o-2024-11-20. All de-identified student responses were scanned and input to LLMs through the model API interaction in Python.

Specifically, we designed a prompt engineering framework to simulate the human scoring process. Each prompt includes the following components: a description of the context, a description of the performance task description and scoring rubric identical to those used in training human raters, the chain-of-thought prompt for scoring the student's response. Every individual response was scored using the same prompt framework to control possible drifts in LLM output.

3.3 Data Analysis

For all samples, descriptive statistics, exact and adjacent agreement rates, and the QWK were calculated by task for Rater 1 (R1) vs. Rater 2 (R2) and R1 vs. GPT-4. R1's ratings served as the reference. As a chance-corrected agreement measure that weighs disagreements based on their severity, QWK offers further insight into the extent of agreement and disagreement between the raters (Cohen, 1968). To address RQ 3, the statistics of rater agreement and the QWK regarding the two different tasks were compared to investigate the impact of task type on the agreement between human and GPT-4 scoring.

In addition, confusion matrices were constructed for each rater pair (R1 vs. R2 and R1 vs. GPT-4) by task and research question to analyze the frequency with which rating categories from one rater corresponded to those of the other. The QWK provides a single summary value that adjusts for chance agreement and the severity of disagreements, while the confusion matrix helps identify where and how raters disagree.

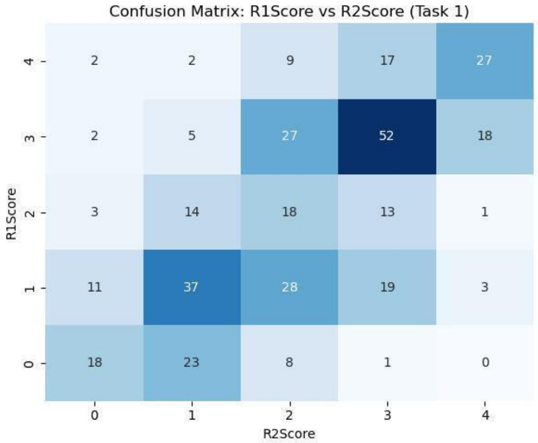
4 Results

Table 2 presents summary statistics of scores assigned by both human raters and GPT-4 for each task, using Sample 1 for Task 1 and Sample 2 for Task 2. Table 4 provides the same statistics based on the common sample (Sample 3). Table 3 reports agreement rates between R1 and R2 and between R1 and GPT-4 by task, while Table 5 presents the corresponding agreement rates based on the common sample. Finally, Table 6 summarizes the differences in agreement levels between Task 1 and Task 2. Figures 1–4 display confusion matrices for Task 1 and Task 2, comparing R1 with R2 and GPT-4 on the 0–4 scale. Figures 5–8 repeat these comparisons using the RQ 3 subsample.

4.1 RQ 1: Task 1 – Art Writing

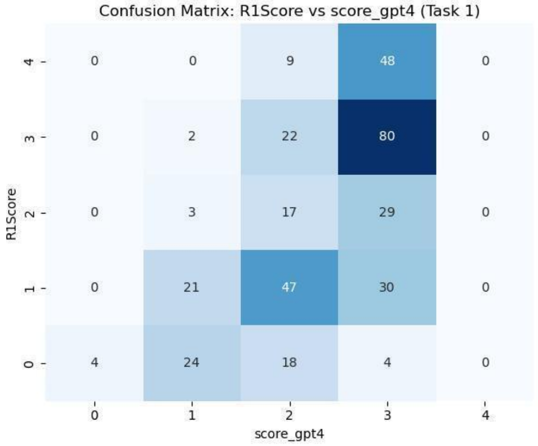
Results in Table 2 show that GPT-4 produced slightly higher average scores than both human raters and with less variability on Task 1 (N=358). In terms of assigning scores, GPT-4 is the most lenient and consistent (M = 2.37, SD = 0.76), while R1 is the harshest (M = 2.08, SD = 1.31).

For Task 1, the exact agreement between human raters was 34%, with an adjacent agreement of 48%, totaling 82% for within-one-point agreement. Comparably, the agreement between GPT-4 and R1 yielded to 28% exact and 54% adjacent, also totaling 82% (see Table 3).



R1Score \ R2Score	0	1	2	3	4
4	2	2	9	17	27
3	2	5	27	52	18
2	3	14	18	13	1
1	11	37	28	19	3
0	18	23	8	1	0

Figure 1: RQ1: Confusion Matrix for Task 1 (Human Rater 1 vs Human Rater 2).



R1Score \ score_gpt4	0	1	2	3	4
4	0	0	9	48	0
3	0	2	22	80	0
2	0	3	17	29	0
1	0	21	47	30	0
0	4	24	18	4	0

Figure 2: RQ1: Confusion Matrix for Task 1 (Human Rater 1 vs GPT-4).

Task (N)	Rater	Mean	SD	Min	Max
1 (358)	Rater 1	2.08	1.31	0	4
	Rater 2	2.13	1.20	0	4
	GPT-4	2.37	0.76	0	3
2 (190)	Rater 1	2.19	0.77	0	4
	Rater 2	2.35	0.79	0	4
	GPT-4	2.59	0.55	1	4

Table 2: Descriptive Statistics of Scores by Task and Individual Rater.

However, the QWK coefficient for R1 vs. GPT-4 (0.49, indicating moderate agreement) was smaller than that for human-human comparison (0.62, indicating substantial agreement), suggesting slightly lower consistency between human- and machine-scoring.

Results of the confusion matrix can visually identify exact/adjacent agreement, observe patterns of disagreement, and detect systematic bias. Figure 1 displays confusion matrix comparing R1 with R2 and Figure 2 displays R1 and GPT-4 on the 0–4 scoring scale. The R1–R2 matrix shows relatively consistent scoring across the full score range, with disagreements mostly concentrated in adjacent categories, suggesting moderate to strong alignment between human raters. In contrast, the R1–GPT-4 matrix shows high agreement primarily around score 3, but GPT-4 avoids assigning extreme scores (0 and 4), resulting in compressed scoring toward the middle. This indicates that while GPT-4 is more consistent with R1 at mid-range scores than R1 and R2 agreement, it demonstrates a conservative bias, especially at the scoring extremes.

Task	Rater Pair	Exact	Adjacent	QWK
1	Rater 1 vs Rater 2	0.34	0.48	0.62
	Rater 1 vs GPT-4	0.28	0.54	0.49
2	Rater 1 vs Rater 2	0.51	0.46	0.61
	Rater 1 vs GPT-4	0.43	0.51	0.44

Table 3: Rater Agreement and Interrater Reliability by Task and Rater Pair.

4.2 RQ 2: Task 2 – Drawing

For the drawing-based performance task (N=190), GPT-4 again yielded slightly higher average scores and a narrower distribution (M = 2.59, SD = 0.55) (see Table 2). As shown in Table 3, GPT-4 exhibited a slightly lower rate of exact agreement with the human rater compared to the human-human agreement (43% vs. 51%). The adjacent + exact agreement remained high for both pairs (97% for the R1-R2 pair vs. 94% for the R1-GPT-4 pair), but the QWK coefficient (0.44) for the R1-GPT-4 pair again fell short of the human-human benchmark (0.61).

Figure 3 and Figure 4 display confusion matrices comparing R1's scores with those from R2 and GPT-4, respectively, on Task 2. The R1–R2 matrix shows strong agreement around score 2 (n=77) but with notable dispersion at higher score levels; for example, many R1 scores of 3 were rated as 2 (n = 15) or 4 (n = 8) by R2. This indicates moderate agreement with some variability, especially at the upper end. In contrast, the R1–GPT-4 matrix shows high agreement for scores of 2 and 3 (n = 51 and 48 respectively), with fewer deviations but a visible tendency to avoid the extremes—GPT-4 rarely assigns scores of 0 or 4. These patterns suggest that while both R2 and GPT-4 align with R1 in the mid-range, GPT-4 displays a narrower scoring range with a conservative bias.

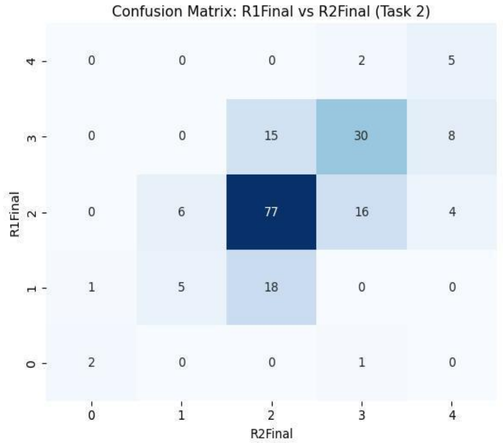


Figure 3: RQ2: Confusion Matrix for Task 2 (Human Rater 1 vs Human Rater 2).

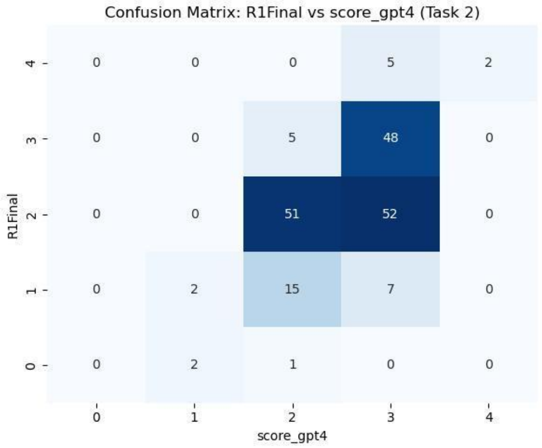


Figure 4: RQ2: Confusion Matrix for Task 2 (Human Rater 1 vs GPT-4).

4.3 RQ 3: Difference between Task 1 and Task 2

To evaluate differences in scoring agreement across task types, we compared R1–R2 and R1–GPT-4 agreement statistics for both Task 1 and Task 2, using the common sample (N = 166). As shown in Table 4, GPT-4 is again the most lenient and consistent (M = 2.39, SD = 0.73 for Task 1; M = 2.61, SD = 0.54 for Task 2), while R1 is the harshest (M = 2.07, SD = 1.28 Task 1; M = 2.22, SD = 0.77 for Task 2). Additionally, on average, both the human raters and GPT-4 assigned higher scores to responses for Task 2 than to those for Task 1.

Task	Rater	Mean	SD	Min	Max
1	Human Rater 1	2.07	1.28	0	4
	Human Rater 2	2.20	1.22	0	4
	GPT-4	2.39	0.73	0	3
2	Human Rater 1	2.22	0.77	0	4
	Human Rater 2	2.38	0.77	0	4
	GPT-4	2.61	0.54	1	4

Table 4: Descriptive Statistics of Scores by Task and Individual Rater (Sample 3: N = 166).

Task	Rater Pair	Exact	Adjacent	QWK
1	Rater 1 vs Rater 2	0.34	0.43	0.56
	Rater 1 vs GPT-4	0.27	0.52	0.41
2	Rater 1 vs Rater 2	0.49	0.46	0.57
	Rater 1 vs GPT-4	0.43	0.51	0.43

Table 5: Rater Agreement and Interrater Reliability by Task and Rater Pair (Sample 3: N = 166).

As shown in Table 5, for both tasks, the exact agreement between R1 and GPT-4 was lower than that between human raters (27% vs. 34% for Task 1; 43% vs. 49% for Task 2), while the adjacent agreement between R1 and GPT-4 was higher than that between human raters (52% vs. 43% for Task 1; 51% vs. 46% for Task 2). Overall, the exact and adjacent agreement rates were higher for Task 2 than Task 1, with R1–R2 increasing from 77% to 95% and R1–GPT-4 from 79% to 94%. In contrast, QWK values remained relatively stable across tasks, rising only slightly from 0.56 for Task1 to 0.57 for Task 2 for R1–R2 and from 0.41 to 0.43 for R1–GPT-4. The stable QWK coefficients for both the R1-R2 and the R1-GPT-4 pairs across tasks suggest overall reliability in scoring despite differences in task type.

Difference across tasks	Exact	Adjacent	QWK
Rater 1 vs. Rater 2	-0.15	-0.03	-0.01
Rater 1 vs. GPT-4	-0.16	0.01	-0.02

Table 6: Difference in Rater Agreement and Interrater Reliability Across Tasks (N = 166).

As observed in Table 6, for the R1–R2 pair from Task 1 to Task 2, the exact agreement increased by 15%, and the adjacent agreement increased by 3%. However, the comparison of QWK statistics indicates nearly no difference in agreement between the two tasks (1%). For the R1–GPT-4 pair from Task 1 to Task 2, the exact agreement increased by 16, and the adjacent agreement decreased by 1%. The comparison of QWK statistics shows a slight increase in agreement (2%) from Task 1 to Task. While the QWK coefficients remained largely stable, the increases in exact agreement indicate more consistent scoring on the drawing task.

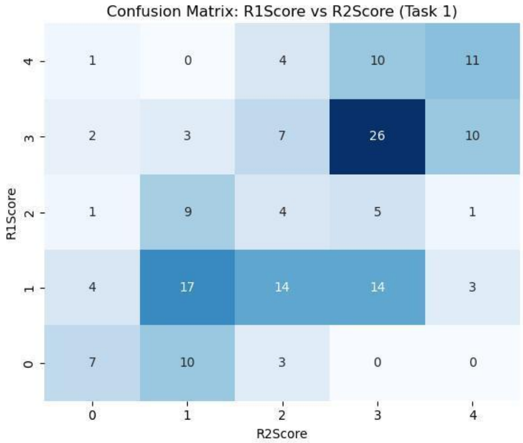


Figure 5: RQ3: Confusion Matrix for Task 1 (Human Rater 1 vs Human Rater 2).

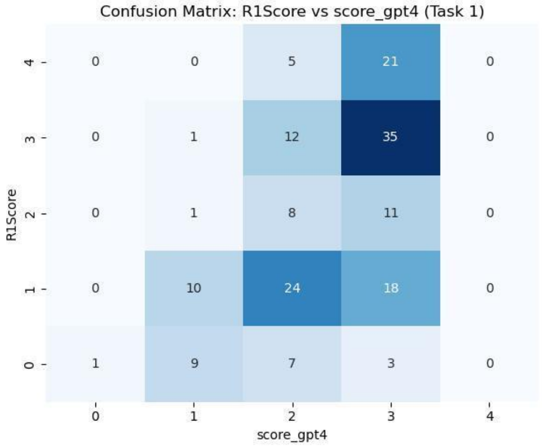


Figure 6: RQ3: Confusion Matrix for Task 1 (Human Rater 1 vs GPT-4).

We also generated confusion matrices for all scoring comparisons and tasks based on Sample 3, comparing R1's scores with those from R2 and GPT-4 on Task 1 (see Figures 5 and 6). The patterns are similar to those results in RQ 1. The R1–R2 matrix indicates moderate agreement, with the highest concentration along the diagonal, particularly at score 3 (n=26). However, off-diagonal cells reveal notable adjacent and distant mismatches, such as R1 assigning a point of 4 when R2 gave a 1 or 2, suggesting some inconsistency at higher scores. In contrast, the R1–GPT-4 matrix shows a narrower distribution, with GPT-4 scores concentrated at 2 and 3 and no scores assigned at the extremes (0 or 4). Although GPT-4 shows alignment with R1 in the mid-range (e.g., R1 = 3 most often matched GPT-4 = 3), it avoids the use of the full score scale, pointing to a compressed and conservative scoring tendency.

Figures 7 and 8 show confusion matrices comparing R1's scores with those from R2 and GPT-4 on Task 2 using the subsample for RQ 3. The patterns are similar to those results in RQ 2. The R1–R2 matrix shows strong agreement at score 2 (n = 62) and moderate alignment at score 3 (n = 26), though some off-diagonal variation appears—particularly when R1 assigned a 3 but R2 gave a 2 or 4, suggesting some upper-end disagreement. The R1–GPT-4 matrix reveals a narrower scoring pattern, with GPT-4 clustering scores tightly around 2 and 3, and avoiding extreme values (0 and 4). While GPT-4 shows high agreement with R1 in the mid-range (e.g., R1 = 2 and GPT-4 = 2 or 3), its reluctance to assign the highest score may reflect a conservative or compressed scoring pattern.

There are some noticeable differences between Task 1 and Task 2 in both R1–R2 and R1–GPT-4 scoring patterns:

- Human-human consistency: in both tasks, R1 and R2 show the strongest agreement at score 2, but Task 2 generally shows tighter clustering along the diagonal—especially at scores 2 and 3, suggesting slightly higher inter-rater consistency than Task 1. In Task 1, disagreements are more spread out, including more extreme mismatches (e.g., R1 = 4 vs R2 = 0 or 1), whereas in Task 2, disagreements tend to stay within adjacent scores.
- GPT-4 scoring behavior: GPT-4 is conservative across both tasks, avoiding extreme scores, but this effect is more pronounced in Task 2. In Task 2, GPT-4

rarely assigns a score of 4 and leans heavily toward scores of 2 and 3, even when R1 gave higher scores. In Task 1, while GPT-4 still avoids extremes, the spread is slightly broader, especially around score 3.

In general, GPT-4 is capable of approximating human scoring with high adjacent agreement (>0.5) in both text-based and drawing-based tasks. However, the QWK statistics for the R1–GPT-4 pair indicate that GPT-4 does not replicate the full scoring pattern of human raters, particularly at extremes (i.e., 0 or 4). The agreement patterns between tasks are similar, but Task 2 (drawing) shows slightly stronger alignment in both adjacent and overall agreement, possibly due to fewer extreme scores in that task.

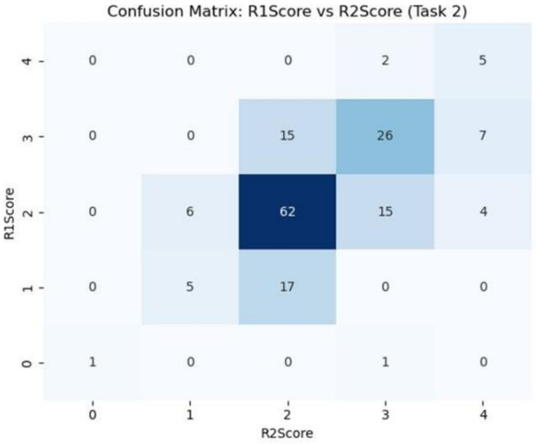


Figure 7: RQ3: Confusion Matrix for Task 2 (Human Rater 1 vs Human Rater 2).

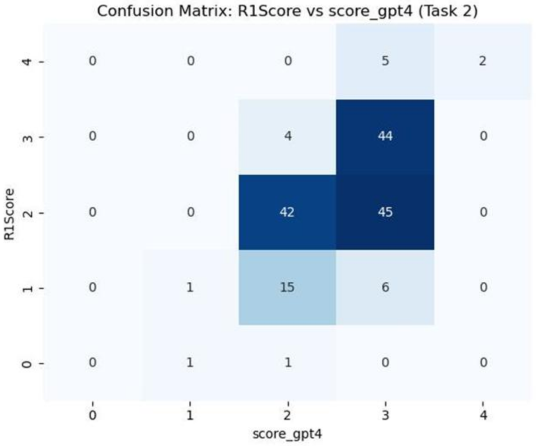


Figure 8: RQ3: Confusion Matrix for Task 2 (Human Rater 1 vs GPT-4).

5 Discussion

In the text-based task, GPT-4 achieved comparable adjacent agreement with human raters but demonstrated a narrower scoring range and a reluctance to assign extreme scores. This centralizing tendency results in lower QWK values, reflecting a more conservative scoring pattern. In the drawing-based task, although adjacent agreement remained high, GPT-4's performance was slightly weaker, particularly in its interpretation of complex visual elements such as depth, composition, and emotional nuance.

Furthermore, the observed patterns in the confusion matrices suggest that while human raters showed moderate consistency across the full score range, discrepancies, particularly at higher scores, highlighted the subjective nature of human scoring. In contrast, GPT-4 demonstrated strong alignment with human raters in mid-range scores but consistently avoided extreme ratings, indicating a conservative bias. These findings imply that while GPT-4 may be a reliable tool for scoring typical responses, caution is needed when using it to evaluate very high- or low-quality work, as it may underrepresent performance extremes and affect decisions tied to those score ranges.

A consistent trend across both tasks is that GPT-4 tended to avoid assigning the lowest and highest score points, which has implications for high-stakes assessments where performance extremes are critical. Additionally, human raters showed greater variability in scoring, especially at higher performance levels, which may reflect the inherent subjectivity in assessing creative work—a dimension GPT-4 is currently limited in replicating.

These findings suggest that while GPT-4 can serve as a reliable supplemental tool in scoring student work, it should not yet be considered a full substitute for human judgment in art assessment. The performance of GPT-4 was task-sensitive, further emphasizing the need for content-specific prompt tuning and calibration.

6 Limitations

While this study provides valuable insights into the comparability of GPT-4 and human scoring in visual arts assessments, several limitations should be acknowledged. First, the analysis focused on fourth-grade student responses from a single assessment year and context (SCAAP 2016), which may limit the generalizability of the findings to other grade levels, content areas, and assessment frameworks. Second, the scoring outcomes reflect a single implementation of GPT-4 using a specific prompt engineering approach; different prompt designs or fine-tuned model configurations may produce different results. Third, although GPT-4 was instructed to incorporate augmentation scores (e.g., +0.33 or –0.33), it did not apply this rule during scoring. This unexpected behavior highlights a potentially important limitation in the model's ability to consistently follow nuanced scoring rules, and future research should investigate the causes and implications of this issue. Finally, human-human agreement was not high, and only R1 was randomly selected as the baseline rater for all comparisons, which may constrain generalizability. One possible reason for the moderate human-human agreement is that Task 1 was new and first administered in 2016. In addition, Task 2's prompt was changed in 2015, and 2016 was only the second year it was used. Therefore, the raters were not yet sufficiently familiar with the new tasks and the rubric. However, this study is strengthened by the inclusion of cross-comparisons across two tasks and repeated analyses on multiple subsamples, thereby providing a multidimensional assessment of scoring consistency from diverse perspectives.

7 Future Research

Future research should explore the reasoning behind GPT-4's scoring decisions through qualitative content analysis. Specifically, analyzing GPT-4's rationale in comparison to the scoring rubric and underlying pedagogical goals may help illuminate how well the model interprets key assessment criteria. Additionally, it would be valuable to examine why GPT-4 consistently avoided assigning extreme scores—whether due to probabilistic constraints in its language modeling, uncertainty in interpreting creative responses, or an overly cautious alignment with prompt wording. Understanding these issues may help refine prompt engineering or model tuning for better alignment with human evaluative standards.

References

Eisner, E. W. (1999). The national assessment in the visual arts. Arts Education Policy Review, 100(6), 16-20.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2014). Validity issues for performance-based tests scored with computer-automated scoring systems. In Advances in Computerized Scoring of Complex Item Formats (pp. 413-432). Routledge.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256.

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A Comprehensive Review of Automated Essay Scoring (AES) Research and Development. Pertanika Journal of Science and Technology, 29(3). https://doi.org/10.47836/pjst.29.3.27.

Perlman, C. C. (2003). Performance Assessment: Designing Appropriate Performance Tasks and Scoring Rubrics.

Vetrivel, S. C., Arun, V. P., Ambikapathi, R., & Saravanan, T. P. (2025). Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments. In *Adopting Artificial Intelligence Tools in Higher Education* (pp. 21-61). CRC Press.

Explainable Writing Scores via Fine-grained, LLM-Generated Features

James V. Bruno
Pearson
US School Assessment Technology
`jimmy.bruno@pearson.com`

Lee Becker
Pearson
US School Assessment Technology
`lee.becker@pearson.com`

Abstract

Transfomer-based models like BERT have increasingly been employed for automated essay scoring, as their high-dimensional representations of text are effective at capturing complex patterns in language. However, transformer-based representations are opaque and difficult to trace to the underlying human-defined constructs being assessed. This paper investigates the ability of LLMs to generate scores according to a rubric constructed from academic standards and evaluates the utility of these scores as features in a supervised regression model. We show that this produces a model that is reliable, construct-relevant, and interpretable. We evaluate this approach on six narrative writing items and find that, even with only 5 features, models can achieve QWKs exceeding 0.8, while also giving a concise and interpretable score explanation.

1 Background

Prior to advances in deep learning, the prevailing approach for Automated Essay Scoring (AES), relied on pairing supervised machine learning (ML) with a set of manually-crafted features (Attali and Burstein, 2006) that aimed for construct relevance. Feature engineering consisted of extracting linguistic phenomena which could serve as proxies for the underlying construct or assessed skill. For example, type-token ratio was used to capture vocabulary richness and semantic similarity measures approximated human ratings of essay cohesion (Graesser et al., 2004). In some cases, features may come from other models trained to predict a subtrait score (Somasundaran et al., 2018).

Advancements in NLP and ML have rapidly evolved the state-of-the art in automated essay scoring (AES). The shift toward dense language representations including semantic vectors (Deerwester et al., 1990), word embeddings (Mikolov et al., 2013a,b), and contextual embeddings (Peters et al., 2018) have yielded steady gains in AES performance (Foltz et al., 1999; Alikaniotis et al., 2016), typically measured using metrics like Quadratic Weighted Kappa (QWK). Transformers (Vaswani et al., 2017) and especially variants of BERT (Devlin et al., 2019) are now considered the de-facto approach for training AES systems (Mayfield and Black, 2020; Wang et al., 2022; Wang, 2024; Elmassry et al., 2025). However, these approaches are complex with features that are not directly interpretable and which number in the hundreds or thousands. Modern, deep-learning AES systems are effectively “black-box” solutions.

A growing body of research has applied explainable AI (xAI) to AES, including approaches such as attention visualization (Yang et al., 2020), multiple instance learning (Hellman et al., 2020) and post hoc explanation methods like LIME (Ribeiro et al., 2016), which surface links between regions of the text and model outputs. However, interpreting these explanations often requires subjective inference to connect model decisions to the constructs being assessed, and may lack direct construct relevance. The capacity for reasoning exhibited by Large Language Models (LLMs) presents new possibilities for explainability. LLMs can be prompted to generate auxiliary information such as rationalization of score (Li et al., 2023) or corresponding feedback (Stahl et al., 2024). While impressive, querying LLMs to provide justifications for their scoring decisions introduces the risk of self-referential explanations.

This work approaches explainability through the lens of subtrait scoring wherein the scored construct is broken down into sub-components with their own scores (Andrade-Lotero et al., 2025). Our framework is most similar to TRATES (Eltanbouly et al., 2025) which predicts rubric elements via LLM generation. Unlike TRATES, we limit the features of our models to only construct-relevant subtrait scores. By pairing LLM-generated subtrait

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 155–165
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

scores with linear regression, we can not only produce ”dead simple” explainability, but also help build trust in the use of LLMs for essay scoring.

2 Aims

The primary aim of this study is to explore how using LLM-outputs as input features to train automated scoring models provides a straightforward path toward interpretable scores. Specifically, we explore the ability of LLMs to produce subtrait scores aligned to components of academic standards defining grade-level expectations, such as those set by Common Core State Standards Initiative (2010). We build, evaluate, and inspect simple linear models using the LLM-produced subtrait scores as features.

This work is part of an overarching goal to develop a collection of models that can assess subtraits and/or skills in support of a wide variety of writing items and rubrics (Andrade-Lotero et al., 2025). In this work we focus on assessing Common Core standards because the standards are decomposed into elements that align well with our notion of subtrait; however, nothing about the approach we illustrate here is limited to using standards.

Our specific research questions are as follows:

- How can we leverage academic standards to generate subtrait scores via an LLM?
- How does the performance of models trained with LLM-generated subtrait scores as features compare to operational models?
- Is an explainable linear regression model with only a handful of subtrait features operationally viable?
- Can the use of a simple, transparent linear regression model enhance the interpretability and trustworthiness of the LLM features?

3 Data

The experiment dataset consists of 6 eighth grade writing prompts administered as part of state-run, year-end summative assessments. Written responses and corresponding human scores come from the train and test data used to build and evaluate the operational scoring models. Responses flagged with codes such as blank, gibberish, passage copy or off-topic are excluded from this dataset. The scoring process is such that there are a minimum of 2 ratings per response in the model-building data set, with a third resolution rating as needed. We model the final score, that is, the score assigned as the end-result of the human scoring process for the item. The items in our experiment have high human agreement and extremely strong operationally deployed models, as shown in Table 1.

	Human	Deployed model
Prompt 1	0.940	0.918
Prompt 2	0.914	0.899
Prompt 3	0.897	0.893
Prompt 4	0.912	0.903
Prompt 5	0.888	0.886
Prompt 6	0.936	0.934

Table 1: Quadratic weighted kappa representing human-human agreement and human-machine agreement for operationally deployed models.

We focus on narrative writing items, as scoring in this genre is often perceived as subjective and multi-dimensional. Explainable models add transparency to this subjective process by linking assessment of narrative elements to the score. Additionally, narrative elements are not well captured by surface level features like n-grams or word count. This presents an opportunity to highlight how the deeper semantics of LLMs can flexibly accommodate a wide variety of subtraits.

We aim to take a uniform sample across score points and use 66% of the data for training and the remainder for testing, for a total of 1125 responses per score point. However, responses at the highest score points were underrepresented in the population for some prompts, and in these cases a uniform distribution is not possible. The score distribution for the experimental dataset appears in Table 2. Additionally, it was not possible to extract subtrait features from for every response, as in some cases the LLM returned malformed JSON or raised content filters. The response counts of the final train-test partitions for the 3 LLMs we use in our experiments appear in Table 3.

4 Method

As our goal is explainability, we wish to build the simplest, most interpretable model possible using the outputs from the LLM. As discussed below, we construct an LLM query with scoring instructions and a rubric based on the Common Core standard for 8th-grade narrative writing (Common Core State Standards Initiative, 2010).

	Train						Test					
	0	1	2	3	4	Total	0	1	2	3	4	Total
Prompt 1	150	150	150	150	117	717	75	75	75	75	58	358
Prompt 2	150	150	150	150	51	651	75	75	75	75	25	325
Prompt 3	150	150	150	121	28	599	75	75	75	60	14	299
Prompt 4	150	150	150	150	68	668	75	75	75	75	35	335
Prompt 5	150	150	150	103	43	596	75	75	75	52	23	300
Prompt 6	150	150	150	148	150	748	75	75	75	74	75	374
Total	900	900	900	822	457	3979	450	450	450	411	230	1991

Table 2: Score distribution of the training and test sets.

	Original Sample		GPT-4o Mini		Llama3.1 8B		Gemma 7B	
	Train	Test	Train	Test	Train	Test	Train	Test
Prompt 1	717	358	690	343	717	358	692	347
Prompt 2	651	325	643	323	651	325	640	322
Prompt 3	599	299	589	292	599	299	582	284
Prompt 4	668	335	645	328	668	335	655	327
Prompt 5	596	300	583	298	596	300	577	290
Prompt 6	748	374	743	372	748	374	733	367
Total	3979	1991	3893	1956	3979	1991	3879	1937

Table 3: Number of responses in Train/Test partitions. It was not possible to obtain LLM-based features from GPT-4o Mini or Gemma 7B for all responses due to content-filters and the LLM returning improper JSON. Therefore, the samples for some of the models tested are non-identical.

The LLM provides feature values in the form of subtrait scores, which we use to train and evaluate highly-interpretable linear regression models. The models predict the operational score using LLM-generated subtrait scores as features. We do not have ground-truth annotations for the subtrait scores.

4.1 LLM-generated features

The feature space centers around "subtrait scoring", wherein finer-grained scores reflect performance on a facet of a larger trait. Following the approach detailed in Andrade-Lotero et al. (2025), we query an LLM to score an essay given a rubric. To maintain consistency between items, we use standards-based rubrics instead of the items' original trait rubric. As we aim to make an assessment grounded in the Common Core standard for the narrative genre, we construct a 5-point rubric from the standard elements of `CCSS.ELA-LITERACY.W.8.3`, that is, the Common Core standard for 8th grade narrative writing. The standard reads "Write narratives to develop real or imagined experiences or events using effective technique, relevant descriptive details, and well-structured event sequences." The standard is further decomposed into the 5 standard elements that appear in Table 4, which we treat as subtraits.

We use an LLM to construct a rubric from these standard elements. Specifically, we embed each of the standard elements into instructions to create criteria for 5 score points, query Claude Haiku 3.5 (Anthropic, 2024) with the instructions, and manually verify the result. We choose a score range from 0 to 4 to reflect the original range on which the responses were scored. Example system instructions to create rubrics appear in Figure 1 and an example rubric appears in Figure 2 in the Appendix.

To produce subtrait scores, we submit the response and the rubric to 3 LLMs. We choose OpenAI's GPT-4o-Mini (OpenAI, 2024) and also two open-source models of similar size: Gemma 7B (DeepMind, 2024), and Llama 3.1 8B (Meta, 2024).

The LLM query to produce subtrait scores includes an instruction to provide feedback. This likely has a positive effect on the output given that eliciting reasoning is known to improve LLM results (see for example, Huang and Chang, 2023); however, we set aside an examination of this effect

W.8.3.A	Engage and orient the reader by establishing a context and point of view and introducing a narrator and/or characters; organize an event sequence that unfolds naturally and logically.
W.8.3.B	Use narrative techniques, such as dialogue, pacing, description, and reflection, to develop experiences, events, and/or characters.
W.8.3.C	Use a variety of transition words, phrases, and clauses to convey sequence, signal shifts, and show the relationships among experiences and events.
W.8.3.D	Use precise words and phrases, relevant descriptive details, and sensory language to capture the action and convey experiences and events.
W.8.3.E	Provide a conclusion that follows from and reflects on the narrated experiences or events.

Table 4: 8th grade narrative writing standard elements.

for future work. The prompt appears in Figure 3 in the Appendix.

4.2 Model training, evaluation, and explanation.

As our overarching goal is intepretability, we train a non-negative ridge regression model with L2 regularization for each item-LLM model pair, using cross-validation within the training set to tune the regularization parameter. Ridge regression is used instead of simple linear regression to minimize the effects of multicollinearity between subtrait scores, and the coefficients are required to be positive for ease of interpretability.

Models are evaluated on quadratic weighted kappa (Cohen, 1968), and we examine the coefficients for each of the 5 subtraits for the purpose of explaining the models.

5 Results

The quadratic weighted kappa for the 6 prompts and the 3 LLMs appear in Table 5. All models were able to predict the human score, with average test QWKs of 0.81, 0.78, and 0.59 for GPT-4o-Mini, Llama 3.1 8B, and Gemma 7B, respectively. GPT-4o-Mini had the best overall performance, but we are particuarly encouraged that an open source model with few parameters such as LLama 3.1 8B is viable. We note that there is only a 3-point difference between GPT-4o-Mini and LLama 3.1 8B, compared to the 22-point difference between GPT-4o-Mini and Gemma 7B.

5.1 Explainability

With respect to explainability, the simplest approach with a non-negative linear model is to examine the relative weights, normalized to 1 to make them more intuitively interpretable. This tells us what percentage of the final score is due to each of subtrait scores from the LLM. As we have 6 prompts and 3 models per prompt, we present the means and standard deviations of the relative weights in Table 6. The relative weights for all the individual models appear in Table 8 in the Appendix.

The ability to inspect the relative weights is what we regard as the main benefit of this approach. We can see, for example, that the GPT and Llama models place more weight on establishing context, the use of narrative techniques, and the fluidity of transitions; and less weight on linguistic descriptiveness and the quality of the conclusion. We hypothesize that this is related to the models' superior performance, and we highlight that the weights can be subject to examination by a subject matter expert (SME) in writing who may not have a great deal of expertise in machine learning.

The weights of the Gemma 7B model are particularly illustrative. The model weights linguistic descriptiveness very heavily at almost half of the score, and the conclusion quality is hardly part of the model at all. This might raise validity concerns for an SME scrutinizing the model and could serve as an early and easily interpretable cautionary signal before moving forward with such a model.

5.2 LLM comparison

Gemma 7B's lack of parity with the other two models in terms of the predictiveness of its subtrait scores is striking. Furthermore, we note that the standard deviations of the per-prompt models are double those of GPT-4o-Mini and Llama 3.1 8B, as shown in Table 6. This suggests that the subtrait scores produced from Gemma 7B are less stable than those of the other two models.

	Train			Test		
	GPT-4o Mini	Llama3.1 8B	Gemma 7B	GPT-4o Mini	Llama3.1 8B	Gemma 7B
Prompt 1	0.802	0.786	0.629	0.843	0.799	0.654
Prompt 2	0.770	0.742	0.616	0.761	0.718	0.603
Prompt 3	0.815	0.781	0.468	0.852	0.791	0.467
Prompt 4	0.794	0.788	0.411	0.815	0.801	0.417
Prompt 5	0.767	0.734	0.593	0.775	0.782	0.617
Prompt 6	0.830	0.815	0.618	0.831	0.804	0.712

Table 5: Quadratic weighted kappa for models trained on subtrait features from 3 LLMs.

	W.8.3.A Context	W.8.3.B Narrative Tech.	W.8.3.C Transitions	W.8.3.D Descriptiveness	W.8.3.E Conclusion
GPT-4o Mini	0.24 (0.05)	0.34 (0.05)	0.17 (0.06)	0.08 (0.06)	0.16 (0.04)
Llama3.1 8B	0.13 (0.07)	0.30 (0.02)	0.30 (0.05)	0.15 (0.03)	0.13 (0.04)
Gemma 7B	0.19 (0.12)	0.17 (0.10)	0.16 (0.11)	0.47 (0.13)	0.01 (0.01)

Table 6: Mean and standard deviation of model weights across 6 prompts per LLM, where the standard deviations are in parenthesis. The weights are normalized to 1 within each model for intuitive interpretation.

	GPT vs. Llama	GPT vs. Gemma	Gemma vs. vs. Llama
W.8.3.A	0.711	0.545	0.550
W.8.3.B	0.758	0.575	0.603
W.8.3.C	0.678	0.408	0.489
W.8.3.D	0.688	0.573	0.562
W.8.3.E	0.679	0.428	0.462

Table 7: Pearson correlations for pairwise comparisons of subtrait scores produced by the 3 LLMs.

When we examine the Pearson correlations of the subtrait scores for each pairing of models in Table 7, we find that the subtrait scores from GPT-4o-Mini are highly correlated with the subtrait scores from Llama 3.1 8B, and less correlated with the subtrait scores from less performant Gemma 7B. We take this as evidence that GPT-4o-Mini and Llama 3.1 8B are assessing the same or similar subtraits, whereas Gemma 7B is responding to the rubrics and responses in a significantly different way.

6 Discussion

The successful models have a QWK that hovers around 0.80, roughly 10 points under the operationally deployed models. We find this an encouraging result, particularly given the high level of performance optimizations that go into achieving the maximum possible QWKs in operational scoring. The optimizations make the model more complex, and therefore less interpretable. We are able to achieve a viable, transparently explainable model, with just 5 features. Furthermore, a comparison of the train and test QWKs in Table 5 indicates that there is no overfitting.

The simplicity and transparency of the models allows for a straight-forward look into the subtraits. While we do not have ground-truth annotations for the subtrait scores, we are reassured by how well the weights from the most successful models match with our intuitions with respect to construct relevance. Both GPT-4o-Mini and Llama 3.1 8B weighted the subtrait feature `W.8.3.B` as highest. This is the standard element that, to us, reflects the heart of the narrative genre: “use narrative techniques, such as dialogue, pacing, description, and reflection, to develop experiences, events. and/or characters.” A lower-weighted feature in these models was `W.8.3.D`, “use precise words and phrases, relevant descriptive details, and sensory language…”

According to the most successful models, it is of highest importance that readers are able to understand who the characters are, what happens to them, and what they do; and it is of lesser importance how vividly these things are described. This aligns with our intuitions: a vivid description matters less if readers cannot understand what happened.

In contrast, the least successful model weights `W.8.3.D` the highest, at almost half of the score. Furthermore, it hardly assesses the quality of the conclusion at all.

It may happen that our intuitions about the relative importance of the standard elements of the narrative genre are underinformed, but we would like to emphasize that being able to inspect and reflect upon the weights of a simple linear model gives us useful tools to scrutinize the LLM output. In the case at hand, the tools would lead us to abandon Gemma 7B, given that we have evidence to suggest that the outputs from GPT-4o-Mini and Llama 3.1 8B are more trustworthy with respect to construct relevance.

A final note is that the weights themselves can be useful in downstream tasks. For example, feedback systems can use the weights in algorithms to select the most valuable feedback to display to the student. One could also build models of the scores assigned by individual raters and use the weights to understand rater behavior, revealing, for example, that one rater places more emphasis on descriptiveness whereas another places more emphasis on the quality of the conclusion.

7 Conclusion and future work

We present a method to build, evaluate, and inspect simple explainable models on the basis of subtrait-scores from LLMs, where the rubric criteria for the subtrait scores are derived from the Common Core Standard elements for 8th grade narrative writing. The models using subtrait scores from GPT-4o-Mini and Llama 3.1 8B are able to predict the holistic scores with a QWK of approximately 0.80, which we find particularly noteworthy because the standard-based rubrics were not what the humans used during scoring. We are encouraged that Llama 3.1 8B, an open source model, performs to within 3 points of the GPT model, and that its subtrait scores are highly correlated with GPT's.

The models we present are generally 10-points under the QWK for the operational models; nonetheless, at 0.80 QWK, we find that these simple, transparent, linear regression models with only 5 features may be operationally viable.

We find that the model weights from GPT-4o-Mini and Llama 3.1 8B are in alignment with our own intuitions about the narrative construct, whereas the weights from Gemma 7B are not. On the basis of this evidence, together with the raw model performance, we find that we can trust the LLM output of GPT-4o-Mini and Llama 3.1 8B much more than Gemma 7B. We take this example as an illustration of how this approach allows an inspection of the model by an SME in writing who may not have a great deal of expertise in machine learning, and as a means of understanding how the output of one LLM may differ from others overall.

We would like to better understand the impact of subtrait score accuracy on these regression models. Our previous study on subtraits found low to modest agreement between human ratings of subtrait scores and LLM-produced ones (Andrade-Lotero et al., 2025). As we did not have human-labeled narrative subtrait scores, we can not speak to the accuracy of the LLM-generated scores. In future work, we would like to work with subject matter experts to validate the accuracy and to understand if the resulting weights align with their expert judgment.

This LLM-plus-regression approach also provides a framework for not only for explaining automated scores, but human ones as well. By modeling individual raters, we can glean insights into sources of rater disagreement. We save exploration of this topic for future research.

Lastly, we are encouraged by the possibilities this framework presents for operationalizing AES models in both high stakes and formative settings. The direct interpretability of features allows for improved monitoring and transparency. As reliability of LLM subtrait assessment improves, this approach opens opportunities to enable scoring for more complex constructs and writing behaviors.

8 Limitations

The first limitation to note is that we neither have ground-truth annotations of the subtrait scores nor an in-depth understanding of the subtrait scores produced by the LLM. While the high QWK and alignment of the model weights with our intuitions is highly suggestive, we have not provided strong empirical evidence with respect to the degree to which the LLM is accurately applying the rubric. Related to this, we make a large assumption that our features are indeed construct-relevant and have not explored the impact of including distractor features.

Another important limitation is that the data is from one genre and one grade level. It is not known how well our results generalize to other grades and genres. Similarly, the human agreement is unusually high, suggesting a strong clear signal in the response text itself. It may be that LLMs are less able to assess responses of a more ambiguous nature.

The final limitation is that there were small differences in the datasets used to extract subtrait scores to train and evaluate the 3 LLMs because the sets of responses that the LLMs were able to process successfully were not identical (shown in Table 3). We believe that our sample size is large enough to overcome this limitation; but nonetheless, the cleanest experiment would make comparisons using data sets that are absolutely identical.

9 Acknowledgments

We would like to thank Scott Hellman, Josh Southerland, and Melanie Sharif for their suggestions on the analysis of the results and Scott Hellman for comments on the final draft.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Alejandro Andrade-Lotero, Lee Becker Becker, Joshua Southerland, and Scott Hellman. 2025. Toward subtrait-level model explainability in automated writing evaluation. Paper presented at the 2025 Annual Meeting of the National Council on Measurement in Education (NCME).

Anthropic. 2024. Claude 3.5 haiku: Our fastest model, delivering advanced coding, tool use, and reasoning at an accessible price. `https://www.anthropic.com/claude/haiku`.

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

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

Common Core State Standards Initiative. 2010. Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects. `https://corestandards.org/wp-content/uploads/2023/09/ELA_Standards1.pdf`.

Google DeepMind. 2024. Gemma: Open models built from the same research and technology used to create gemini. `https://ai.google.dev/gemma`.

Scott Deerwester, Susan T Dumais, Thomas K Landauer, George W Furnas, and Robert A Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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.

Ahmed M. Elmassry, Nazar Zaki, Negmeldin Alsheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, 13:121902–121917.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. Trates: Trait-specific rubric-assisted cross-prompt essay scoring. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Scott Hellman, William Murray, Adam Wiemerslage, Mark Rosenstein, Peter Foltz, Lee Becker, and Marcia Derr. 2020. Multiple instance learning for content feedback localization without annotation. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 30–40, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Meta. 2024. Llama 3.1 8b. `https://www.llama.com/models/llama-3/`.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1532–1543. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. `https://openai.com/blog/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. 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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Shixiao Wang. 2024. Deberta with hats makes automated essay scoring system better. *Applied and Computational Engineering*, 52:45–54.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Appendix

```
You are an expert in academic standards with deep knowledge of assessment and
rubric design.  You will be given a standard_id along with information about the
standard as well as parameters for the output like min_score and max_score.  Your
job is to interpret the standard and provide a set of criteria for each score
point that will help to assess the level of student writing with respect to the
standard.

Please provide the criteria in a clear and concise manner, ensuring that
they are:
1. specific to the standard
2. relevant to the grade level of the students.
3. appropriate for the type of writing being assessed
4. written in a way that guides an LLM to evaluate a student's response in a
   reliable and consistent manner.
5. (This is very important) the criteria are written to ensure non- overlapping
    behaviors to encourage the LLM to use the full score range.
```

Figure 1: System instructions that were used to create subtrait rubrics from the narrative standard elements. Additional instructions had to do with JSON formatting; definitions of the keys and values; specifications for the min and max score points, grade, and genre; and encouragements to use the full score range.

		W.8.3.A Context	W.8.3.B Narrative Tech.	W.8.3.C Transitions	W.8.3.D Descriptiveness	W.8.3.E Conclusion
GPT-4o Mini	Prompt 1	0.231	0.285	0.157	0.129	0.198
	Prompt 2	0.276	0.381	0.084	0.096	0.163
	Prompt 3	0.172	0.304	0.243	0.151	0.131
	Prompt 4	0.323	0.310	0.112	0.076	0.178
	Prompt 5	0.257	0.418	0.214	**0.000**	0.111
	Prompt 6	0.199	0.371	0.215	0.022	0.193
Llama3.1 8B	Prompt 1	0.075	0.290	0.302	0.183	0.150
	Prompt 2	0.193	0.272	0.248	0.152	0.135
	Prompt 3	0.127	0.275	0.340	0.169	0.090
	Prompt 4	0.015	0.325	0.347	0.155	0.159
	Prompt 5	0.217	0.326	0.236	0.134	0.088
	Prompt 6	0.130	0.300	0.305	0.085	0.180
Gemma 7B	Prompt 1	0.228	0.068	0.172	0.532	**0.000**
	Prompt 2	**0.000**	0.257	0.218	0.525	**0.000**
	Prompt 3	0.124	0.107	0.311	0.458	**0.000**
	Prompt 4	0.181	0.131	0.025	0.663	**0.000**
	Prompt 5	0.336	0.326	0.033	0.304	**0.000**
	Prompt 6	0.279	0.138	0.202	0.347	0.033

Table 8: Relative weights of LLM-generated subtrait features in regression models. We observe a fair amount of consistency in the weights across prompts within models, and similar weights between GPT-4o-Mini and Llama3.1 8B. We also observe that Gemma 7B's assessment of the conclusion did not offer a unique contribution to the prediction of the score.

```
{
  "4": {
    "criteria": [
      "Exceeds Expectations",
      "Masterfully establishes a complex and engaging narrative context",
      "Provides a highly sophisticated and nuanced introduction of narrator and/or
      characters",
      "Creates an exceptionally clear and compelling point of view",
      "Organizes events with remarkable logical flow and natural progression",
      "Demonstrates advanced narrative techniques that immediately capture the
      reader's interest"
    ]
  },
  "3": {
    "criteria": [
      "Meets Expectations",
      "Effectively establishes a clear narrative context",
      "Introduces narrator and/or characters with sufficient detail",
      "Presents a distinct and appropriate point of view",
      "Organizes events in a logical and coherent sequence",
      "Provides a solid foundation for the narrative that guides the reader's
      understanding"
    ]
  },
  "2": {
    "criteria": [
      "Approaching Expectations",
      "Provides a basic narrative context with some gaps or lack of clarity",
      "Partially introduces narrator and/or characters with minimal details",
      "Demonstrates an inconsistent or somewhat unclear point of view",
      "Attempts to organize events, but the sequence may have some minor logical
      inconsistencies",
      "Shows an emerging understanding of narrative introduction"
    ]
  },
  "1": {
    "criteria": [
      "Below Expectations",
      "Offers a minimal or confusing narrative context",
      "Provides little to no introduction of narrator and/or characters",
      "Lacks a clear or coherent point of view",
      "Events are poorly organized or difficult to follow",
      "Struggles to establish a meaningful narrative foundation"
    ]
  },
  "0": {
    "criteria": [
      "Insufficient",
      "No discernible narrative context",
      "No introduction of narrator or characters",
      "No identifiable point of view",
      "No coherent event sequence",
      "Fails to create any meaningful narrative structure"
    ]
  }
}
```

Figure 2: Example rubric for the `W.8.3.A` standard element subtrait.

```
Assess the student's ability to effectively introduce a narrative by
establishing a clear context, point of view, and characters. Evaluate how well
the writer sets up the story and creates a logical, natural progression of
events. Consider the sophistication of the narrative setup, the clarity of
the introduction, and the coherence of the event sequence.

Return the chosen score_point as well as up to three small excerpts from the
response as evidence, without any modification or additional reasoning. The
excerpts should only be subsets of the original response text. It is okay to
return fewer than the max amount of excerpts, if some aren't good relative
to the others. Also, don't return the same excerpt twice. If the student got
the highest score_point, you should provide feedback summarizing what they
did well. If they did not get the highest score_point, you should give
feedback with a high level suggestion on how to improve. Feedback should be
worded to communicate with a student in grade 8 and limited to the specific
criteria in the rubric. You should not mix in unrelated analyses Return only
JSON containing the score_point, optional feedback, and each optional excerpt.

You should specifically evaluate the response based only on the following
scorepoint criteria:
//{Rubric appears here}

Score and provide feedback for this response:
//{Response text appears here.}
```

Figure 3: Prompt that was used to elicit subtrait assessments from 3 LLMs.

Validating Generative AI Scoring of Constructed Responses with Cognitive Diagnosis

Hyunjoo Kim
University of Illinois Urbana-Champaign
hyunjoo5@illinois.edu

Abstract

Generative AI has been investigated as a tool for scoring constructed responses (CRs). Although generative AI can provide both numeric scores and qualitative feedback on written tasks effectively and efficiently, its lack of transparency in output makes it challenging to build strong validity. Validity evidence for outputs from generative AI scoring is evaluated mainly through expert reviews or statistical concordance measures with human raters. As additional validity evidence for CR scores produced by generative AI, particularly for essay-type tasks, this research examines the feasibility of applying the Cognitive Diagnosis (CD) framework in psychometrics. The results of the study indicate that the classification information of CRs and item-parameter estimates from cognitive diagnosis models (CDMs) could provide a new perspective as additional validity evidence for CR scores and feedback from generative AI with less human oversight.

1 Introduction

Constructed responses (CRs), ranging from short sentences to full essays, are widely used to assess understanding of subject-matter knowledge. Unlike multiple-choice questions, which require selecting a given option, CRs ask examinees to generate answers in their own words. This task not only requires applying knowledge but also engaging in critical thinking and constructing logical arguments. As CRs can reveal a deeper level of understanding than simple recall questions, they purportedly provide strong evidence of the skills examinees possess. Accurately scoring CRs and providing personalized, detailed feedback is hence crucial, as it can support more subsequent learning (Gan and Liu, 2021; Kochmar et al., 2020; Parr and H.S., 2010). However, scoring CRs with high-quality feedback from human raters is labor-intensive, and issues such as fatigue and subjectivity are inevitable. These challenges limit the widespread use of CRs, especially in large-scale assessments.

Recent advances in generative AI, such as ChatGPT, have been explored as potential alternatives to address these limitations (Casabianca et al., 2025; Mizumoto and Eguchi, 2023; Yoon et al., 2023). Generative AI offers efficiency and scalability for CR scoring, but concerns about its sensitivity and lack of transparency make building strong validity evidence challenging. To date, most validity evidence for generative AI-based scoring systems has come from expert reviews or statistical concordance measures comparing AI outputs with human ratings or traditional Natural Language Processing (NLP) scores (Bui and Barrot, 2025a; Casabianca et al., 2025; Seßler et al., 2024; Tate et al., 2024).

As an additional source of validity evidence, this research draws on the Cognitive Diagnosis (CD) framework in psychometrics. CD is designed to evaluate examinees' mastery of instructional content and to provide feedback on their strengths and weaknesses in terms of learned and yet-to-be learned cognitive skills. This way, CD offers fine-grained diagnostic information, unlike traditional assessments that typically provide a single numeric score. It has been successfully applied to language assessments, including English writing, demonstrating its value in diagnosing both examinees' skills and the characteristics of test items (Lee and Sawaki, 2009; Mei and Chen, 2022). By bridging the CD framework with generative AI-based CR scoring, this study explores the extent to which CD can increase the interpretability and transparency of generative-AI scoring systems, thereby improving their validation.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 166–177
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

2 Background

2.1 Generative AI for CR scoring and its validity evidence

The use of generative AI has the potential to automate CR grading by providing both quantitative scores and personalized feedback. Recent studies have investigated the performance of several generative AI systems for this purpose (Bui and Barrot, 2025b; Pack et al., 2024). However, a key distinction lies in the process by which traditional methods and generative AI produce scores. While human raters and traditional Natural Language Processing (NLP) approaches rely on trained expertise and handcrafted features, generative AI derives scores through processes that are largely opaque. This lack of transparency necessitates strong validity evidence for the use of generative AI in CR scoring.

Because the scoring process of generative AI is often difficult to interpret, one integral approach to validation has been to measure agreement between human raters and generative AI outputs. To date, the most common evaluation metrics have been Spearman rank correlation coefficients (r) and Quadratic Weighted Kappa (QWK) between scores assigned by generative AI and those assigned by human raters. The Spearman rank correlation coefficient measures the monotonic association between ordinal data, and QWK evaluates inter-rater agreement on ordinal scales by assigning heavier penalties for larger disagreements and lighter penalties for smaller ones, with penalties increasing quadratically as rating differences grow. Landis and Koch (1977) provided interpretive benchmarks for QWK: below 0.0 = Poor; 0.00–0.20 = Slight; 0.21–0.40 = Fair; 0.41–0.60 = Moderate; 0.61–0.80 = Substantial; 0.81–1.00 = Almost perfect. Although these metrics reflect the degree of agreement and thus serve as evidence of validity, they are limited. Specifically, they capture only overall agreement without offering insight into the reasoning behind scores or into systematic differences between responses that receive similar scores. Hence, complementary approaches are necessary to provide stronger validity evidence of generative AI's outputs.

2.2 Cognitive diagnosis models and their use in writing assessments

Cognitive Diagnosis (CD) provides fine-grained information about examinees' mastery of specific cognitive skills by describing ability in a knowledge domain as a composite of K (specific) binary skills, called "attributes" (denoted α_k, $k = 1, 2, ..., K$) (Sessoms and Henson, 2018; von Davier and Lee, 2019). These attributes form profiles that define proficiency classes, written as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)'$–mastered (1) or not (0)–with 2^K possible combinations.

Cognitive diagnosis models (CDMs) classify examinees into one of these proficiency classes based on their observed item responses, Y_{ij}, where $Y_{ij} = 1$ if examinee $i = 1, 2, ..., N$ answered item $j = 1, 2, ..., J$, $Y_{ij} = 0$ correctly, and $Y_{ij} = 0$ otherwise. As such, CDMs are restricted latent class models in which the latent constructs are attributes, and the discrete latent classes are examinees' proficiency classes.

Similarly, test items are also characterized by K-dimensional attribute profiles $\mathbf{q}_j$, which specify the attributes required for a correct response. The q_{jk} entries indicate whether the jth item requires the kth attribute (1) or not (0). These $\mathbf{q}$-vectors form a "Q-matrix" of size $J \times K$ (Tatsuoka, 1985).

Using the observed item-response data $\mathbf{Y}_{N \times J}$ and Q-matrix $\mathbf{Q}$, typically provided by domain experts, CDMs estimate each examinee's attribute profile by selecting the most likely class. The functional relation between attribute mastery and the probability of a correct response to an item, $P(Y_j = 1|\boldsymbol{\alpha})$, is modeled according to the chosen CDM, which specifies how attributes interact with items. Various CDMs differ in complexity and assumptions (e.g., compensatory vs noncompensatory; conjunctive vs disjunctive), but they all share the core principle that the probability of success on an item depends on the interaction between an examinee's attribute profile and the item's $\mathbf{q}$-vector (Henson et al., 2009; Williamson, 2024; Dibello et al., 2006; Rupp and Templin, 2008). The selection of a CDM depends on assessment goals.

Several CDMs have been applied to writing assessments, with a focus on English as a Foreign Language (EFL) (Effatpanah et al., 2019; He et al., 2021; Kim, 2011; Shi et al., 2023; Xie, 2016). Unlike typical applications of CD, which use direct item responses (correct/incorrect), writing assessments first require transforming examinees' constructed responses into binary response data. Prior CD studies on writing have commonly used rating checklists for this transformation, a task done by writing experts. Kim (2011) developed an empirically derived, descriptor-based (EDD) checklist of 35 fine-grained descriptors measuring five constructs–content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics (see Appendix A.2 for the descriptors of EDD). In her study, they had ten ESL (English as a second language) teachers assess 480 TOEFL (Test of English as a Foreign Language) essays using the EDD checklist and converted them into binary item-response data. Then, using the reduced reparameterized unified model (R-RUM) (Hartz, 2002), they diagnosed learners' writing ability and demonstrated that the CD framework could reliably identify strengths and weaknesses. Building upon this work, Xie (2016) applied the same EDD checklist with R-RUM in a different assessment context and found that it provided more diagnostic information than raw scores alone. The checklist was further validated using the additive CDM (ACDM) (de la Torre, 2011) and R-RUM (Effatpanah et al., 2019; Shahsavar, 2019). More recently, He et al. (2021) employed polytomous CDMs to extend binary models, diagnosing English learners' writing ability with a different checklist measuring four constructs. Under this approach, each constructed response is treated as an "examinee," each checklist descriptor as an "item," and each latent construct targeted in the rubric as an "attribute." When a response meets a checklist criterion, $Y_{ij} = 1$; otherwise, $Y_{ij} = 0$. With this mapping, CDMs estimate each constructed response's attribute profile/proficiency class, indicating which rubric-defined attributes are present.

3 Methodology

3.1 Dataset

Ideally, this study would have used essay datasets from previous writing assessment studies, such as the TOEFL independent writing, along with the corresponding binary item-response data coded by expert raters using the empirically derived descriptor-based diagnostic (EDD) checklist and the Q-matrix. However, because those datasets were not publicly accessible, we sought an open essay dataset that closely resembled the formats and tasks used in prior research, so that the EDD checklist and Q-matrix developed by (Kim, 2011) and subsequent studies could still be applied.

As a surrogate, we identified Set 2 from the Kaggle "The Hewlett Foundation: Automated Essay Scoring (AES)" dataset[1]. Set 2 consists of persuasive essays averaging 350 words, written by grade 10 native English speakers. The scoring scale ranged from 2 to 10, with two domains: Writing Applications (1-6) and Language Conventions (1-4). A detailed comparison of the formats between TOEFL Independent Writing and AES Set 2 is presented in Table 1.

The dataset contains 1,800 essays, each scored by two human raters, along with a detailed rubric and exemplar essays for each score. Based on the prior studies, we selected 500 essays from Set 2 for analysis. The specific essay prompt is provided in Appendix A.1.

Table 1: Comparison between TOEFL Independent Writing Task and AES Set 2 Dataset

Feature	TOEFL Ind.	AES Set 2
Task type	Persuasive "agree/disagree"	Persuasive "agree/disagree"
Avg length	300–400 words	350 words
Time limit	30 minutes	Not specified
Population	EFL learners (Primarily ages 21–25)	Grade 10 native English speakers
Score scale	1–5	2–10
# of Raters	2	2
Domains	Single holistic score	Writing Applications (1–6) + Language Conventions (1–4)

3.2 Procedure

We used the auto version of ChatGPT-5 from OpenAI, with the default temperature setting (which controls the variability of responses) for all stages of this study.

First, ChatGPT-5 was customized with the instructions to act as a rater evaluating student essays. The essay scoring guide and rubric file, including exemplar essays from the original AES website, were provided via Retrieval Augmented Generation (RAG). This setup enabled ChatGPT-5 to assign numerical scores and generate qualitative feedback of the selected 500 essays. The scoring prompt was adapted from Casabianca et al. (2025) and tailored to this study (see Appendix A.1). In this setting, generation followed a zero-shot Chain of Thought (CoT) approach, since no worked examples were included. Each essay was scored using a single

---

[1] `https://www.kaggle.com/c/asap-aes/data`

prompt that requested both the score and feedback.

The major challenge was the absence of binary item-response data for the AES dataset based on the EDD checklist, which are required inputs for CDMs. Without resources to recruit expert raters to transform essays into item-response data, we relied on ChatGPT-5 for this task. To mitigate concerns about reliability, we used a few-shot Chain of Thought (FsCoT) design. Three essays, drawn from outside the 500-essay sample, were manually rated with the EDD checklist and provided as exemplars in the prompt to ChatGPT-5. The model was then asked to generate a binary vector of size 1×35 for each essay, indicating whether each descriptor was met. This task was conducted in a separate session from the scoring task. The full prompt for this task is included in Appendix A.2.

This process yielded a 500×35 item-response matrix, ready for CDM analysis with the Q-matrix by Kim (2011). The Q-matrix specifies five latent attributes that the EDD checklist aims to measure: Content Fulfillment (CON), Organizational Effectiveness (ORG), Grammatical Knowledge (GRM), Vocabulary Use (VOC), and Mechanics (MCH). Based on the initial Q-matrix by Kim (2011), we empirically validated it using the method suggested by de la Torre and Chiu (2016), as in previous studies, and refined it to reduce misspecification and enhance CDM performance for datasets used in the study.

3.3 Data Analysis

The constructed item-response matrix and the refined Q-matrix were used as inputs for the Reduced Reparameterized Unified Model (R-RUM), which has been widely applied in prior CDM-based writing assessment research. The item response function of the R-RUM is given by:

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{ik})},$$

where $0 < \pi_j^* < 1$ represents the probability of a correct response when examinee i has mastered all attributes required by item j, and $0 < r_{jk}^* \leq 1$ is the penalty parameter for not mastering attribute k. π_j^* and r_{jk}^* correspond to item difficulty and item discrimination, respectively (Kim, 2011). Values of $\pi_j^* < 0.6$ suggest overly difficult items, and $r_{jk}^* < 0.5$ indicates that an item discriminates well between mastery and non-mastery on skill k (Roussos et al., 2007).

The validity evidence for the generative AI scoring was evaluated by examining the following:

1. the agreement between ChatGPT scores and human rater scores via Spearman correlation and QWK

2. the consistency between ChatGPT scores & feedback and the attribute mastery profiles estimated by the CDM.

For example, generative AI scoring can be supported as valid if essays grouped into the same proficiency class by CDMs also receive consistent scores from ChatGPT, and if the qualitative feedback aligns with mastered and non-mastered attributes.

We also analyzed the estimated item parameters from R-RUM to enhance the interpretability of ChatGPT's scoring process. This analysis allowed us to examine the extent to which latent attributes (from the Q-matrix) or items (descriptors from the checklist) influenced ChatGPT's scoring. For example, two constructed responses might receive the same overall score from ChatGPT, yet differ in their attribute profiles or in the number of mastered attributes, providing insight into how ChatGPT's scoring reflects specific attributes.

Although generating item-response datasets with expert raters would be more rigorous in practice, our approach demonstrates the feasibility of using CDMs to evaluate generative AI scoring in greater depth and to strengthen validity evidence by moving beyond numerical scores to attribute-level reasoning.

4 Results

As a preliminary check, inter-rater reliability between two human experts was examined using Spearman correlation coefficients and quadratic weighted kappa (QWK). Both indices were 0.82, indicating strong agreement. A t-test comparing the two raters' mean raw scores showed no statistically significant difference (p-value $= 0.55$).

Next, the validity of ChatGPT-5 scores was evaluated against the human ratings. For total scores, the Spearman coefficient was 0.67 and QWK was 0.56, indicating moderate agreement. At the domain level, Domain 1 showed 0.58 (Spearman) and 0.55 (QWK), while Domain 2 was lower–0.56 (Spearman) and 0.26 (QWK). The relatively weak agreement in Domain 2 reflects ChatGPT's tendency to score more strictly on language conventions.

The cross-tabulation of human mean scores and ChatGPT-5 scores shown in Table 2 confirms this pattern. ChatGPT frequently assigned scores about one point lower than the human raters, particularly in the mid-range of the scale.

Table 2: Agreement matrix between human mean total scores (Hu) and ChatGPT-5 (AI) total scores

Hu \AI	2	3	4	5	6	7	8	9	10
2	1	2	0	0	0	0	0	0	0
2.5	0	2	0	0	0	0	0	0	0
3	1	1	2	0	0	0	0	0	0
3.5	1	1	2	0	0	0	0	0	0
4	0	4	3	4	1	0	0	0	0
4.5	0	5	6	2	1	0	0	0	0
5	0	4	5	14	5	0	0	0	0
5.5	0	2	1	8	1	1	0	0	0
6	0	7	10	33	34	10	5	0	1
6.5	0	2	2	12	16	7	4	0	0
7	0	2	4	20	30	15	11	1	0
7.5	0	0	3	3	9	14	12	1	0
8	0	0	0	12	27	28	46	9	3
8.5	0	0	0	0	1	2	11	3	1
9	0	0	0	1	0	0	4	3	0
10	0	0	0	0	0	0	0	0	1

Applying R-RUM, essays were classified into proficiency profiles (Table 3). Fourteen of the 32 possible classes were observed, with the distribution concentrated in higher-mastery profiles. Across all classes, ChatGPT-5 assigned lower average scores than human raters. Profiles with fewer mastered attributes tended to receive lower scores overall, while broader mastery was associated with higher scores from both humans and AI.

Table 3: Summary of human (Hu) and ChatGPT-5 (AI) scores by proficiency class

Class	Count	Hu Mean (SD)	AI Mean (SD)
(00000)	25	4.64 (1.56)	3.96 (1.67)
(01000)	6	6.58 (0.58)	5.33 (0.63)
(00100)	42	5.15 (1.27)	4.05 (0.96)
(00001)	8	6.94 (0.82)	6.38 (1.19)
(11000)	4	7.00 (1.83)	5.75 (1.50)
(10001)	3	8.00 (0.00)	6.33 (1.53)
(01100)	67	6.37 (0.96)	5.30 (0.98)
(00101)	14	5.50 (1.14)	5.14 (0.86)
(11100)	73	6.84 (0.89)	5.60 (1.04)
(11010)	1	7.00 (NA)	6.00 (NA)
(01101)	55	6.74 (1.08)	5.98 (0.83)
(11110)	9	7.06 (1.10)	6.78 (1.39)
(11101)	39	6.90 (0.93)	6.82 (0.97)
(11111)	154	7.64 (0.80)	7.49 (1.20)

Beyond serving as validity evidence for ChatGPT's scores, the estimated proficiency classes of essays also aid in the interpretation of ChatGPT's scoring with respect to mastery status. Table 4 presents three essays' estimated proficiency classes, scores, and ChatGPT feedback. Essays from higher-level classes received feedback acknowledging strengths, while essays from lower-level classes received feedback identifying weaknesses consistent with their attribute profiles. For example, Essay 3213–mastering CON, ORG, GRM, and MCH but not VOC–received relatively positive comments, with vocabulary flagged as the main area for improvement. In contrast, Essays 2995 and 3103, from lower classes, received feedback emphasizing underdeveloped reasoning, weak transitions, and frequent grammar issues.

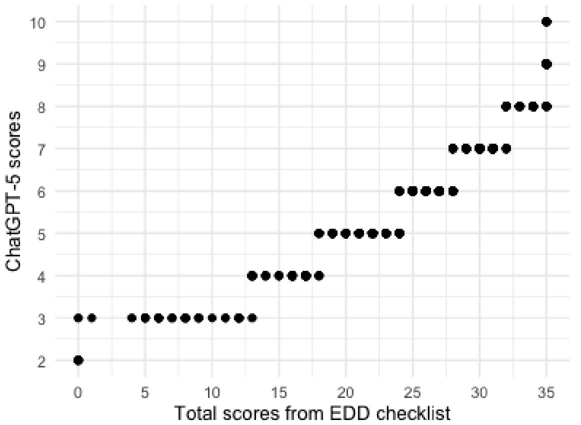


Figure 1: Total EDD checklist score vs. ChatGPT-5 score

Furthermore, even when two essays shared the same AI and/or human score, their estimated proficiency classes show more fine-grained diagnostic information with different combinations of attributes. The high alignment between classification results and ChatGPT feedback strengthens the point that the generative AI's comments are sufficiently valid to guide targeted revisions.

Figure 1 further illustrates the relation between EDD checklist totals (0–35) and ChatGPT-5 scores (2–10). A clear positive trend emerges: as more checklist items were met, ChatGPT assigned higher scores. This convergence with the checklist also supports the construct validity of ChatGPT's scoring.

Drawing on the CD framework provides further insight into factors that may influence ChatGPT's scoring. Based on the R-RUM estimates, the proportion mastering each attribute was: CON$= 0.57$, ORG$= 0.82$, GRM$= 0.91$, VOC$= 0.33$, and MCH$= 0.55$. Thus, many essays satisfied GRM

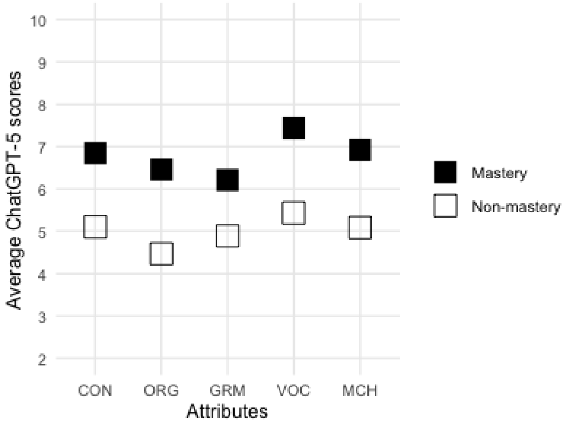


Figure 2: Average ChatGPT-5 scores by mastery status for each attribute

but not VOC (e.g., limited vocabulary variety). As shown in Figure 2, ChatGPT scores were higher when an attribute was mastered, with the largest mastery-non-mastery gap for VOC–indicating that vocabulary mastery was especially influential on ChatGPT-5 scoring.

The checklist-descriptor parameter estimates (Table 5) corroborate these patterns. Descriptors linked to VOC exhibit comparatively strong discrimination, consistent with the larger VOC-related score gap in Figure 2. Based on Table 5, most descriptors showed high π^* values (near 1) in general, indicating that once the relevant attributes were mastered, the probability of meeting the corresponding criteria was very high. Some $\gamma*$ parameters showed little or no penalty, implying that certain descriptors were less diagnostic of underlying mastery.

Overall, ChatGPT-5 scoring agrees reasonably well with human raters, although it tends to be more strict, particularly in certain domains. Together with numerical scores, its narrative feedback corresponds meaningfully with latent attribute (non)mastery, showing where specific strengths and weaknesses lie in each essay.

5 Discussion

This study aimed to adapt psychometric modeling, grounded in cognitive diagnosis theory, to introduce new forms of validity evidence for generative-AI scoring. Specifically, we demonstrated how cognitive diagnostic models and checklist-based item-response data of essays can serve as a framework for investigating ChatGPT-generated scores at both the holistic and attribute levels.

Several limitations should be noted. First, we did not examine the consistency of AI scores across multiple time points. Given that large language models are sensitive to prompt wording and contextual framing, test–retest reliability remains an open question. Future work should estimate intraclass correlation coefficients (ICCs) to evaluate score stability over repeated administrations (Seßler et al., 2024).

Second, our results may be suboptimal because (i) the EDD checklist was not originally designed for the essay samples in this study, and (ii) our artificial item-response dataset has not undergone validation. Access to established item-response datasets from prior studies would provide stronger grounding and allow more robust validation of this approach. Moreover, a formal implementation would require sustained expert involvement–from rubric design and Q-matrix specification to interpreting constructed-response evaluations.

Despite these limitations, the findings are promising. While challenges remain in applying CDMs within AI-automated assessment, the results suggest that such models can enhance the transparency of AI scoring. By linking scores to specific latent attributes, this framework provides an additional source of validity evidence-—helping explain not just what score was assigned, but also why.

In sum, this study demonstrates the feasibility of using psychometric frameworks to support the validation of AI-generated scores. By combining the interpretive strengths of cognitive diagnosis with the efficiency of generative AI, this approach offers a novel pathway toward transparent, evidence-based scoring systems in educational assessment.

Future research should extend this work in several directions. Different CDMs could be compared by model fit across diverse testing contexts. Beyond extended essays, the approach could also be adapted to shorter constructed-response tasks, such as sentence-level prompts. For these tasks, the checklist and Q-matrix would be smaller and easier to operationalize, reducing reliance on expert judgment while still producing meaningful validity evidence.

References

N. M. Bui and J. S. Barrot. 2025a. Chatgpt as an automated essay scoring tool in the writing classrooms:

how it compares with human scoring. *Education and Information Technologies*, 30:2041–2058.

N. M. Bui and J. S. Barrot. 2025b. Using generative artificial intelligence as an automated essay scoring tool: a comparative study. *Innovation in Language Learning and Teaching*, pages 1–16.

J. M. Casabianca, D. F. McCaffrey, M. S. Johnson, N. Alper, and V. Zubenko. 2025. Validity arguments for constructed response scoring using generative artificial intelligence applications. *Preprint*, arXiv:2501.02334.

J. de la Torre. 2011. The generalized dina model framework. *Psychometrika*, 76:179–199.

J. de la Torre and C-Y. Chiu. 2016. A general method of empirical q-matrix validation. *Psychometrika*, 81:253–273.

L.V. Dibello, L.A. Roussos, and W.F. Stout. 2006. Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao and S. Sinharay, editors, *Handbook of Statistics*, pages 979–1030. Elsevier.

F. Effatpanah, P. Baghaei, and A.A. Boori. 2019. Diagnosing efl learners' writing ability: a diagnostic classification modeling analysis. *Language Testing in Asia volume 9*, 9, 12.

Z. Gan and F. Liu. 2021. Teacher feedback practices, student feedback motivation, and feedback behavior: How are they associated with learning outcomes? *Frontiers in Psychology*, 12:697045.

S. Hartz. 2002. *Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

L. He, Z. Jiang, and S. Min. 2021. Diagnosing writing ability using china's standards of english language ability: Application of cognitive diagnosis models. *Assessing Writing*, 50:100565.

R.A. Henson, J.L. Templin, and J.T. Wilse. 2009. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210.

Y-H. Kim. 2011. Diagnosing eap writing ability using the reduced reparameterized unified model. *Language Testing*, 28(4):509–541.

E. Kochmar, D. Vu, R. Belfer, V. Gupta, I. Serban, and J. Pineau. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. *Artificial Intelligence in Education*, 12164:140–146.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174.

Y.W. Lee and Y. Sawaki. 2009. Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, 6(3):172–189.

H. Mei and H. Chen. 2022. Cognitive diagnosis in language assessment: A thematic review. *RELC Journal*, 55(2):529–537.

A. Mizumoto and M. Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2:100050.

A. Pack, A. Barrett, and H. 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.

J.M. Parr and Timperley H.S. 2010. Feedback to writing, assessment for teaching and learning and student progress. *Assessin Writing*, 15:68–85.

L. A. Roussos, L. V. DiBello, W. Stout, S. M. Hartz, R. A. Henson, and J. L. Templin. 2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications*, chapter The fusion model skills diagnosis system. Cambridge University Press.

A.A. Rupp and J. Templin. 2008. Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6(4):219–262.

K. Seßler, M. Fürstenberg, B. Bühler, and E. Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *Preprint*, arXiv:2411.16337.

J. Sessoms and R. Henson. 2018. Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, 16:1–17.

Z. Shahsavar. 2019. Diagnosing english learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education*, 6(1).

X. Shi, X. Ma, W. Du, and X. Gao. 2023. Diagnosing chinese efl learners' writing ability using polytomous cognitive diagnostic models. *Language Testing*, 41(1):109–134.

T.P. Tate, J. Steiss, D. Bailey, S. Graham, Y. Moon, D. Ritchie, W. Tseng, and M. Warschauer. 2024. Can ai provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

K.K. Tatsuoka. 1985. A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics*, 12:55–73.

M. von Davier and Y.S. Lee. 2019. *Handbook of diagnostic classification models*. Springer International Publishing, Cham, CH.

J. Williamson. 2024. Cognitive diagnostic models and how they can be useful. *Cambridge University Press & Assessment*.

Q. Xie. 2016. Diagnosing university students' academic writing in english: Is cognitive diagnostic modelling the way forward? *Educational Psychology*, 37(1):26–47.

S-Y. Yoon, E. Miszoglad, and L.R. Pierce. 2023. Evaluation of chatgpt feedback on ell writers' coherence and cohesion. *Preprint*, arXiv:2310.06505.

A Appendix

A.1 ChatGPT prompt for scoring essays

A student is assigned to an essay question as follows:

"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf—that work I abhor—then you also have exactly the same right, and so does everyone else. And then we have no books left on the shelf for any of us." (Katherine Paterson Author)

Write a persuasive essay for a newspaper reflecting your views on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.

In addition to the rubric files in the "File" folder, the following are the specific guidelines for scoring the essay question above. Grade the essay based on all the rubrics provided. Remember that there are two types of scores: Domain 1 (Writing Applications) and Domain 2 (Language Conventions). As mentioned in the rubrics, give a score from 1 to 6 for Domain 1 and a score from 1 to 4 for Domain 2. Then, provide a 20-30 word feedback highlighting strengths and areas for improvement. Return the response in a JSON format of:

{ Score for Domain 1, Score for Domain 2, "Reasons": [{ reasons }] } The reasons should be an array of 3 objects. Each object should be in the structure shown above and described below. For each object in the reasons array, a reason must be provided. This reason should be one of the reasons for giving the score based on the rubric. The reason should not be a full sentence but in 20-30 words and be suitable to be displayed as bullet points to a person with a high school-college-level education, rather than copied directly from the rubric.

This is the end of the explanation. Now, I'll give you an essay from each student.

A.2 ChatGPT prompt for the checklist

You are an essay rater. Using the provided 35-item EDD checklist below, evaluate 10th-grade essays. For each item on the checklist, determine if the essay satisfies the criterion.

The EDD checklist is presented as follows:

1. This essay answers the question.
2. This essay is written clearly enough to be read without having to guess what the writer is trying to say.
3. This essay is concisely written and contains few redundant ideas or linguistic expressions.
4. This essay contains a clear thesis statement.
5. The main arguments of this essay are strong.
6. There are enough supporting ideas and examples in this essay.
7. The supporting ideas and examples in this essay are appropriate and logical.
8. The supporting ideas and examples in this essay are specific and detailed.
9. The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.
10. Each body paragraph has a clear topic sentence tied to supporting sentences.
11. Each paragraph presents one distinct and unified idea.
12. Each paragraph is connected to the rest of the essay.
13. Ideas are developed or expanded well throughout each paragraph.
14. Transition devices are used effectively.
15. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.
16. This essay demonstrates an understanding of English word order.
17. This essay contains few sentence fragments.
18. This essay contains few run-on sentences or comma splices.
19. Grammatical or linguistic errors in this essay do not impede comprehension.
20. Verb tenses are used appropriately.
21. There is consistent subject-verb agreement.
22. Singular and plural nouns are used appropriately.
23. Prepositions are used appropriately.

24. Articles are used appropriately.
25. Pronouns agree with referents.
26. Sophisticated or advanced vocabulary is used.
27. A wide range of vocabulary is used.
28. Vocabulary choices are appropriate for conveying the intended meaning.
29. This essay demonstrates facility with appropriate collocations.
30. Word forms (noun, verb, adjective, adverb, etc) are used appropriately.
31. Words are spelled correctly.
32. Punctuation marks are used appropriately.
33. Capital letters are used appropriately.
34. This essay contains appropriate indentation.
35. Appropriate tone and register are used throughout the essay.

Score each criterion with a binary value—1 if the essay meets the criterion or 0 if the essay does not. Your final evaluation should be a binary vector of size (1 x 35).

Let me give some examples of a student's essays and the corresponding evaluations:

Example 1: Certain materials being removed from libraries such as books, music and magazines, shouldn't be removed from the libraries. It gives people a chance to understand how the real world @CAPS2. Having certain materials such as books and music definitly should not be removed, because most books and music can show most people how bad the statement in the book @CAPS2 or how bad the lyrics are in a song, and help that person to avoid that type of thing that the book or song @CAPS2 saying to the reader or listener. People should give every type of music at least a try and not always doubt what they hear about what people say about that type of music. I always hear about people saying how bad the band @PERSON1 A.M. @CAPS2, just because in the lyrics it talks about drugs and how much cursing each song has. Really the band @CAPS2 talking about one mans life and how he turns his life from being a drug addict to having the best life someone could ever live. People always doubted him and never gave his music a chance. Another example would be @PERSON1's book, '@CAPS1 @CAPS2 @CAPS3 @CAPS4' for it talks about drug addicts, homeless people, people who have been born with disfigured arms or even someone who lost there legs, and telling how beautiful each and everyone of them really are. His book taught me a few things and made me think different about people. It doesn't matter how they look or how they talk, no matter what, that person @CAPS2 beautiful. As far as movies and magazines has gone within the last few years, I think that the also shouldn't be taken from libraries. I think @CAPS1 for the same reason of how I feel about the books and music. Of course we see previews of movies and think that they @MONTH1 not be good, but libraries shouldn't keep leave them out. Movies @CAPS2 a great way to learn how to treat others and how to act around other people when you don't know how to act. If you act differently around people that you've never been around before, then you could feel embarassed or maybe even get @CAPS4. Movies can help people learn about the real world by seeing how to do those type of things as we get older. Same goes with the magazines, they also help people see what not to do or to help them understand the consequences of something that shouldn't be done. Knowing what to do from a magazine could possible save your life or perhaps maybe even someone elses life. I don't understand why some libraries would want to banned certain materials to help people understand the things that happen in someone elses life and to help them not make the same mistakes as that person once did.

Evaluation based on the EDD checklist: $(1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1)$

Example 2: Do you believe that certain should be removed i think so be no that yes i think should no person that in chager the book, music, movies, magazines, ect., that be no agure why do i think if you need that please think i no thank you please if i need why do we if know that if i failure the this test i who need to graduate please the children allow to home please yes. Why do we need to be a prafece person please why do we need to do this why write this assgiment because you mean to be the best teaches ever and ever facebook is my password is @PERSON1 @NUM1 that why i need my myspace is the same thingh but different at same time please know that i need to know i really i need to my e-mail address is @EMAIL1 that is my e-mail please work m

Evaluation based on the EDD checklist: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0)$

Example 3. Do you think that libraries should remove certain materials off the shelves? People have different oppions, of whats good and whats bad. I have read and seen a lot of books in my life *time. I hear people telling me, 'oh dont read that book its a bad book.'But I ask myself, @CAPS2 do I know it's a bad book when I haven't even given it a chance?' @CAPS1 are some books, music, movies, and magazines out @CAPS1 that are offensive. Yet we still want to read, listen, watch, and look at them. If we tried to remove all the offensive books, from the libraries we wouldn't have anything left on the shelves. Katherine Paterson said, 'If I have the right to remove that book from the shelf that work i abhor- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us.' Katherine Paterson makes a great point out of her quote. Why should we have to remove a book if just some people think its offensive? Ask yourself the question again, '@CAPS2 do you know it's a bad book when you haven't even given it a chance?' @CAPS3't judge a book by what you hear. Find out what your own oppion is*

Evaluation based on the EDD checklist: $(1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1)$

Refer to the above evaluation examples, rate the provided essay in the same form using the EDD checklist. When you evaluate, ignore words starting with an "@" symbol in essays, as it is marked for privacy and copyright issues. Also, one thing to remember regarding the indentation of the essay–which is the 34th criterion of EDD—is that consider tab-spacing between two sentences as the indentation for a new paragraph.

This is the end of the instruction. Now, I'm providing each essay.

Table 4: Essay classifications, human and AI scores, and excerpts of ChatGPT feedback

Essay #	Class	Hu Score	AI Score	AI Feedback
2995	11100	6	6	• Essay clearly takes a stance against censorship with two main arguments, but reasoning is repetitive and lacks deeper development or specific supporting details. • Organization includes introduction, body, and conclusion, but transitions are weak, and the structure is somewhat formulaic with limited expansion of ideas. • Frequent grammar and spelling errors (e.g., shout for should, there selves, ofeneded, intersting) and awkward phrasing disrupt fluency and reduce overall polish.
3103	01000	6	6	• The essay presents a clear opinion in favor of censorship, with supporting reasons such as protecting children from adult content and the influence of media. Personal experience is included, which adds some strength to the argument, though examples are somewhat vague and underdeveloped. • Organization is generally clear with an introduction, supporting paragraphs, and conclusion, but transitions are weak and ideas repeat instead of expanding. Some claims lack detail or evidence, which reduces the persuasiveness. • Language control is limited. There are numerous errors in grammar, spelling, and word choice (e.g., ’boundries,’ ’consent,’ ’hazardess,’ ’drasticly,’ ’cencorships,’ ’portarying’). Sentence fluency is uneven, and phrasing is sometimes awkward. These issues interfere somewhat with clarity.
3213	11101	8	7	• Essay presents a clear stance against censorship with supporting arguments and examples, though some ideas are repetitive and lack deeper exploration or nuanced development. • Organization includes introduction, body, and conclusion, but transitions are uneven and some paragraphs meander, reducing the overall clarity and impact of progression. • Language shows good control overall, but errors in spelling and grammar (‘Weather’ vs. ‘Whether,’ ‘twards,’ subject-verb agreement) slightly disrupt flow. Vocabulary is basic and could be more varied.

Table 5: Parameter estimates for checklist descriptors

Descriptor	π^*	r^*_{CON}	r^*_{ORG}	r^*_{GRM}	r^*_{VOC}	r^*_{MCH}
D01	0.99	1.00	0.98	0.78		
D02	0.99		0.81	0.65		
D03	0.81				**0.25**	
D04	0.99	0.81	0.59			
D05	0.95	0.88	**0.16**			
D06	0.99	0.82	**0.08**			
D07	0.99		**0.03**			
D08	0.95	**0.17**				**0.37**
D09	0.99		**0.46**			
D10	0.95				**0.07**	
D11	0.93	**0.08**				
D12	0.98		**0.25**			
D13	0.99	**0.11**				
D14	0.96				**0.04**	
D15	0.97	0.80	**0.43**	**0.35**		0.84
D16	0.99		1.00	**0.28**		
D17	0.99			**0.39**		
D18	0.88			**0.05**		**0.32**
D19	0.99		0.98	**0.23**		
D20	0.99		0.98	**0.33**		
D21	0.98			**0.31**		
D22	0.99			**0.22**		
D23	0.96			**0.13**		
D24	0.78			1.00		**0.15**
D25	0.99			**0.30**		0.89
D26	0.31				**0.02**	1.00
D27	0.80				**0.09**	
D28	0.99		0.92	0.58		0.98
D29	0.98		0.79	**0.24**		
D30	0.94		0.71	**0.29**		0.87
D31	0.54	**0.46**	1.00			**0.28**
D32	0.99		1.00			**0.34**
D33	0.95			**0.13**		
D34	0.97		0.83	0.56		0.98
D35	0.99		0.93	0.82		0.99

Notes. $\gamma*$ estimates with strong discriminant power (< 0.5) are highlighted in bold.

Automated Diagnosis of Students' Number Line Strategies for Fractions

Zhizhi Wang
Rutgers University
zw393@cs.rutgers.edu

Dake Zhang
Rutgers University
dake.zhang@gse.rutgers.edu

Min Li
University of Washington
minli@u.washington.edu

Yuhan Tao
Columbia University
yt2905@tc.columbia.edu

Abstract

Fraction number lines are widely recognized as an effective tool for supporting students' conceptual understanding of rational numbers, yet their abstract nature makes them challenging for students to use and for teachers to evaluate reliably. This study introduces Visual Translator (VT), an AI-based tool designed to automatically diagnose students' problem-solving strategies and error types in handwritten fraction number line tasks. VT employs object detection models trained on a curated subset of 1,134 annotated student responses from the ASSISTments Mathnet dataset, and generates structured text descriptions of key visual elements, which can then be processed by large language models (LLMs) for higher-level reasoning. Preliminary results indicate that VT outperforms GPT-4o and Grok in fraction value identification (Precision = 0.611, Recall = 0.582), while achieving substantially higher accuracy in location detection (mAP@50 = 0.88) compared to Gemini-2.5 Pro (0.11). These findings demonstrate the promise of combining computer vision with generative AI to improve automated diagnosis of students' mathematical strategies.

1 Introduction

Fractions represent a critical transition in mathematics learning, serving as a bridge between whole numbers and rational numbers and laying the foundation for later understanding of proportional reasoning, algebra, and real numbers (Siegler et al., 2011; Siegler and Pyke, 2013). However, decades of research have shown that students frequently struggle with fractions, often treating them as two whole numbers rather than as magnitudes on a continuous scale (Ni and Zhou, 2005). To address these difficulties, instructional researchers have emphasized the use of number lines as a visual and conceptual tool for representing fractions. A number line highlights relative magnitude, density, and equivalence of rational numbers, making it particularly effective for developing conceptual understanding (Stewart et al., 2008). As such, number lines have been widely recommended in standards and curricula, including the Common Core State Standards for Mathematics (Association et al., 2010).

Despite their benefits, number lines also present challenges for students. Siegler and colleagues identified two primary strategies (Siegler et al., 2011) used by students: segmentation strategies, which involve partitioning the interval between 0 and 1, and numerical transformation strategies, which involve mapping fractions onto known reference points. Errors commonly observed include uneven segmentation, the use of incorrect units, or misapplied transformations (Bright et al., 1988; Lamon, 2007). These errors are not merely procedural slips; they reflect deeper misconceptions about the nature of rational numbers (Lamon, 2007). For teachers, especially novice teachers, diagnosing these misconceptions from handwritten number line representations is both cognitively demanding and time-consuming (Zhang et al., 2016). Consequently, there is a pressing need for scalable tools that can assist teachers in analyzing student responses and identifying error types with accuracy and consistency.

The rapid advancement of artificial intelligence (AI) provides new opportunities to address this need. Automated scoring systems have demonstrated success in domains such as essay grading, short-answer evaluation, and mathematical problem solving (Lockwood, 2014; Dikli, 2006). Recent work has also explored the use of computer vision and large language models (LLMs) to interpret drawn diagrams and models. For example, Lee and Zhai reported limited success in using GPT-4o to grade student-drawn science models, with accuracy ranging from 0.2 to 0.6, highlighting the challenges of reliably recognizing children's handwritten and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 178–184
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

diagrammatic representations (Lee and Zhai, 2023). Similarly, early experiments with GPT-4o and related multimodal models suggest that, while LLMs excel in natural language reasoning, their image-processing capabilities remain insufficient for fine-grained educational diagnostics such as interpreting number lines.

In this study, we evaluate VT against both human-annotated ground truth and state-of-the-art LLMs (GPT-4o, o3, Gemini-2.5, Grok) on two key tasks: (a) recognizing handwritten fraction values, and (b) detecting the locations of visual elements on number lines. Our preliminary results show that VT achieves substantially higher accuracy in location identification (mAP@50 = 0.88 vs. 0.11 for Gemini-2.5 Pro) and competitive accuracy in fraction value recognition, outperforming GPT-4o and Grok. Beyond empirical results, our contributions are threefold:

- We manually labeled over one thousand student responses featuring fraction number lines, creating a domain-specific dataset with fine-grained annotations of key visual elements (ticks, endpoints, and handwritten values).
- We trained the Visual Translator (VT) on this dataset, tailoring it specifically for fraction number line tasks to detect and interpret key visual information from students' handwritten solutions.
- We designed metrics to assess the accuracy of models in capturing key visual information in students' work, and conducted extensive experiments comparing VT with leading multimodal LLMs, including GPT-4o, GPT-o3, Gemini-2.5 Pro, and Grok-2.

2 VT Model

2.1 Data Preparation

The first step in developing the VT model was to identify student responses containing fraction number lines. From the full MathNet dataset of 3.8 million images, we initially filtered 139,000 fraction-related items using keywords extracted from the associated JSON metadata. From this subset, we manually labeled a small number of images that clearly contained number lines to serve as seed data.

To expand the labeled dataset efficiently, we trained a YOLOv8 model on the seed images to detect number lines and applied it to additional candidate images. Predictions from the model were then manually verified to confirm their relevance. This iterative process, in which the model guides the selection of images for human annotation, effectively implements an active learning strategy, concentrating labeling effort on the most informative samples and improving data collection efficiency.

Table 1: Statistics of Key Elements.

Key Elements Type	Number of Instances
Fraction	8199
Tick	8385
0	2474
1	4447
2	2236
3	2054
4	2182
5	1732
6	2387
7	606
8	1204
9	476

After identifying 1,134 confirmed images featuring 0–1 fraction number lines, we conducted fine-grained annotations of key visual elements, which are identified by our educational experts, including tick marks, digits(0-9), and fractions. Finally, the dataset contains more than 8,000 fraction labels, over 8,300 tick marks, and thousands of digit labels (0–9). Detailed statistics of the labeled dataset are summarized in Table 1. All annotations were completed by our graduate assistants using Roboflow[1], a comprehensive platform for data annotation, model training, and deployment. Each key element was enclosed within a bounding box of a distinct color and assigned a unique label, as shown in Figure 1. The platform allows export of labeled information into various formats, including .txt, .json and other supported formats, enabling users to directly download the annotation files. These annotations serve as the foundation for subsequent model training, evaluation, and automated diagnostic tasks.

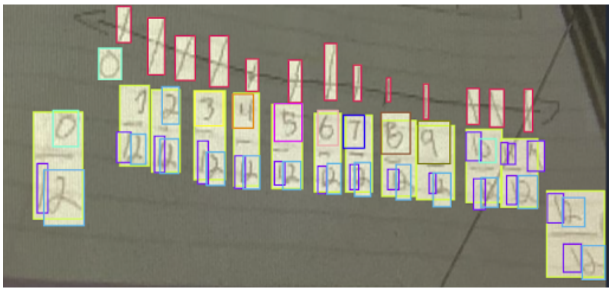


Figure 1: An example for labeling work.

---

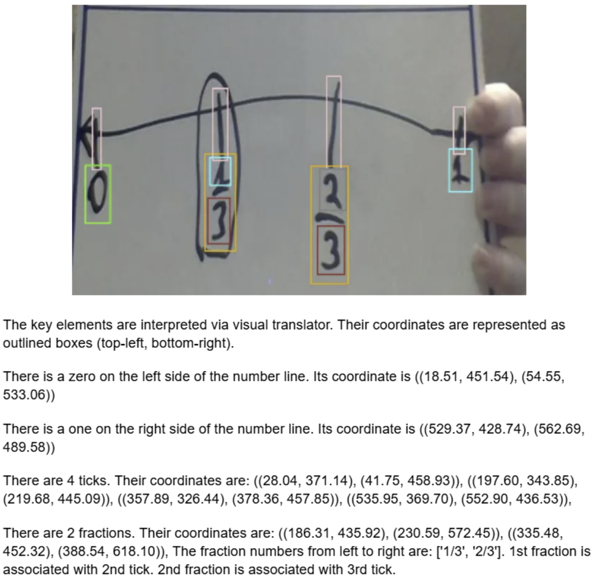
[1] `https://roboflow.com`

2.2 Textual Description Generation

After preparing and annotating the dataset, the VT model was trained to generate textual descriptions of a student's work from a given input image. For each image, the model produces a description that includes the key visual elements, their corresponding labels, and spatial coordinates. Additionally, the model identifies whether the number line is a 0–1 number line by detecting the presence of leading 0 and ending 1.

When a key element corresponds to a fraction, the model further derives the fraction value from the detected digits in the bounding box of a fraction. Specifically, a post-processing clustering algorithm groups digits associated with each fraction into two sets, corresponding to the numerator and denominator. The final fraction value is then computed from these clusters, ensuring accurate reconstruction of the student's intended fraction. Besides that, for a detected fraction, it is linked to its corresponding tick (if present) in the form of "F0-T1", where F0 denotes the first fraction and T1 denotes the second tick. The indices of fractions (i.e., F0, F1, F2 . . .) and ticks (T0, T1, T2 . . .) were automatically generated based on the left-to-right order of the top-left coordinates of their bounding boxes.

Overall, this process provides a structured textual summary of the detected key elements in a student's response, including their spatial information and numerical content. These textual summaries serve as standardized inputs for downstream diagnostic tasks, enabling automated error analysis and strategy classification. An example for textual generation is shown in Figure 2.



```
The key elements are interpreted via visual translator. Their coordinates are represented as
outlined boxes (top-left, bottom-right).

There is a zero on the left side of the number line. Its coordinate is ((18.51, 451.54), (54.55,
533.06))

There is a one on the right side of the number line. Its coordinate is ((529.37, 428.74), (562.69,
489.58))

There are 4 ticks. Their coordinates are: ((28.04, 371.14), (41.75, 458.93)), ((197.60, 343.85),
(219.68, 445.09)), ((357.89, 326.44), (378.36, 457.85)), ((535.95, 369.70), (552.90, 436.53)),

There are 2 fractions. Their coordinates are: ((186.31, 435.92), (230.59, 572.45)), ((335.48,
452.32), (388.54, 618.10)), The fraction numbers from left to right are: ['1/3', '2/3']. 1st fraction is
associated with 2nd tick. 2nd fraction is associated with 3rd tick.
```

Figure 2: A demo example for textual generation.

2.3 Model Development

We developed the VT model in the following pipeline:

2.3.1 Key Element Detection

We trained an object detection model from the YOLO series to identify ticks, digits, and fractions. Training was conducted on the Roboflow platform, which provides resources optimized for small object detection in real-time. The labeled dataset enabled the model to learn the visual appearance and spatial layout of key elements. Our best-performing model achieved a mean Average Precision at IOU 0.5 (mAP@50) of 0.88 on the validation set, demonstrating high accuracy in detecting fine-grained handwritten components.

2.3.2 Model Deployment

The trained VT model is deployed to detect key elements—ticks, digits, and fractions—while returning their corresponding labels and spatial coordinates. The deployment is hosted on Roboflow and accessible via an API, which allows external systems to directly query the model. This design enables smooth integration into various downstream applications without requiring local installation or complex setup.

2.3.3 Web-based Interface

To further enhance usability, we developed a web application hosted on Hugging Face Spaces, offering an interactive interface for educators and researchers. Through this platform, users can:

- Upload an image of a student's work.
- Visualize detection results superimposed on the original image.
- Automatically reconstruct fraction values by clustering detected digits into numerators and denominators.
- Generate a textual summary of all identified key elements along with their coordinates.
- Download the complete results as a JSON file for integration into other pipelines.

The web service is hosted on Hugging Face Spaces and can be accessed at MathNet VT Model Web Platform. Access is granted via the invitation token **RU_MATHNET_VT**.

3 Evaluation

Unlike conventional object detection models, the VT system is designed not only to detect visual elements but also to generate structured textual descriptions that capture the key information in students' handwritten number line tasks. While standard detection metrics such as mAP@50 provide useful references, they are insufficient to directly reflect performance on our educational task, where the ultimate goal is to recover meaningful mathematical content (e.g., fractions, their values, and their associations with ticks).

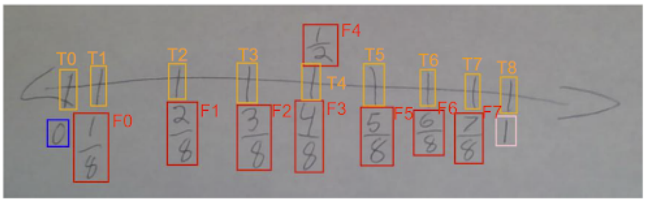
To address this gap, we manually annotated additional ground truth information–including fraction values ordered from left to right according to the bounding box locations of fractions, as well as the associations between fractions and their corresponding ticks–and developed a set of tailored evaluation metrics that complement traditional object detection measures. These metrics assess (i) the accuracy of recognized fraction values and (ii) the correctness of their associations with ticks. Finally, we designed a weighted composite score that integrates our custom metrics with conventional detection metrics. This composite score serves as a more intuitive and holistic evaluation of the performance of VT, balancing both visual detection accuracy and the recovery of meaningful mathematical content from student work.

3.1 Additional Ground Truth Construction

In addition to the bounding box annotations used for model training, two further steps were conducted to establish reliable ground truth:

- **Fraction Value Annotation.** Expert graduate assistants manually read each student's handwritten work and labeled fraction values (e.g., 1/8, 2/8, 3/8). These values were aligned with bounding boxes to create a verified mapping between visual regions and fraction numbers.
- **Fraction–Tick Association.** Fractions were linked to their corresponding ticks (if present) using index pairs (e.g., F0–T1), where indices were assigned based on the left-to-right order of their bounding boxes.

As illustrated in Figure 3, this multi-level annotation ensured that ground truth covered both fraction recognition and structural relationships in the number line.



fraction values: 1/8, 2/8, 3/8 4/8, 1/2, 5/8, 6/8, 7/8
relationship between fractions and ticks: F0-T1, F1-T2, F2-T3, F3-T4, F4-T4, F5-T5, F6-T6, F7-T7

Figure 3: Additional ground truth annotation.

3.2 Metrics for Fraction Value Identification

To evaluate fraction recognition, we designed two complementary accuracy indices:

- **Jac Index (Order-independent).** Measures the set overlap between predicted and ground-truth fraction values, ignoring their order. It reflects the completeness of detection.
- **Seq Index (Order-sensitive).** Measures the length of the longest subsequence of correctly predicted fractions that also appear in the correct order. This is stricter than the Jac Index, as it penalizes out-of-order predictions.

In addition, we computed precision and recall for fraction values:

$$\text{Precision} = \frac{\#\text{ of correctly detected fraction values}}{\#\text{ of all detected fraction values}}$$

$$\text{Recall} = \frac{\#\text{ of correctly detected fraction values}}{\#\text{ of all ground-truth fraction values}}$$

For example, assume the ground truth fractions are 0/3, 1/3, 2/3, 3/3, If the model predicts 0/3, 3/3, 2/3, then three of the predicted fractions are correct under the jac index metric. In this case, the model achieves a precision of 1.0 (since three out of three predictions are correct) and a recall of 0.75 (since it misses one ground-truth fraction, 1/3). While under the seq index, its precision is only 0.67 because 2/3 is out of the right order and its recall is 0.5 (since it misses two ground-truth fractions, 1/3 and 2/3).

3.3 Metrics for Relationships between Fractions and Ticks

In addition to evaluating individual fraction values and key element locations, a crucial aspect of analyzing students' number line work is capturing the spatial and logical relationships between fractions and their corresponding ticks. Correctly identifying these relationships ensures that each fraction is accurately mapped to its intended position on the number line, which is essential for subsequent diagnostic analyses.

To assess this, we introduce relationship-specific metrics that compare the associations generated by VT with ground-truth annotations. Each fraction in the ground truth is linked to a specific tick (if present) using index pairs (e.g., F0–T1), where the indices are determined based on the left-to-right ordering of the top-left coordinates of their bounding boxes. VT's predicted fraction–tick pairs are then matched against these ground-truth pairs. Accuracy is calculated as the proportion of correctly identified fraction–tick relationships over all annotated pairs. In addition, we compute precision and recall for these relationships to provide a more detailed assessment of VT's performance in capturing fraction–tick associations The overall fraction–tick score is summarized by its F1-score, which we will introduce later.

3.4 Composite Score

To provide a single, interpretable measure of VT's overall performance, we designed a composite score that integrates both elemental and relational information extracted from student work. To account for both precision and recall, we adopt the F1 score as a comprehensive performance metric, which is computed as follows:

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The textual description score is calculated as a weighted average of these per-element F1 scores:

$$\text{Textual Description Score} = \sum_i 100 * F1_i * W_i$$

where i indexes the key element categories and W_i denotes their user-defined weights and $\sum_i W_i = 1$. This formulation provides flexibility, as users can assign greater importance to specific elements (e.g., fractions over digits) depending on instructional or diagnostic needs. In addition, detailed per-element results (precision, recall, and F1) are retained to facilitate fine-grained analysis.

To achieve a comprehensive assessment, we integrate the textual description score with the accuracy of fraction–tick relationship detection. Specifically, the final composite score is computed as:

$$\begin{aligned}\text{Composite Score} = {} & W_{ft} * \text{Fraction-Tick Score} \\ & + W_{ts} * \text{Textual Description Score}\end{aligned}$$

where W_{ts} is the weight ratio of textual description score, W_{ft} is the weight ratio of fraction-tick score and $W_{ts} + W_{ft} = 1$. Since fraction–tick relationships depend on the accurate detection of both fractions and ticks, missing either element directly leads to a missing relationship. Therefore, it is advisable to assign a much higher weight W_{ts} than to W_{ft}, e.g. $W_{ts} = 0.8$ and $W_{ft} = 0.2$. By combining elemental accuracy and relational correctness into a single metric, this composite score provides an interpretable and holistic evaluation of VT's effectiveness in extracting both low-level visual details and higher-order relational information from student work.

4 Experiments

We conducted a series of experiments to evaluate the effectiveness of VT in comparison with state-of-the-art large language models (LLMs). Specifically, we first benchmark VT against Grok-2, Gemini-Pro 2.5, GPT-4o, and GPT-o3 on the task of fraction value detection, where accuracy serves as the evaluation metric. In addition, we assess the composite score of VT and Gemini-Pro 2.5, as Gemini-Pro 2.5 is the only publicly available LLM known to support image segmentation, i.e., the ability to process element-level bounding box coordinates.

4.1 Accuracy on Fraction Values Detection

To ensure fairness, the evaluation was conducted on a set of 227 images sampled from both the validation and test datasets, none of which were used to train VT. All models were tested under the same experimental settings.

The results are shown in Table 2. Gemini-2.5 Pro achieved the highest precision and recall across both the Jaccard index and sequential index metrics, demonstrating its strong capability in fraction detection. However, VT consistently outperformed GPT-4o and Grok-2,though its performance was slightly lower than that of GPT-o3. Overall, all models performed somewhat worse on the sequential index than on the Jaccard index, suggesting that capturing the correct ordering of elements remains a challenging aspect.

4.2 Comprehensive Evaluation

The comprehensive evaluation was performed on a combined set of 177 images drawn from our validation and test data. We evaluated both models using our proposed composite score metric, with

Table 2: Accuracy results on fraction values detection.

Model	precision (jac index)	recall (jac index)	precision (seq index)	recall (seq index)
Grok-2	0.320	0.387	0.243	0.293
GPT-4o	0.459	0.408	0.357	0.332
Gemini-2.5-pro	**0.646**	**0.668**	**0.520**	**0.538**
VT	0.560	0.499	0.446	0.403
GPT-o3	0.594	0.527	0.490	0.446

Table 3: Composite scores of VT and Gemini-2.5Pro

Model	VT	Gemini-2.5Pro
Composite score	**66.8**	15.8
Textual description score	**73.0**	18.1
Fraction-Tick score	**41.9**	6.2

the detailed results presented in Table 3, including the composite score along with textual description score and fraction-tick score. The Textual description score itself is a weighted average reflecting the accurate detection of key elements, with weights defined as 'fraction': 0.5, 'tick': 0.4, 'one': 0.05, 'zero': 0.05. The final Composite score is then calculated by combining the Textual description score (representing key element detection, 'ke') and the Fraction-Tick score (representing relationship between a fraction and its corresponding tick, 'tick2frac') with weights of $W_{ke} = 0.8$ and $W_{tick2frac} = 0.2$, respectively.

As illustrated in Table 3, our VT model demonstrates a commanding lead across all metrics. VT achieved a composite score of 66.8, which is more than four times higher than the 15.8 scored by Gemini-2.5Pro. This significant gap is consistent across the sub-metrics: VT scored 73.0 on textual description and 41.9 on fraction-tick relationships, compared to Gemini-2.5Pro's scores of 18.1 and 6.2.

The stark performance disparity underscores the critical importance of domain-specific training for specialized, high-precision tasks. While LLMs like Gemini-2.5Pro possess extensive general knowledge, they struggle to accurately parse the fine-grained, structured information required by our task without targeted fine-tuning. This outcome strongly indicates that Large Language Models (LLMs) do not serve as an infallible "oracle" or a universal solution for all problems.

5 Conclusion

In this paper, we presented VT, a specialized vision-language model designed to parse key semantic information from student-produced diagrams of fraction number lines. Departing from conventional object detection methods that focus primarily on localization, VT generates a structured textual representation that encapsulates not only elemental components (e.g., digits, ticks, fractions) but also their crucial relational associations. To facilitate a rigorous and fair evaluation, we have contributed a manually annotated dataset of over 1,000 student drawings and proposed a suite of tailored metrics, including accuracy on fraction values, fraction-tick relationship metrics, and a weighted composite score that provides a comprehensive assessment of model performance. Our empirical results demonstrate that VT significantly outperforms general-purpose Large Language Models. This finding suggests that while LLMs offer broad capabilities, they are not a universal solution; for domain-specific tasks requiring fine-grained interpretation of private data, developing and training specialized models remains a necessary and effective approach for robust information extraction.

References

National Governors Association and 1 others. 2010. Common core state standards. *Washington, DC*.

George W Bright, Merlyn J Behr, Thomas R Post, and Ipke Wachsmuth. 1988. Identifying fractions on number lines. *Journal for research in mathematics education*, 19(3):215–232.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Susan J Lamon. 2007. Rational numbers and proportional reasoning: Toward a theoretical framework for research. *Second handbook of research on mathematics teaching and learning*, 1(1):629–668.

Gyeong-Geon Lee and Xiaoming Zhai. 2023. Nerif: Gpt-4v for automatic scoring of drawn models. *arXiv preprint arXiv:2311.12990*.

Jane Lockwood. 2014. Handbook of automated essay evaluation current applications and new directions mark d. shermis and jill burstein (eds.)(2013). *Writing & Pedagogy*, 6(2):437–442.

Yujing Ni and Yong-Di Zhou. 2005. Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational psychologist*, 40(1):27–52.

Robert S Siegler and Aryn A Pyke. 2013. Developmental and individual differences in understanding of fractions. *Developmental psychology*, 49(10):1994.

Robert S Siegler, Clarissa A Thompson, and Michael Schneider. 2011. An integrated theory of whole number and fractions development. *Cognitive psychology*, 62(4):273–296.

James Stewart, Lothar Redlin, and Saleem Watson. 2008. *Precalculus: Mathematics for calculus*, 5 edition. Brooks/Cole.

Dake Zhang, Pamela Stecker, Sloan Huckabee, and Rhonda Miller. 2016. Strategic development for middle school students struggling with fractions: Assessment and intervention. *Journal of Learning Disabilities*, 49(5):515–531.

Medical Item Difficulty Prediction Using Machine Learning

Hope Adegoke[1], **Ying Du**[2], **Andrew Dwyer**[2]
[1]Educational Research Methodology, University of North Carolina, Greensboro.
[2]American Board of Pediatrics, 111 Silver Cedar Court, Chapel Hill, NC 27514.
[1]`hoadegoke@uncg.edu`, [2]`ydu@abpeds.org`, [2]`adwyer@abpeds.org`

Abstract

This study examines the prediction of medical exam item difficulty using NLP and machine learning. A dataset of 1,154 MOCA-Peds items was represented with lexical features, cohesion metrics (TAACO), blueprint encodings, BioMedBERT embeddings, semantic similarity, and unsupervised semantic indicators. Regression models predicted IRTb parameters with RMSE ≈ 1.37 and R² $\approx .29$, a $\sim 15\%$ improvement over baseline. Ordinal classifiers reached 30–34% accuracy, doubling the baseline (17%), with $\sim$68% adjacent accuracy. Feature importance showed blueprint, item structural features (number of options, option length), and semantic embeddings as the strongest predictors. These findings demonstrate the feasibility of difficulty prediction in medical assessments and highlight the value of combining domain-specific semantics with latent difficulty signals for efficient test development.

1 Introduction

Having an idea of the difficulty of items before operational use supports efficient test construction, helps to achieve target score precision, and reduces the cost and security risks of pre-testing (Settles et al., 2020). Previous work using text features alone has shown mixed results, often only modestly surpassing naive baselines (Štěpánek et al., 2023). Shared-task evidence suggests that transformers can help when carefully tuned and combined with complementary features (Yaneva et al., 2024; Gombert et al., 2024). In medical licensing contexts where items require specialized knowledge and clinical reasoning, prediction of difficulty has been particularly challenging (Ha et al., 2019). We extend this literature by integrating different categories of linguistic and semantic feature families, including domain-specific biomedical embeddings (Gu et al., 2021), and by encoding content blueprints, then evaluating both continuous and ordinal formulations of difficulty.

2 Related Works

Early studies of automated item difficulty prediction focused on surface-level text features such as length, readability, and lexical counts, but these typically explained little variance (McNamara et al., 2014; Štěpánek et al., 2023; Ha et al., 2019). With the rise of NLP, transformer-based models achieved top performance in the BEA 2024 Shared Task on difficulty prediction (Gombert et al., 2024), while methods using traditional linguistic/cohesion features (including clinical embeddings and principal component features) also showed competitive performance though generally lower than the top transformer-based approaches (Yaneva et al., 2024; Tack et al., 2024).

Medical assessments present unique challenges. Ha et al., 2019 reported only modest gains over baseline for medical MCQs, reiterating the importance of incorporating domain expertise. Domain-specific embeddings like BioMedBERT (Gu et al., 2021) provide richer representations of medical terminology. Furthermore, content-based encodings such as blueprints or cognitive-level taxonomies (Anderson and Krathwohl, 2001) can serve as structured priors for difficulty. Our work extends these lines by combining linguistic, semantic, and content features, evaluating both continuous prediction of IRTb and ordinal classification into difficulty bins.

3 Method

3.1 Data & Targets

We analyze 1,154 multiple-choice items from the Maintenance of Certification Assessment for Pediatrics (MOCA-Peds), a longitudinal, online assessment that allows pediatricians to demonstrate ongoing knowledge through periodic testing rather

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 185–190
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

than a single high-stakes exam (Leslie et al., 2018). Each item consists of a clinical stem and four or five answer options (A–D or A–E). Every item has a calibrated IRT difficulty parameter (b), which serves as the outcome variable for the prediction. In the regression setting, the task is to predict the continuous IRTb value directly. In the classification setting, items are grouped into five ordinal difficulty categories by dividing the IRTb distribution into quintiles, ranging from the easiest to the hardest items. Because there are five equally sized bins, random guessing of the algorithms would yield a uniform-chance accuracy of 20% for predicting the exact difficulty class.

3.2 Features

We construct five feature families:

Lexical features: Stem and option lengths (in both characters and tokens), punctuation counts, total number of options (4 and 5), and lexical overlap indices. Overlap was measured with Jaccard similarity between the stem and the correct option, and between the stem and the distractors.

TAACO cohesion features: Cohesion and readability indices computed with the Tool for the Automatic Analysis of Cohesion - TAACO (Crossley et al., 2016, 2019). These include lexical diversity, referential overlap, connectives usage, and standard readability metrics.

BioMedBERT semantic features: Mean-pooled contextual embeddings of the stem and options from BioMedBERT (Gu et al., 2021). From these embeddings, we derived cosine similarities (stem–correct option and stem–distractors), dispersion among option embeddings (mean pairwise cosine distance), and principal components of the stem embeddings to provide lower-dimensional semantic factors.

Blueprint encoding: Smoothed target encoding of Level-2 content domains from the MOCA-Peds blueprint. Each item's category was assigned the smoothed mean difficulty from training folds, providing a structured content-based prior.

Unsupervised difficulty features: Embedding-derived features that do not use the target difficulty, such as stem "uniqueness" relative to the corpus, cluster-based indicators (cluster size, distance to centroid), and coherence/variance scores. These are designed to capture latent difficulty signals without relying on calibrated b parameters.

3.3 Modeling & Validation

We approach difficulty prediction through two complementary pipelines: regression for continuous IRTb values and ordinal classification for quantile-based difficulty bins. For both pipelines, we used a nested cross-validation strategy. In the outer loop, five folds of data ensured every item served once as test data to estimate generalization error. Within each training partition, randomized hyperparameter search with inner cross-validation selected the best model configuration. This setup prevents information leakage between tuning and evaluation, which produces a robust and unbiased performance estimates (Varma and Simon, 2006; Arlot and Celisse, 2010). For regression, we combined interpretable linear models with flexible non-linear methods. Elastic Net and Ridge provided regularized linear baselines (Štepánek et al., 2023), while Random Forest and gradient boosting methods (XGBoost, LightGBM, CatBoost) captured non-linear interactions. Evaluation emphasized Root Mean Square Error (RMSE), supplemented by Mean Absolute Error (MAE), R^2, and Spearman rank correlation to reflect both the magnitude and the ranking of difficulty. Calibration slopes were also inspected to assess systematic under- or overestimation.

RMSE (Root Mean Square Error): average magnitude of prediction errors, penalizing larger errors more heavily.

MAE (Mean Absolute Error): average absolute difference between predicted and true values.

R^2 (Coefficient of Determination): proportion of variance in item difficulty explained by the model.

Spearman's ρ: correlation between predicted and true rankings of item difficulty, reflecting how well the model preserves ordering.

Calibration Slope: regression slope of observed versus predicted difficulty; values close to 1 indicate unbiased, well-scaled predictions.

For classification, we tested Logistic Regression, SVMs, and Random Forest, consistent with prior work on medical MCQ difficulty prediction (Ha et al., 2019). We additionally evaluated gradient boosting classifiers (XGBoost, LightGBM, CatBoost), which have been shown to be competitive in recent difficulty prediction tasks (Yousefpoori-Naeim et al., 2024). We reported overall accuracy, macro-F1, Ordinal Mean Absolute Error, Adjacent Accuracy (crediting predictions within one difficulty level), and Top-2 Accuracy. Such metrics are recommended in ordinal classification/assessment prediction settings, though prior work uses subsets of them (Ha et al., 2019; Gombert et al., 2024).

Accuracy: proportion of items where the predicted bin exactly matched the true bin.

Macro-F1: unweighted average of F1-scores across bins, combining precision and recall.

Ordinal MAE: mean absolute difference between predicted and true bin indices, capturing distance on the ordered scale.

Adjacent Accuracy: percentage of predictions that were exact or within one difficulty level.

Top-2 Accuracy: percentage of items where the correct bin was among the model's two highest-scored predictions.

All models were implemented in standardized Python pipelines with preprocessing, scaling, and leakage-safe encoding. Embeddings were precomputed and cached to ensure efficiency, and fixed seeds were used for reproducibility.

4 Results

4.1 Regression (predicting IRTb)

In the regression task, models were trained to predict continuous IRT b parameters directly from item features. The baseline predictor, which always returned the mean item difficulty, yielded RMSE $\approx$ 1.63 and explained virtually none of the variance ($R^2 \approx 0$). In contrast, all feature-based models substantially outperformed this baseline. As shown in Table 1, the best results were achieved by Elastic Net and Random Forest, which attained RMSE $\approx$ 1.37 and explained about 28–29% of variance ($R^2 \approx 0.28$–0.29). Both also showed moderate rank-order correlations with true difficulty ($\rho \approx 0.45$), indicating that they not only approximated difficulty values but also captured relative ordering among items. Gradient boosting methods (LightGBM and CatBoost) performed nearly as well, with RMSE $\approx$ 1.37 and $R^2 \approx 0.288$. Ridge regression was slightly weaker ($R^2 \approx 0.276$), while XGBoost lagged behind with the highest RMSE ($\approx$ 1.40) and the lowest explained variance ($R^2 \approx 0.26$).

Calibration analysis confirmed that model predictions were well aligned with observed values: Random Forest achieved a slope close to 1.0, while Elastic Net slightly underestimated extreme difficulties (slope $\approx$ 1.1). In practical terms, given the observed range of b values ($\approx$ –5 to +5), the error reduction from 1.63 to 1.37 translates into roughly a 15% gain in predictive precision from the baseline.

4.2 Ordinal classification (5 difficulty bins)

For the 5-class ordinal classification task, models substantially outperformed the baselines. The majority-class baseline reached only 17.2% accuracy, while a uniform random predictor would achieve $\approx$20% accuracy by chance.

As shown in Table 2, the Random Forest classifier achieved the strongest performance with 34.5% accuracy, approximately double the majority-class baseline. Its Macro-F1 (0.35) was aligned with accuracy, reflecting fairly uniform performance across difficulty bins. The ordinal-specific metrics confirmed its usefulness: the Ordinal Mean Absolute Error was 1.14 (vs. 1.48 for baseline), and Adjacent Accuracy reached 68%, indicating that two-thirds of predictions were either exact or within one difficulty level. The Top-2 accuracy of 54% further shows that the true class was frequently among the two highest-scored bins.

LightGBM and XGBoost followed closely (overall accuracy $\approx$0.34 and $\approx$0.34 respectively), while SVM (RBF kernel) and Logistic Regression trailed modestly ($\approx$0.30–0.32 overall accuracy). Importantly, even the weaker models still exceeded baseline performance, confirming that item features contain reliable ordinal difficulty signals.

4.3 Feature Importance

We examined feature importance across regression and classification models. Figure 1 show us that for Random Forest regression, the strongest predictors were the blueprint encoding, the number of answer options, and the length of option E. Each accounted for around 10% of the model's explanatory variance, confirming that both content area and item format influence difficulty.

Embedding-based features also played a key role. Several principal components from BioMedBERT stem embeddings and stem–distractor cosine similarities ranked among the top predictors, indicating that semantic complexity and distractor plausibility strongly shaped difficulty. In contrast, cohesion indices from TAACO and traditional readability measures contributed little when richer semantic and content features were available.

For the ordinal classification Random Forest, the same pattern emerged: option E length and blueprint encoding dominated, followed by embedding-derived factors and unsupervised simi-

Table 1: Cross-validated regression results (5-fold outer CV). Lower RMSE is better.

Model	**RMSE**	**MAE**	**R^2**	**Spearman ρ**	**Calib. slope**
Elastic Net	**1.3719**	1.0759	**.2881**	0.4584	1.10
Random Forest	**1.3685**	1.0744	**.2916**	0.4464	0.98
LightGBM	1.3715	1.0763	.2885	0.4577	1.12
CatBoost	1.3717	1.0752	.2882	0.4530	1.07
Ridge	1.3838	1.0852	.2757	0.4400	0.96
XGBoost	1.3976	1.1034	.2612	0.4174	1.04
Baseline (mean)	1.6262	1.2685	$\sim$0.00	$\sim$0.00	—

Table 2: Ordinal classification results (5-fold outer CV).

Model	**Accuracy**	**Macro-F1**	**Ordinal-MAE**	**Adjacent Acc.**	**Top-2 Acc.**
Random Forest	**.345**	**.347**	**1.136**	**.679**	.538
LightGBM	.343	.338	1.187	.657	**.555**
XGBoost	.337	.335	1.156	.677	.536
SVM (RBF)	.322	.324	1.174	.666	.538
Logistic Reg.	.309	.296	1.247	.638	.524
Baseline (maj.)	.172	.150	1.478	.546	.382

larity features (e.g., stem–nearest neighbor cosine distance). This shows that difficulty is not just a function of surface text complexity but is rooted in the interaction of content domain, item structure, and semantic relationships among options.

5 Discussion

Our models achieved modest but meaningful predictive power: regression explained about 28–29% of variance in IRTb, and ordinal classification reached 34% accuracy with 68% adjacent accuracy. These gains over baseline suggest that automated difficulty prediction can support item development, though the strength of the predictions remains limited.

Importantly, our results are consistent with prior studies. Štepánek et al., 2023 reported similar variance explained when predicting reading comprehension item difficulty, and Ha et al., 2019 found only modest gains for medical MCQs. Recent BEA-2024 shared-task findings (Yaneva et al., 2024; Gombert et al., 2024; Tack et al., 2024; Yousefpoori-Naeim et al., 2024) likewise show that even transformer-based systems reach only moderate correlations, underscoring a common ceiling in this line of work.

The drivers of prediction in our study, which are option structure, blueprint encoding, and biomedical embeddings, mirror some of those highlighted in other research (Ha et al., 2019; Tack et al., 2024). Readability and cohesion features offer little contribution once richer, domain-specific features are available. This points to why the ceiling persists: difficulty depends not just on text but also on broader context, reasoning steps, and examinee knowledge, factors not fully captured by textual features.

From a practical perspective, these models are best used for screening and triage: flagging potentially too-easy or too-hard items, or giving item writers feedback about content areas and option structures. They are unlikely to replace psychometric calibration (at least not yet), but can reduce workload and guide review.

Looking ahead, progress will likely come from incorporating richer modalities (stimuli, visuals), domain-adapted embeddings, and design-aware features that better align with the cognitive processes behind item difficulty. Until then, automated prediction should be seen as an assistive tool that complements, rather than substitutes the current process.

6 Conclusion

This study shows that predicting the difficulty of medical multiple-choice items is feasible when models combine diverse linguistic, semantic, and content-informed features. By integrating domain-specific biomedical embeddings and blueprint encodings alongside lexical and cohesion measures, our models achieved measurable improvements over baselines in both continuous and ordinal formulations of difficulty. Importantly, the results highlight that difficulty prediction is not driven by surface text length alone but by deeper signals of what the item is about and how it is structured.

The practical implication is that automated prediction can serve as a support tool in item devel-

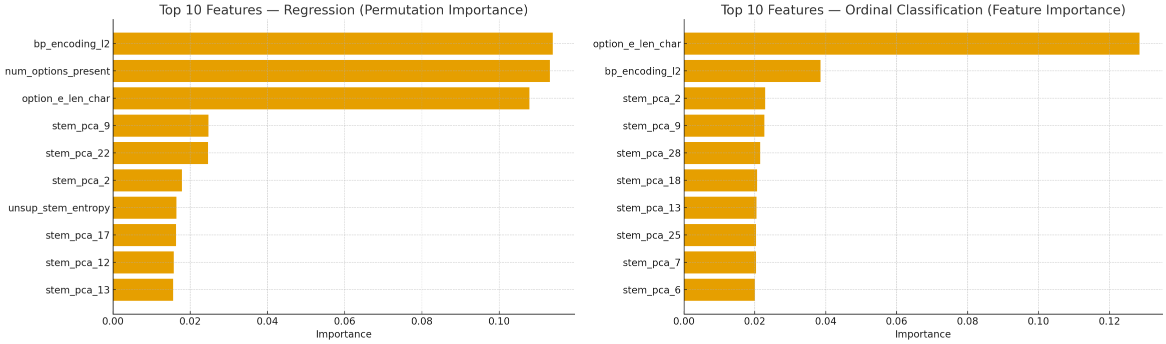


Figure 1: Top 10 predictive features identified by importance analysis. Left: Regression models predicting continuous IRTb. Right: Ordinal classification models predicting difficulty bins.

opment pipelines, sorting items for expert review, guiding test assembly, and reducing reliance on costly pretesting. At the same time, the modest explanatory power and the prominence of dataset-specific signals (such as Option E length) remind us that these models should complement, not replace, expert judgment and psychometric validation.

Future work should extend these methods to larger and more diverse item pools, explore fine-tuned transformer models, and incorporate additional sources of difficulty such as cognitive complexity ratings (Bloom's taxonomy) or multimedia elements. Taken together, the findings provide evidence that machine learning can play a constructive role in modern test development, enhancing efficiency while respecting the central role of human expertise.

7 Limitations and future work

This study was limited by the use of a single dataset of 1,154 MOCA-Peds items, which may constrain generalizability. Some highly ranked predictors, such as Option E length, applied to only a small fraction of items (88/1154) and may reflect dataset-specific patterns rather than universal drivers of difficulty. In addition, the models considered only text and blueprint features, without incorporating multimedia content or group-level differences.

Future work should validate these findings across larger and more diverse item banks, explore explicit cognitive-level annotations, and investigate fine-tuned transformer models trained on exam text. Embedding difficulty prediction into item development workflows to provide real-time feedback to item writers is a promising application.

References

Lorin W. Anderson and David R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. Taaco 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Simon Gombert, Lasse Menzel, Dimitrios 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*, pages 483–492.

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*, 3(1):1–23.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Jenny 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*, pages 11–20.

Laurel K Leslie, Murrey G Olmsted, Adam Turner, Carol Carraccio, Andrew Dwyer, and Linda Althouse. 2018. Moca-peds: development of a new assessment of medical knowledge for continuing certification. *Pediatrics*, 142(6):e20181428.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine-learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. ITEC at BEA 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, Mexico City, Mexico. Association for Computational Linguistics.

Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91.

Victoria Yaneva, Kate North, Peter Baldwin, Le An Ha, Saba Rezayi, Yiming Zhou, Sourav Roy Choudhury, Padma 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*, pages 470–482.

Mohammad Yousefpoori-Naeim, Shirin Zargari, and Zahra Hatami. 2024. Using machine learning to predict item difficulty and response time in medical tests. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 551–560, Mexico City, Mexico. Association for Computational Linguistics.

Lukáš Štepánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19):4104.

Examining decoding items using engine transcriptions and scoring in early literacy assessment

Zachary Schultz and **Mackenzie Young** and **Debbie Dugdale** and **Sue Lottridge**
Cambium Assessment
{zachary.schultz, mackenzie.young, debbie.dugdale, susan.lottridge}@cambiumassessment.com

Abstract

We attempt to improve a transcription-based early childhood speech scoring approach by implementing allowable variations, which are phonemes that can be swapped out for those in a target word to create adjacent words that would receive a passing score. This approach is based off of how a child acquires the ability to pronounce sounds in the English language and how certain phonemes can be easily confused during transcribing, whether done by a human or a machine. Testing both a set of general allowable variations and a set specifically based on the American southern dialect against human scoring, we find that the general allowable variations improve performance, especially at item-specific levels. The performance when utilizing dialect-specific variations does not change much, although this is likely from the use of a word-based transcriber. This approach can be useful for specific words that may have phonemes easily misheard for their voiced or voiceless counterpart (e.g. "have" and "half") but, overall, a modeling approach for training an AI engine would more likely result in higher performance.

1 Introduction

Automated scoring of text-based items is common in K-12 assessment. Automated scoring of speech-based items is used in English Language learner assessments (e.g., Texas English Proficiency Assessment Program) as well as some early literacy screeners (e.g., Soapbox labs, Amira). In automated scoring of speech-based items, automated speech recognition systems are utilized, sometimes taking a transcriber-based approach. Transcribers, whether human or machine, are prone to mistakes, with many human transcribers requiring professional training in order to achieve accurate and quality transcriptions. Even the most reliable, open-source transcriber model, Whisper, can experience vastly different Word Error Rates (WER) depending on the acoustic environment of the audio and the speaker themselves (Kuhn et al., 2024.)

Standard Whisper-based models are not usually trained on child speech and are therefore more prone to errors when transcribing audio of children speaking (Jain et al., 2023.) When utilizing this transcribing approach for speech scoring, in which speech is transcribed and then a rules-based scoring process is applied, one should take into account linguistic features. In particular, for early literacy assessment or the assessment of young children's speech, understanding how children develop their articulatory skills and how phonemes are connected in their place or manner of articulation can contribute to potentially more robust scoring and results that can more accurately inform about a student's speaking ability.

In bridging the gaps between machine scoring, psychometrics, and linguistics, we explore the ways in which one transcription-based approach could be improved by the use of "allowable variations" in early literacy verbal tasks.

2 Background

Early literacy assessments are becoming a critical piece of K-12 large scale assessment to support evidence-based reading instruction (Brunetti et al., 2025). Most early literacy assessments consider reading fluency as a combination of word recognition and language comprehension (Gough & Turner, 1986; Scarborough, 2001; Duke & Cartwright, 2021). Word recognition can be divided into three broad strands: phonological awareness, sight recognition, and decoding, with the latter being the focus of this study.

Decoding is the process of linking printed letters to spoken sounds and includes recognition of phonology, orthography, and morphology (Clemens et al., 2020). During decoding, readers might sound out and blend individual letters

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 191–196
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

into phonemes or combine larger letter groups to form syllables and recognize whole words (Ehri, 2005). Garcia & Cain's meta-analysis (2014) analyzed decoding assessment characteristics and found that the accurate decoding of real words (vs pseudowords) was more predictive of reading comprehension than other measures.

In one decoding item type, students are shown a word and asked to say it aloud. Scoring involves determining whether the student verbalized the target word accurately and whether and what "variations" are allowed, in order to recognize multiple factors that can influence this determination. For instance, words are typically interpreted by humans (and engines) in the context of other words; without this context, both humans and engines can interpret a word slightly differently with both likely being correct representatives. Additionally, the acoustic and linguistic properties of very young children's speech can impact how both humans and engines interpret pronunciation. Acoustically, children's speech falls in a higher register and can have prosodic characteristics that differ from adults. Linguistically, children have underdeveloped articulatory systems and some may struggle to pronounce more advanced English phonemes. This can lead to children replacing a more difficult sound with an easier one (e.g., /r/ vs. /w/). Finally, dialect or regional pronunciations can impact how words are pronounced.

When tests are administered remotely via computer in a classroom setting, testing conditions can also impact scoring (Oberle & Powers, 2025). Often, tests are administered at the same time within a classroom; there can be substantial background noise and chatter, multiple speakers, as well as variations in how loudly or quickly a student speaks. And, young students' ability to interact with the test can also contribute to the demonstration of their decoding skill. The determination of "allowable variations" thus needs to consider these factors relative to each target word and an acceptable pronunciation.

There are three ways to score these items. First, the student speech can be scored by trained human raters using a rubric. Second, humans or machines can transcribe the student speech, and then apply explicit scoring rules. Third, AI systems can be modeled directly on speech to predict human scoring.

In this study, we aim to explore the second approach while addressing the previously stated factors that can add difficulty to this method.

3 Methods

We use data from seventeen decoding items administered across kindergarten and grade one during a Spring 2024 operational field test in one southern state. Students could earn a score of 1 for a correct pronunciation or a score of 0 if incorrect. A correct score required an exact pronunciation with little to no variations allowed. In these data, responses were scored by trained, human raters and a subset (100 per item) was transcribed by both a human and a Whisper-based model trained on adult speech (Radford et al., 2022.)

In transcribing, neither the humans nor the machine had knowledge of the target word for each item. Once transcribed, a score of 0 or 1 was given depending on if the transcription contained the target word, with a score of 1 indicating that the target word is present. Initially, only the target word can trigger a score of 1. Then, the list of acceptable words expands once allowable variations are added.

To determine acceptable variations, we first look at the literature concerning how young children may differ in their pronunciations of various phonemes and how their articulatory systems develop. For example, children develop the ability to pronounce consonants such as /b, p, m, n, h, w, d/ around two years of age, whereas consonants such as /ɹ, ʒ, ð, θ/ are acquired at an age between five and seven years old (Crowe & McLeod, 2020.) Because of this, students may replace one of these later-stage sounds with one they acquired earlier. We also consider manner and place of articulation, with the assumption that phonemes that are close in one or both traits may be misinterpreted when transcribing. In terms of vowels, those that are close to one another in the physical vowel space can be considered as allowable variations.

Allowable variations are determined and listed using the International Phonetic Alphabet (IPA), a collection of symbols each representing one unique possible sound in human speech. Sounds in English are sometimes composed of multiple letters but represented as one symbol using the IPA. For example, the English sound written as "th" is represented in the IPA either as /ð/ or /θ/ depending on if it is a voiced sound or not. Evidently, the IPA allows one to represent a specific sound with one character and it is therefore useful in both representing allowable variations and in implementing them computationally.

Once the list of allowable variations is determined, we then apply each of the variations to the IPA transcription of a given response. To convert a transcription to its IPA representation, we use the Python package 'eng-to-ipa.'[1] Each phoneme within the transcription is swapped out with each of its acceptable variations until all possible combinations have been created. This leads to a long list of non-existent words, so we cross reference the created list with the Carnegie Mellon University (CMU) Pronouncing Dictionary[2] and only include those produced transcriptions that are valid and present in the dictionary. This both limits the number of targets that we are accepting beyond the given and it accounts for the transcribers acting more in a word-based manner than a phonemic manner.

The list of variations can also be modified to fit specific frameworks, such as the phonetic inventory of a dialect. While some variations in the general list we created may apply to a dialect, focusing on specific features of a given dialect should produce a more specific set of variations. In this study, we create a list of variations using features of the standard American southern dialect. Features such as monophthongization, diphthongization, triphthongization, non-rhoticity, the "pin-pen" vowel merger, and the distinction between words such as "which" and "witch" are all incorporated into the list. In generating the variant targets with this list, we skip the cross reference with the CMU Pronouncing Dictionary to maintain all the features of the dialect even if they lead to non-standard words.

With both a general list of acceptable variations[3] created for each target and a list of variations utilizing the features of the American southern dialect, we then rescore by looping through the variation lists alongside the transcriptions. If any of the variations or the original target word are present in the transcription, a score of 1 is given. If the target word nor any of the variations are present, a score of 0 is given. We then use these scores to calculate comparative statistics in order to gauge changes in item performance.

We use three statistical measures to gauge agreement in this study. Firstly, we calculate exact agreement between two sets of scores. Exact agreement is a percentage of the scores that are the same for a given response for both sets. Secondly, we use Quadratic Weighted Kappa (QWK or Kappa, for short). This is another agreement statistic with a more robust calculation which takes into account the possibility that an agreement occurred by chance. It also penalizes disagreements that are further from one another on an ordinal scale; however, this is irrelevant in this study as there are only two possible labels for the data. Finally, we calculate the p-value, which in this case is the mean score. These measures were then averaged across grade level and averaged overall. We compare agreements between the human raters, the first human rater and both types of transcription, and between the transcriptions. The goal in this study is to be comparable to, or better than, the agreement values between the two human raters, which are outlined in Table 1.

Grade	**N**		**H1H2 Scores**		
	Items	Responses	Exact Agr.	K	p-val.
K	3	203	89%	0.74	0.66
1	14	182	93%	0.73	0.81
All	17	186	92%	0.73	0.79

Table 1: Agreement statistics between the two human raters

4 Results

4.1 General Variations

Table 2 provides an overview of the agreement statistics when a score of 1 is strictly given for the target word and no variations are included. One can see that a transcribing method with similar scoring rules to the human raters does not perform as well as humans. The low Kappa values here are primarily due to machine transcription error. This prompted an attempt to improve these results through the use of allowable variations.

Table 3 provides an overview of the agreement statistics when allowable variations are used. These variations are those from a set of general variations. Table 4 shows the agreement statistics when the variations are focused on the American southern dialect.

When using general variations, performance improves. For exact agreements, there is a slight improvement of 1-3%. Similarly, kappa values show slight improvement as well, ranging from an increase of 0.01 to 0.05. Overall, this is not a large difference, but individual items experienced more

---

[1] https://pypi.org/project/eng-to-ipa/

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] See Table 5 for an example of consonant variations utilized in this study.

Grade	N		H1 Score-Human Transcription			H1 Score-Engine Transcription			Human-Engine Transcriptions	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	78%	0.58	0.45	73%	0.51	0.42	84%	0.65
1	14	89	79%	0.45	0.64	72%	0.36	0.57	78%	0.45
All	17	88	79%	0.47	0.60	72%	0.39	0.54	79%	0.49

Table 2: Agreement statistics without the use of variations

Grade	N		H1 Score-Human Transcription			H1 Score-Engine Transcription			Human-Engine Transcriptions	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	76%	0.53	0.52	72%	0.47	0.48	83%	0.66
1	14	89	82%	0.48	0.69	76%	0.40	0.62	79%	0.49
All	17	88	81%	0.49	0.66	75%	0.41	0.59	80%	0.52

Table 3: Agreement statistics with the use of general variations

drastic improvement or decay.

For example, an item with the target word of "have" increased from 65% to 71% in exact agreement and from 0.28 to 0.36 in its kappa value. When examining the data, the acceptable variation "half" appears to be responsible for these increases. This suggests that either students have difficulty distinguishing or pronouncing the voiced and voiceless dental fricatives or that these phonemes sound similar on recording and can be hard to distinguish by listeners and transcribers.

Another item with the target word of "what" experienced similar increases in performance. The exact agreement value increased from 82% to 86% and the kappa value increased from 0.55 to 0.57. This improvement mainly came from vowel variations, especially when swapping the phoneme /ə/ with /ɛ/, which produce the words "what" and "wet" respectively. There are also instances in this case where the word "wood" contributes to performance changes, showing, again, that voiced and voiceless consonant pairs can cause problems for students or listeners.

Finally, the item with the target word of "your" gained large improvements. The exact agreement value increased from 74% to 86% and the kappa value increased from 0.31 to 0.35. In this instance, we produced variations that included the removal of the initial sound and this seemed to lead to this performance improvement. This suggests that initial sounds may be missed in the recording or transcribing process or that words with many minimal pairs may have a higher rate of mistakes by students or listeners and transcribers.

4.2 Dialectical Variations

The use of variations which only involve changes associated with the American southern dialect yielded little change. The only item that experienced significant performance improvement was the item with the target word "been." This is most likely due to the "pin-pen merge" feature of the southern dialect in which the vowels in the words "pin" and "pen" are pronounced almost identically.

Most likely, we do not see drastic performance changes using the southern dialect because transcribing was done using a word-based transcriber, which seeks to output a valid English word if possible. While testing was done with a phonetic transcriber, the output was not reliable. The approach with dialectical variations could be very useful in cases where there are reliable phonetic transcriptions and the use of a specific dialect is well documented for the area of testing.

5 Conclusion

Overall, implementing allowable variations can lead to slight overall performance improvements and item-specific improvements ranging from slight to major. When the variations are a general list, the improvements are higher and more widespread across the items. When they are focused in on a dialect, the improvements are minimal. However, this could be due to the use of a word-based transcriber whereas dialects feature varying pronunciations of a word with one acceptable spelling.

With only slight improvement overall, this approach may only be useful in cases where a specific word is being used and it is likely that a phoneme within the word will lead to transcribing errors. This most often seems to be pairs of voiced and voiceless phonemes but can also occur when a word has many minimal pairs.

The other avenue to take when attempting to improve these agreements is to use high-quality,

Grade	N		H1 Score-Human Transcription			H1 Score-Engine Transcription			Human-Engine Transcriptions	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	78%	0.58	0.45	73%	0.51	0.42	84%	0.65
1	14	89	80%	0.46	0.65	74%	0.37	0.59	78%	0.49
All	17	88	80%	0.48	0.61	74%	0.40	0.56	79%	0.52

Table 4: Agreement statistics with the use of dialectical variations

human transcriptions and scores and train an AI engine using them. After looking at the results from this study, we believe this would be the recommended route to take if possible.

Limitations

In this study, the sample sizes per item were fairly small and may not have been representative of the student population. We also used a word-based transcriber which has more difficulty in reporting dialectical features of speech.

Acknowledgments

I am deeply grateful to my coauthors for their help in every aspect of conducting this study. Their expertise, time, and dedication were essential to completing this work. I would also like to thank Amy Burkhardt for her guidance and leadership, not just in this work but in all that I do. Finally, a huge thank you to the rest of the ML Team at Cambium Assessment for their leadership and innovation in this field. I am eternally thankful to be a part of the team and the important work we do.

References

Matthew Brunetti, Meredith Langi, and Sarah Quesen. 2025. Are we on the same page? a discussion on the use and misuse of early literacy assessments.

Nathan H Clemens, Kejin Lee, Maria Henri, Leslie E Simmons, Oi-Man Kwok, and Stephanie Al Otaiba. 2020. Growth on sublexical fluency progress monitoring measures in early kindergarten and relations to word reading acquisition. *J. Sch. Psychol.*, 79:43–62.

Kathryn Crowe and Sharynne McLeod. 2020. Children's english consonant acquisition in the united states: A review. *American Journal of Speech-Language Pathology*, 29(4):2155–2169.

Linnea C Ehri. 2005. Learning to read words: Theory, findings, and issues. *Sci. Stud. Read.*, 9(2):167–188.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *Preprint*, arXiv:2311.00430.

J Ricardo García and Kate Cain. 2014. Decoding and reading comprehension. *Rev. Educ. Res.*, 84(1):74–111.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of whisper models to child speech recognition. *Preprint*, arXiv:2307.13008.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2023. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23.

C. Oberle and S. Powers. 2025. Evaluating early literacy tasks: Insights from a mixed-methods study.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

A Appendix

Original Phoneme	**Sound in English**	**Variations**
/p/	p	p, b
/b/	b	b, p
/m/	m	m, b
/t/	t	t, d
/d/	d	d, t
/k/	k	k, g, t
/g/	g	g, k, d
/f/	f	f, v
/n/	n	n, d, nd, ŋ
/ŋ/	ng	ŋ, n
/w/	w	w
/j/	y	j, ’ ’
/h/	h	h, ’ ’
/v/	v	v, f
/s/	s	s, z, θ
/z/	z	z, s, ð
/ʃ/	sh	ʃ, s, θ
/ʒ/	si (as in ’vision’)	ʒ, ʃ
/tʃ/	ch	/tʃ/, ʃ, k, t, dʒ
/l/	l	l, w, j
/dʒ/	j	/dʒ/, tʃ, d
/θ/	th (voiceless)	θ, ð, t, f
/ð/	th (voiced)	ð, θ, d, v
/ɹ/	r	r, w, l, ’ ’
/tr/	tr	tr, tʃ, t
/dr/	dr	dr, dʒ, d
/kr/	kr	kr, gr, r, k, g
/gr/	gr	gr, kr, r, g, k
/skr/	skr	skr, sk, kr, s, k, r

Table 5: Example of general consonant variations

Addressing Few-Shot LLM Classification Instability Through Explanation-Augmented Distillation

William Muntean[1] **and Joe Betts**[1]
[1]National Council of State Boards of Nursing (NCSBN)
Chicago, IL
Correspondence: wmuntean@ncsbn.org

Abstract

Large language models (LLMs) are increasingly adopted for educational assessment despite evidence that specialized models achieve superior performance. This study compares few-shot in-context learning with explanation-augmented knowledge distillation for exam question classification using medical education data. Few-shot learning exhibited substantial performance instability, with accuracy varying up to 14 percentage points based on example selection, while knowledge distillation provided consistent 70.1% accuracy after proper hyperparameter optimization. Though neither LLM approach matched specialized BERT performance (80.5%), knowledge distillation eliminated the reliability issues plaguing few-shot methods, offering organizations a stable solution for leveraging existing LLM infrastructure in operational assessment applications.

1 Introduction

Large language models (LLMs) have gained widespread adoption across educational assessment applications, driven by their versatility and the appeal of unified infrastructure that can handle multiple tasks without maintaining separate specialized models. However, this adoption occurs despite evidence that task-specific approaches often achieve superior performance. Bucher and Martini (2024) demonstrated that fine-tuned smaller models, including BERT-based classifiers, significantly outperform both zero-shot and few-shot LLM approaches in text classification tasks. This performance gap raises important questions about how organizations already invested in LLM infrastructure can most effectively leverage these capabilities for reliable educational assessment applications, even when accepting that peak performance may require specialized alternatives.

Few-shot in-context learning represents the most straightforward approach to LLM-based classification, requiring no model training while promising reasonable performance through carefully selected examples. However, recent research has revealed substantial instability in few-shot classification performance, with accuracy varying significantly based on example selection, ordering, and prompt construction choices (Nguyen and Wong 2023; Alves et al. 2023; Wan et al. 2023). This variability extends beyond minor fluctuations, with identical examples presented in different orders producing measurably different classification outcomes. For operational assessment systems requiring consistent and reliable performance, such instability undermines the practical utility of few-shot approaches, even when average performance might be acceptable. The sensitivity to configuration choices introduces an additional layer of complexity that conflicts with the apparent simplicity that makes few-shot learning initially attractive.

Knowledge distillation offers a promising solution for organizations committed to LLM-based approaches, enabling the transfer of reasoning capabilities from large models to smaller, more efficient counterparts while maintaining performance consistency. Unlike few-shot learning, knowledge distillation produces stable models that do not depend on carefully curated examples at inference time. Explanation-augmented distillation extends this approach by incorporating the reasoning patterns and decision processes of teacher models, potentially capturing more nuanced classification strategies than traditional output-only distillation methods (Xu et al., 2024). While this approach may not achieve the peak performance of specialized classifiers, it represents an optimization strategy for organizations seeking to maximize the reliability and efficiency of LLM-based classification within existing infrastructure constraints. This study evaluates whether explanation-augmented knowledge distillation can provide the consistency and computational efficiency needed for operational deploy-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 197–203
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

ment while delivering competitive performance relative to unstable few-shot alternatives.

1.1 Knowledge Distillation in Natural Language Processing

Knowledge distillation has emerged as a powerful technique for transferring capabilities from large, computationally expensive models to smaller, more efficient alternatives while maintaining competitive performance. Originally developed for computer vision applications (Hinton et al., 2015), the approach has been successfully adapted to natural language processing tasks, where the computational demands of large language models create significant deployment challenges. Traditional knowledge distillation focuses on matching output distributions between teacher and student models, enabling smaller models to approximate the decision boundaries learned by their larger counterparts (Gou et al., 2021).

Recent advances in explanation-augmented knowledge distillation extend beyond output matching to incorporate the reasoning processes of teacher models. This approach leverages the natural language generation capabilities of large language models to produce detailed explanations alongside predictions, creating richer training signals for student models (DeepSeek-AI et al., 2025). By learning to replicate both the decisions and reasoning patterns of teacher models, student models may achieve better generalization and more robust performance across diverse inputs. However, the effectiveness of explanation-augmented distillation for classification tasks in educational domains remains underexplored.

1.2 Few-Shot Learning Instability

While few-shot in-context learning offers apparent simplicity for LLM deployment, mounting evidence reveals significant performance instability across different configuration choices. Studies have documented substantial variance in classification accuracy based on example selection, with different sets of representative examples producing measurably different results even when controlling for example quality and domain coverage (Nguyen and Wong, 2023). This instability extends to example ordering effects, where identical examples presented in different sequences can alter model predictions.

The sensitivity of few-shot learning to prompt construction choices poses particular challenges for operational deployment in educational assessment. Beyond random variation, systematic biases may emerge when examples exhibit consistent characteristics that do not represent the full complexity of the classification task (Tjuatja et al., 2024). These findings suggest that the apparent simplicity of few-shot learning may be misleading, as achieving reliable performance requires careful curation and validation of example sets—a process that may be as complex as traditional model training approaches.

1.3 Exam Question Classification

Educational assessment systems rely heavily on accurate classification of exam questions into predefined content domains to ensure proper test construction, maintain content validity, and support diagnostic feedback (Kane, 2006). This classification task involves mapping individual questions to taxonomic categories that reflect the knowledge, skills, or competencies being assessed. In medical education, for example, questions must be aligned with clinical domains, procedural categories, or competency frameworks to ensure comprehensive coverage of required learning outcomes (Bridge et al., 2003).

Traditional approaches to question classification have relied on manual expert review or rule-based systems, but the scale of modern item banks and the complexity of question content have motivated automated classification methods. Recent advances in data-driven approaches have also extended to optimizing assessment items themselves, including systematic methods for refining item options (Muntean et al., 2025). While specialized models like fine-tuned BERT classifiers have demonstrated superior performance for this task (Bucher and Martini, 2024), many educational organizations seek to benefit from existing LLM infrastructure for question classification as part of broader assessment workflows.

The stakes for classification accuracy in educational assessment are particularly high, as misclassified questions can compromise test validity, lead to content imbalances, and undermine the reliability of score interpretations (Messick, 1995). This context demands not only reasonable classification performance but also consistent and predictable behavior across diverse question types and content areas.

1.4 Research Questions

This study investigates the effectiveness of explanation-augmented knowledge distillation for exam question classification compared to few-shot in-context learning approaches. Specifically, we address the following research questions:

RQ1 How does few-shot in-context learning performance vary when examples are systematically selected based on question difficulty (easy vs. difficult vs. mixed examples within each content domain)?

RQ2 Can explanation-augmented knowledge distillation produce student models that achieve competitive classification accuracy compared to few-shot learning approaches while maintaining greater performance consistency?

RQ3 How sensitive is explanation-augmented knowledge distillation to hyperparameter choices, and what configurations optimize the trade-off between performance and training efficiency?

2 Methods

2.1 Dataset

We utilized a subset of a medical examination item bank containing 6,839 multiple-choice questions labeled according to eight high-level test blueprint domains (National Council of State Boards of Nursing, 2023). Items were randomly selected from questions that had passed all statistical pretest criteria. The dataset was partitioned using stratified sampling to maintain domain proportions: 4,103 questions (60%) for training, 1,368 questions (20%) for validation, and 1,368 questions (20%) for testing. Question difficulty was determined using population-calibrated item difficulty values, with difficulty distributions roughly equivalent across the eight content domains. The classification task involved mapping individual question to their corresponding content domains based on the medical knowledge and competencies being assessed.

To establish performance benchmarks, we implemented a BERT-based classification model using the all-MiniLM-L6-v2 sentence transformer (Wang et al., 2020). We fine-tuned the model using contrastive learning with 8,000 question pairs (1,000 pairs per domain, consisting of 500 positive and 500 negative pairs), representing the specialized classification method that has been shown to outperform LLM-based approaches in similar text classification tasks.

2.2 Few-Shot Learning Experiments

We conducted few-shot learning experiments primarily using GPT-OSS-20B (OpenAI et al., 2025), with pilot studies on GPT-OSS-120B, LLaMA 4 Maverick (Meta AI, 2025), and Claude Sonnet 3.7 (Anthropic, 2025) to validate that instability patterns generalize across different large language models. All models were accessed through Databricks environment endpoints. To systematically investigate the impact of example difficulty on few-shot performance, we created three experimental conditions based on population-calibrated difficulty values. For easy examples, we selected the 25 easiest items and randomly divided them into 5 sets of 5 items per domain. For difficult examples, we selected the 25 most difficult items and applied the same division strategy. For mixed examples, we randomly selected 5 items per domain, repeated 5 times. Each condition resulted in 5 replications of 40 few-shot examples (5 examples from each of the 8 domains), enabling assessment of both systematic bias effects and random variation. All few-shot examples were drawn from the training set to prevent data leakage.

The prompt structure consisted of task instructions, content domain definitions, few-shot examples with their classifications, and repeated instructions with output format specifications. All models were required to follow structured output formatting to ensure consistent response parsing. Each replication was evaluated on the complete test set to quantify performance variability across different example selections.

2.3 Knowledge Distillation Experiments

We implemented explanation-augmented knowledge distillation using LLaMA 3.1 405B as the teacher model and LLaMA 3.1 8B as the student model. For all 4,103 training examples, we prompted the teacher model to generate detailed rationales explaining why each question belonged to its specified domain and why alternative domains were less appropriate. This process created question-explanation-classification triplets that enabled the student model to learn both the reasoning patterns and classification decisions of the teacher model.

The student model underwent full parameter fine-tuning on these explanation-augmented sequences using Databricks parameter sweep functionality. We conducted systematic hyperparameter optimization across a 2×2 experimental design with learning rates of 1×10^{-6} versus 1×10^{-7} and training epochs of 1 versus 2. This design enabled assessment of hyperparameter sensitivity while maintaining computational feasibility for the full parameter fine-tuning approach.

2.4 Evaluation

We evaluated all approaches using overall accuracy and weighted F_1-score on the held-out test set (1,368 questions) to ensure unbiased performance assessment. For few-shot learning approaches, we measured performance consistency by calculating the standard deviation and range (maximum - minimum accuracy) across the 5 replications within each difficulty condition. This analysis quantifies both the magnitude and variability of performance instability across different example selections, enabling direct comparison with the consistent performance of knowledge distillation approaches.

3 Results

The experimental results demonstrate clear performance differences between approaches and reveal significant instability in few-shot learning methods. The BERT baseline achieved the highest overall performance with 80.5% accuracy and 80.4% weighted F_1-score, confirming prior findings that specialized models outperform LLM-based approaches for text classification tasks. However, the comparison between few-shot learning and knowledge distillation reveals important insights about the viability of LLM-based classification methods.

3.1 Few-Shot Learning Performance and Instability

Few-shot learning performance varied substantially based on example difficulty, with counterintuitive results regarding the relationship between example difficulty and classification accuracy. Models performed best when provided with difficult examples (62.7% accuracy, 60.7% F_1), followed by random examples (56.0% accuracy, 54.0% F_1), and worst with easy examples (52.8% accuracy, 51.4% F_1). This unexpected finding suggests that difficult questions may provide richer contextual information or more distinctive features that help models distinguish between content domains.

More critically, few-shot learning exhibited substantial performance instability across different example selections within each difficulty condition. The difficult examples condition showed the highest variability, with accuracy ranging from 56.8% to 71.0% (standard deviation = 6.1%) and F_1 scores ranging from 53.9% to 68.9% (standard deviation = 6.7%). Easy examples demonstrated moderate instability with accuracy ranging from 47.3% to 58.2% (standard deviation = 4.4%), while random examples showed the most consistent performance with accuracy ranging from 51.7% to 60.7% (standard deviation = 3.9%). Despite this relative consistency, even the random condition exhibited meaningful performance variation that could impact operational deployment reliability.

The instability patterns were consistent across multiple large language models tested in pilot studies, including GPT-OSS-120B, LLaMA 4 Maverick, and Claude Sonnet 3.7, indicating that few-shot learning instability represents a general phenomenon rather than model-specific behavior. This cross-model consistency strengthens the evidence that example selection significantly impacts few-shot classification performance regardless of the underlying architecture.

3.2 Knowledge Distillation Performance and Stability

Explanation-augmented knowledge distillation results revealed extreme sensitivity to hyperparameter selection, with learning rate choice proving critical for successful model training. The optimal configuration using learning rate 1×10^{-6} and 2 training epochs achieved 70.1% accuracy and 70.4% weighted F_1-score, representing competitive performance relative to few-shot learning approaches while completely eliminating the instability associated with example selection.

Hyperparameter analysis revealed dramatic performance differences based on learning rate selection. Models trained with learning rate 1×10^{-6} substantially outperformed those trained with 1×10^{-7}, likely due to catastrophic forgetting effects at the extremely low learning rate that prevented adequate adaptation to the classification task. The 1×10^{-7} learning rate produced poor performance regardless of epoch count (44.0% accuracy with 1 epoch, 47.0% accuracy with 2 epochs), while the 1×10^{-6} learning rate enabled effective learning (60.8% accuracy with 1 epoch, 70.1% accuracy with 2 epochs). The improvement from 1 to 2 epochs at the higher learning rate suggests that additional training time benefits explanation-augmented distillation when appropriate learning rates are used.

Knowledge distillation demonstrated complete stability across evaluation runs, producing identical performance metrics without the variability that characterized few-shot approaches. This stability emerges from the fundamental difference in methodology: rather than relying on a small set of potentially biased examples at inference time, the distilled model learns from the complete training dataset during fine-tuning. The student model internalizes classification patterns through exposure to the full range of question types and difficulty levels, eliminating dependence on the specific characteristics of a limited example set. Once trained, the distilled model produces consistent predictions without requiring carefully curated examples, removing the primary source of instability that affects few-shot learning.

3.3 Comparative Analysis

The knowledge distillation approach addresses the primary limitation of few-shot learning by providing consistent performance without dependence on carefully curated examples. While the best knowledge distillation configuration (70.1% accuracy) did not exceed the maximum few-shot performance (71.0% accuracy with difficult examples), it achieved performance within the range of few-shot results while eliminating the risk of poor performance due to unfavorable example selection. Notably, the knowledge distillation performance exceeded the minimum few-shot performance across all difficulty conditions and matched the performance of the best few-shot condition (difficult examples) while avoiding the substantial variability that makes few-shot approaches unreliable.

The comparison reveals a fundamental trade-off between peak performance potential and performance consistency. Few-shot learning offers the possibility of higher performance when examples are carefully selected, but carries substantial risk of poor performance with different example choices. Knowledge distillation provides predictable performance that falls within the middle-to-upper range of few-shot results, representing a viable solution for organizations requiring reliable classification performance from LLM-based approaches. The elimination of example-dependent variability makes knowledge distillation particularly suitable for operational deployment where consistent performance is more valuable than occasional peak performance.

4 Discussion

This study provides empirical evidence that explanation-augmented knowledge distillation offers a viable solution to the instability problems that plague few-shot in-context learning for educational question classification. While neither LLM-based approach achieves the performance of specialized BERT classifiers, the findings reveal important practical considerations for organizations committed to leveraging existing LLM infrastructure for assessment applications.

The substantial performance variability observed in few-shot learning—with accuracy ranges exceeding 14 percentage points in some conditions—represents a significant barrier to operational deployment. This instability extends beyond random variation to include systematic biases based on example characteristics, as demonstrated by the counterintuitive finding that difficult examples produced better classification performance than easy examples. This result suggests that few-shot learning may be sensitive to the cognitive complexity and feature richness of selected examples in ways that are difficult to predict or control. The consistency of these instability patterns across multiple large language models indicates that the problem is fundamental to the few-shot learning paradigm rather than specific to particular architectures. For educational assessment applications, this variability is particularly concerning as unreliable classification performance can compromise test validity and undermine confidence in automated systems. The observed variability means that identical classification tasks could produce different results depending solely on example selection choices, creating potential fairness and consistency issues in high-stakes assessment environments.

Explanation-augmented knowledge distillation addresses these limitations by fundamentally changing the relationship between examples and model performance. Rather than depending on a small set of potentially biased examples at inference time, the distilled model learns from comprehensive exposure to the full training dataset, internalizing classification patterns that remain consistent across evaluations. This methodological difference eliminates the primary source of instability in few-shot approaches while maintaining competitive performance levels. The extreme sensitivity to hyperparameter selection observed in our distillation experiments, particularly the dramatic performance differences between learning rates, highlights the importance of systematic optimization rather than relying on conventional parameter choices. The poor performance at learning rate 1×10^{-7} likely reflects catastrophic forgetting, where the extremely conservative learning rate prevented adequate adaptation to the classification task. However, once properly configured, the distilled model produces stable and reliable performance without the variability that characterizes few-shot approaches.

The trade-off between peak performance and consistency revealed in our results reflects broader considerations in educational technology deployment. While few-shot learning may occasionally achieve higher performance with optimal example selection, the risk of poor performance with suboptimal examples may be unacceptable in assessment contexts where consistent behavior is essential. Knowledge distillation provides a middle path that sacrifices some performance potential for greater reliability and predictability, making it particularly suitable for operational assessment applications where consistency is paramount.

Several limitations should be considered when interpreting these results. Our evaluation focused on a single domain (medical education) and classification task, and generalization to other educational contexts requires further investigation. The hyperparameter space explored for knowledge distillation was limited, and more comprehensive optimization might yield improved performance. Future research should investigate the effectiveness of explanation-augmented distillation across diverse educational domains, examine different distillation methods, and analyze the quality and utility of generated explanations. Additionally, research into methods for automatically selecting optimal few-shot examples or reducing example dependency could address some of the limitations identified in few-shot approaches.

This study demonstrates that explanation-augmented knowledge distillation provides a practical solution to the instability problems inherent in few-shot learning approaches for educational question classification. The elimination of example-dependent variability, combined with competitive performance levels, makes knowledge distillation particularly suitable for operational assessment applications where consistency and reliability are paramount. These findings contribute to the growing understanding of how to effectively deploy large language models in educational contexts while managing their inherent limitations and operational constraints, offering organizations a viable path to leverage existing LLM infrastructure reliably and consistently.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet [large language model]. `https://www.anthropic.com/news/claude-3-7-sonnet`. Anthropic PBC.

Patrick D Bridge, Joseph Musial, Robert Frank, Thomas Roe, and Shlomo Sawilowsky. 2003. Measurement practices: methods for developing content-valid student examinations. *Medical teacher*, 25(4):414–421.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Michael Kane. 2006. Content-related validity evidence in test development. *Handbook of test development*, 1:131–153.

Samuel Messick. 1995. Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, 14(4):5–8.

Meta AI. 2025. Llama 4 maverick (17b-128e) [large language model]. `https://huggingface.co/meta-llama/`

`Llama-4-Maverick-17B-128E-Instruct`. Meta Platforms, Inc.

William Muntean, Joe Betts, Zhuoran Wang, and Hao Jia. 2025. Comparing data-driven methods for removing options in assessment items. *Journal of Educational Measurement*. Online First.

National Council of State Boards of Nursing. 2023. Next generation nclex: Nclex-rn test plan. `https://www.ncsbn.org/public-files/2023_RN_Test%20Plan_English_FINAL.pdf`. National Council of State Boards of Nursing, Chicago, IL. Accessed September 2025.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *Preprint*, arXiv:2302.11042.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O. Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. *Preprint*, arXiv:2305.14106.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Identifying Biases in Large Language Model Assessment of Linguistically Diverse Texts

Lionel Meng
Shamya Karumbaiah, Ph.D.
Vivek Saravanan
Daniel Bolt, Ph.D.
University of Wisconsin - Madison / 1025 W Johnson St, Madison, WI 53706
Correspondence: lhmeng@wisc.edu, shamya.karumbaiah@wisc.edu

Abstract

The development of Large Language Models (LLMs) to assess student text responses is rapidly progressing but evaluating whether LLMs equitably assess multilingual learner responses is an important precursor to adoption. Our study provides an example procedure for identifying and quantifying bias in LLM assessment of student essay responses.

1 Introduction

The application of Large Language Models (LLMs) for assessing student essays affords numerous avenues of research within learning analytics. Particularly for high stakes assessment contexts where annotated data is often sparse or difficult to acquire, the use of LLMs becomes particularly attractive. However, for LLMs to be ethically applied to educational assessment, they must be evaluated for equity across diverse student subpopulations. One subpopulation of particular concern is multilingual students. In high stakes testing contexts, acquiring sufficient annotated data for multilingual students is often unrealistic for reasons such as test security, student privacy, diversity in linguistic practices, and low population size. Furthermore, traditional methods of algorithmic bias assessment that rely on broad demographic categories such as age or gender are prone to mis-characterize the complex heterogeneous backgrounds of such students, potentially making them ineffective. Direct empirical comparisons across subpopulations can also be complicated by difficulties in separating bias from impact (Ackerman, 1992). Understanding causes of differential item functioning is a notorious challenge in the use of empirical data for evaluating bias (Zumbo, 2007).

2 Aim

Our study aims to illustrate a procedure by which LLM performance can be assessed for equity by systematically manipulating texts with construct-irrelevant linguistic variations and characterizing resultant score change. We refer to these variations as perturbations and the resulting scores as perturbed scores.

3 Sample(s)

Texts are from the Hewlett Foundation: Automated Essay Scoring competition data (Hammer et al., 2012). The sample consists of 5875 actual essay responses written by students in grades 7 through 10 in response to prompts that did not have accompanying reading passages (essay numbers 1, 2, 7, 8). Essays range from 150 to 550 words in length.[1]

4 Methods

Analyses of texts begins by feeding original, monolingual texts to the target LLM, GPT-4o, for scoring.

4.1 LLM Prompting

We employed GPT-4o as the LLM for essay grading. The model was prompted with "Grade the essay below with a score between 0 and 100 based on content, ignoring language errors. Your response must be exactly one number between 0 and 100". Scores were normalized to range from 0 to 1.

4.2 Text Quality and Baseline Reference Values

Each monolingual text was scored twice by the LLM. The first of each of these scores was chosen to be a reference value for subsequent analyses, and is henceforth referred to as the "original score." We refer to the second score as the "replicate score."

Texts were then grouped into quartiles based on original scores. Resultant quartile sizes, from first to fourth, were as follows: 1767, 2600, 619,

---

[1] This dataset is openly accessible at [https://www.kaggle.com/competitions/asap-aes/overview].

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 204–210
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

889. For each quartile, the first three statistical moments of the difference between original and perturbed scores were calculated. The inclusion of higher order moments reflects the notion that equity in measurement transcends expected score differences, and includes equivalence in precision as well as potential for outlier scores, etc.

Corresponding moments were also calculated for the difference between original and replicate scores to serve as a baseline. Specifically, an expected signed difference between the original score and replicate score (first order), an expected squared deviation between the original score and replicate score (second order), and an expected signed cubed deviation between the original and replicate score were calculated (third order).

4.3 Construct-Irrelevant Linguistic Variations

Construct-irrelevant linguistic variations here are defined as linguistic features of the text that are not directly related to the content proficiency intended to be measured. For a student essay response on a science test question, for example, it might refer to spelling errors that the student makes. The idea is that the underlying student response may be scientifically accurate despite the linguistic variation.

While it is true that large proportions of linguistic variations may impede LLM scoring of the text, not unlike barriers to comprehension that may occur with a human scorer, the label of "construct-irrelevant" is used to highlight that these are not the intended target construct of measurement. Indeed, where the LLM scoring becomes difficult due to perturbations, this difficulty itself becomes a form of inequity (Prabhakaran et al., 2019).

The four linguistic variations analyzed in this study were: 1) spelling errors, 2) noun transfers (i.e., borrowing nouns across languages; e.g., “Tierra” instead of “Earth”), 3) cognates (i.e., borrowing words with similar meaning, spelling, and pronunciation; e.g., “océano” instead of “ocean”), and 4) Spanglish (a hybrid use of both languages; e.g., “en la Earth” instead of “on the Earth”).

We build an algorithm in which eligible words or phrases at which the above linguistic variations could occur is first determined for each text. Then, the linguistic variations above are randomly introduced to each of the texts at these words or phrases, resulting in transformed, or perturbed versions of the text with the same underlying response meaning. Under this scheme, the magnitude of the perturbation can be controlled. For this study, we introduce perturbations of the following magnitudes: 20, 40, 60, and 80 percent.

4.4 Assessment of Inequity

The perturbed texts are scored by GPT-4o, such that each text has not only an original score, but a perturbed score as well. Using these values, we apply procedures conceptually derived from Lord’s (1980) notions of equity.

4.4.1 First Order Inequity

We refer to the signed difference between the original score and perturbed score as "error" for each text. By calculating the expected error (original - perturbed) across texts for each quartile, we can determine a quartile-specific bias value attributable to the linguistic perturbations. Expected perturbation error values greater than the expected replicate error values suggest first order inequity.

4.4.2 Second Order Inequity

By calculating the expectation of squared deviations between the original and perturbed score across essays within each quartile, we can get quartile-specific variances of deviations. We take the square root of these values to get standard deviations, and compare to the corresponding standard deviation for replicate scores as a reference. Standard deviation values that surpass the reference values suggest second-order inequity.

4.4.3 Third Order Inequity

We also calculate the expectation of signed cubed deviations between the original and perturbed score across essays for each quartile. Values greater than quartile-specific third order baseline reference values defined using replicate scores suggest third order inequity.

5 Results

A Wilcoxon test (see Table 2) was conducted to compare original scores and perturbed scores, confirming that differences in scores seen due to linguistic perturbations of varying magnitudes were statistically significant. Our sample size of 5875 texts naturally predisposes the test to be significant, even with small average deviations. While practical significance of such LLM audits are context-specific and best determined on a case-by-case basis, for this study readers are referred to Figure 1, Figure 2, and Figure 3 where the effect of linguistic perturbations are quantified on the scale of the text

perturbation & magnitude	**Wilcoxon test statistic**	**p value**
spanglish 20	3533900	$p < 0.05$
spanglish 40	3581410	$p < 0.05$
spanglish 60	3481962	$p < 0.05$
spanglish 80	3727746	$p < 0.05$
cognates 20	4146682	$p < 0.05$
cognates 40	4147836	$p < 0.05$
cognates 60	3969052	$p < 0.05$
cognates 80	3847161	$p < 0.05$
noun transfer 20	3944342	$p < 0.05$
noun transfer 40	3794908	$p < 0.05$
noun transfer 60	3524270	$p < 0.05$
noun transfer 80	3498297	$p < 0.05$
spelling 20	2669144	$p < 0.05$
spelling 40	1732722	$p < 0.05$
spelling 60	1436490	$p < 0.05$
spelling 80	1132988	$p < 0.05$

Table 1: Wilcoxon test results comparing original and perturbed scores by linguistic variation and magnitude.

scores (0 to 1), and as such may serve as effect size measures.

Figure 1 displays quartile-specific results for bias in each linguistic perturbation, faceted by magnitude of perturbation. A general trend whereby increasing magnitudes of perturbation result in greater mean error can be observed. Additionally, for all linguistic variations, mean error values trend positive as successive quartile results are compared for all magnitudes of perturbation. Mean error values for all perturbations exceed the first order baseline reference level in the positive direction, although error values are still negative in the first quartile for all perturbations aside from spelling errors.

Figure 2 displays quartile-specific results for the expected value of squared deviations between original scores and perturbed scores, converted to standard deviations for each linguistic perturbation, faceted by magnitude of perturbation. Quartile-specific baseline reference values are represented as horizontal lines. With the exception of noun transfer at 60% magnitude for the first quartile, all standard deviation values were greater than baseline reference values within their respective quartiles. A moderate trend can be observed such that for texts in the first quartile, standard deviations tend to be high across linguistic variations.

Figure 3 displays quartile-specific results for the expected value of cubed deviations between original scores and perturbed scores for each linguistic perturbation, faceted by magnitude of perturbation. Skewness values exceed baseline reference values in the positive direction for all linguistic perturbations, although skewness values are still negative in the first quartile for all perturbations aside from spelling errors. Additionally, cognate skewness values only barely surpass baseline reference values in the second quartile.

6 Discussion

For texts in the second to fourth quartile of text quality, sensitivity of GPT-4o scores to linguistic perturbations of varying magnitudes in all analyses suggests inequitable assessment of student knowledge in its application. First order results indicate the presence of bias in LLM scoring, second order results further indicate differences in precision, and third order results indicate the an increased likelihood for extreme cases of discrepant results in the positive direction. Additional interpretation for results of texts in the first quartile are presented below.

First order analysis results for the first quartile in Figure 1 show mean error values surpassing the baseline reference value in the positive direction while remaining negative. In interpreting these results, however, it should be appreciated that the negative value associated with the reference condition likely represents a "regression to the mean" phe-

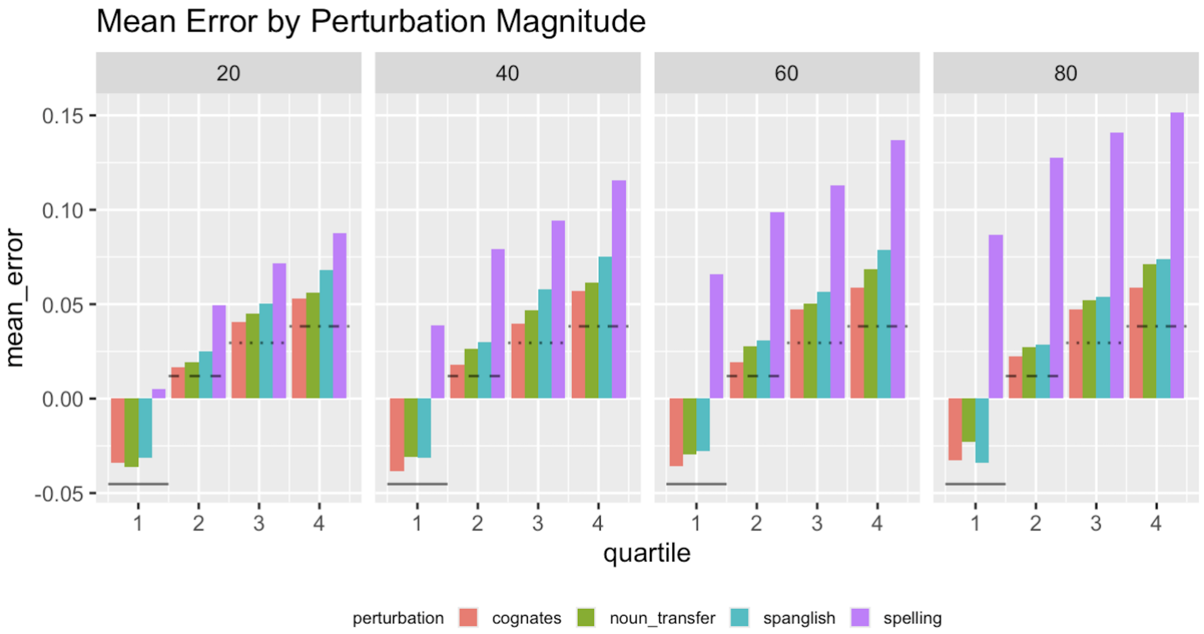


Figure 1: Error (original score - perturbed score) averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for signed difference between original and replicate score are included for each quartile.

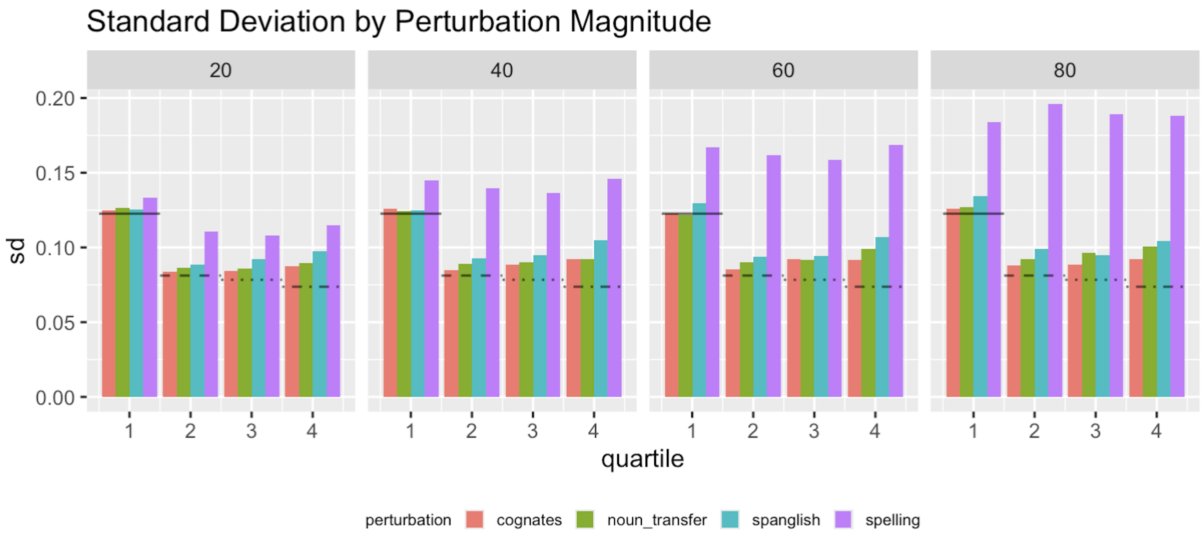


Figure 2: Standard deviation of differences averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for square root of expected squared deviations between original and replicate scores are included for each quartile.

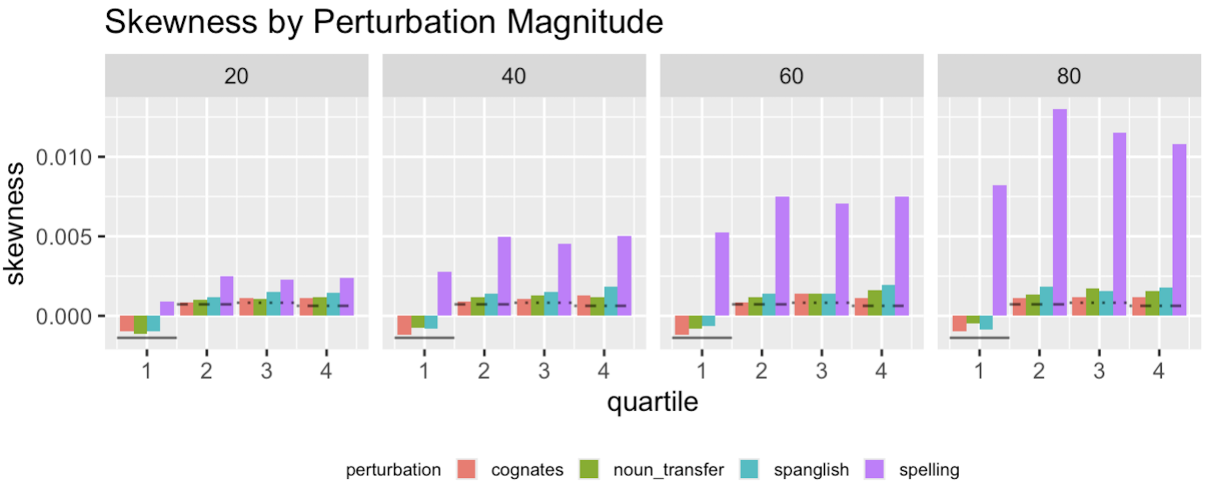


Figure 3: Skewness averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for expected cubed deviations between original and replicate scores are included for each quartile.

Cognate Perturbation 40% Magnitude Sample Texts with Large Error

Text ID	Original Score	Perturbed Score	Error
19047	0.15	0.60	-0.45
240	0.25	0.65	-0.40
18088	0.10	0.50	-0.40
21140	0.10	0.50	-0.40
1245	0.20	0.60	-0.40
19190	0.45	0.15	0.30
19479	0.50	0.20	0.30
276	0.45	0.10	0.35
19320	0.45	0.10	0.35
18578	0.45	0.00	0.45

Table 2: Sample texts with large magnitudes of error for cognate perturbations at 40% magnitude in the 1st quartile. Scores are on a scale from 0 to 1. Error is calculated as original score - perturbed score. Results show potential for error scores in either direction.

nomenon, an expected statistical result, as scoring error is on average negative in the lowest quartile. Thus the larger values observed under perturbation, although still often negative, can nevertheless be viewed as a first order equity violation (albeit generally small), in that less than the expected regression correction is observed under perturbation. Additionally, it is important to keep in mind that mean error scores are muted due to cancellation from the signed nature of the quantity; paired with large variance of scores (see Figure 2), this leaves a nontrivial likelihood of inequities in scoring for particular students. In other words, for texts produced by students of developing proficiency, the target LLM would be expected to grade multilingual student text responses with lower levels of precision than monolingual student text responses. As an example, Table 2 shows the 5 most negative error value texts and 5 most positive error value texts for cognate perturbations at 40% magnitude in the first quartile. Comparably large magnitudes of error in either direction illustrate how noisy assessment of perturbed texts as shown by second order analysis results from Figure 2 can manifest. Given that academic decisions for students occur in consideration of individual scores, not group-aggregated values such as mean scores, this is potential reason for concern. Third order analysis results (see Figure 3) for the first quartile are similar to first order results in that expected cubed deviations are on average negative and regression to the mean corrections are more weakly observed under perturbation, indicating mild third order equity violations. We conclude that the effect of linguistic perturbation results in violations of equity in all three orders, with particularly strong results for spelling errors.

7 Limitation and Future Directions

Our study intended to highlight a methodology for examining the effects of perturbations on LLM scoring. A primary limitation of our results relates to the constantly changing nature of LLMs. It is likely that the validity of audit results for any given LLM will have limited longevity. As such, stakeholders are advised to audit their target LLMs as close to the time of application as possible.

As the authenticity of the algorithmically-introduced linguistic variations can be questioned, results from this audit procedure should be interpreted cautiously. More developed ways to introduce these perturbations can be implemented in the future to improve the validity of the audit procedure.

One challenging aspect of this procedure lies in determining comparable magnitudes of different perturbations. In this study, we opted to use the number of eligible words in the text for a given linguistic variation. For spelling error, this includes all words in the text. However, for noun transfer, only nouns in the text would be included. Thus, a 20% magnitude spelling error perturbation involves more words being perturbed than in a 20% magnitude noun transfer perturbation. This is likely why in our results, spelling error perturbations show the largest bars across all quartiles for all analyses. Depending on the text feature and context, different methods for normalization may be preferable.

While this study focused on a limited selection of linguistic variations, the audit procedure can be applied for analysis of various other text features. This gives stakeholders flexibility to choose those features that are most appropriate to their context and use case. Analysis of additional features of essays (ie. length) could also allow for investigation of potential moderating effects on linguistic perturbations.

When establishing baseline reliability, only two trials of test-retest analysis were conducted. While this decision was made for illustration of concept, for more robust audits of LLMs, more replications should be included. On a related note, perturbed texts were only passed to the LLM for scoring once each in this study - more replications could be considered for improved auditing of LLM scoring. Increasing replications of both original and perturbed scores would additionally afford the opportunity to analyze the effect of text features on scoring at the individual essay level.

Another natural future direction of this study is to incorporate higher moments for analyses, which have the potential to illustrate further nuances of potential inequities in LLM scoring. There is no theoretical limit to moments that can be analyzed.

8 Conclusion

Our study has provided an example procedure for evaluating LLM scoring of texts for equity, incorporating algorithmically introduced linguistic perturbations and higher order moment analyses in characterizing impacts on stakeholders. We believe this procedure to be useful in the following ways:

First, in such contexts as educational testing with multilingual student populations where annotated data is sparse, such a procedure has the potential to augment our ability to evaluate whether LLMs are ethically appropriate for application.

Second, due to the experimental nature of this process whereby the effects of the perturbation can be isolated, sources of LLM bias can be directly studied. Furthermore, LLM scoring does not suffer from carryover effects the way human raters might, allowing true replications of scores to be obtained for study. By further investigating how distributions of target features may vary across groups (e.g. multilingual vs. monolingual students), stakeholders can leverage audit results to infer how LLM scores may manifest as bias at the subpopulation level.

Third, this procedure is accessible in that it can be conducted by stakeholders in various contexts (not just education) for evaluation of their target LLM, and with respect to various features beyond those targeted in this study.

References

Terry A. Ackerman. 1992. A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*, 29(1):67–91.

Ben Hammer, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The Hewlett Foundation: Automated Essay Scoring.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum, Hillsdale, NJ.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*.

Bruno D Zumbo. 2007. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, 4(2):223–233.

Implicit Biases in Large Vision-Language Models in Classroom Contexts

Peter Baldwin
Office of Research Strategy, National Board of Medical Examiners, Philadelphia, USA
pbaldwin@nbme.org

Abstract

We tested whether GPT-4o exhibits bias when rating classroom excerpts embedded in images of teachers differing by race and gender. Four teacher images (Black female, White female, Black male, White male) were paired with short lecture excerpts across four instructional contexts. The model was instructed only to rate the embedded excerpt—without reference to the image itself—on four dimensions: Clarity, Student Engagement, Coherence and Organization, and Pedagogical Effectiveness. Ratings were compared using paired nonparametric tests with multiplicity adjustment. Across 32 factor-specific tests, 18 were significant. Effects favored female teachers at the 3rd-grade level, male teachers in graduate-level advanced mathematics, and generally favored White teachers; effect sizes were small. These findings are relevant for researchers studying bias in vision–language models and for practitioners involved in teacher evaluation or instructional content review.

1 Introduction

> *"Facts are not pure and unsullied bits of information; culture also influences what we see and how we see it."*
> –Gould, *The Mismeasurement of Man* (1981)

Visual bias in humans refers to systematic distortions in the perception or interpretation of visual stimuli based on prior beliefs. As machines increasingly process visual inputs, these systems have also been shown to exhibit visual bias (e.g., Ananthram et al., 2024; Fraser & Kiritchenko, 2024; Hamidieh et al., 2024; Howard et al., 2024; Lee & Jeon, 2024; Greene et al., 2025; Kim et al., 2025). One class of widely-used large multimodal models that couple an image (or video) encoder to a large language model is referred to as *large vision-language models* (LVLMs). Visual bias has been observed in LVLMs and at times, these distortions resemble implicit bias, which in humans operates unconsciously. However, because LVLMs lack consciousness, researchers may use terms such as latent, emergent, or data-driven bias to avoid conflating these effects with human cognition. Regardless of terminology, all refer to the same underlying phenomenon: outputs that are misaligned with intent and not visible in the system architecture. As such, biases of this kind must be identified through empirical testing.

In this study, counterfactual images of teachers differing only by race and gender were created. Classroom lecture excerpts were generated; for each excerpt, the identical text was superimposed onto each image, and a large vision-language model (LVLM) was asked to evaluate the excerpts on four criteria. Because the model was tasked solely with evaluating identical written content, this design isolates whether—and to what extent—a given teacher's visually cued race and gender alter the model's evaluation. The evaluation task was repeated across four teaching contexts, resulting in 32 (4 criteria × 4 contexts × 2 factors) comparisons, of which 18 showed evidence of emergent visual bias. Aside from identifying an especially salient illustration of such bias, this audit-style investigation broadens the scope of bias assessment to include model behaviors that may be particularly relevant in applied contexts such

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 211–217
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

as instructional content review or teacher evaluation.

2 Background

Many evaluations of LLM bias rely on static benchmarks or fixed stereotype probes—such as WEAT (Caliskan et al., 2017), StereoSet (Nadeem et al., 2020), or WinoBias (Zhao et al., 2018)—but these approaches can fail to capture context-sensitive or decision-level forms of bias. Recent work emphasizes evaluation strategies that account for context, intersectionality, and model-specific behavior (Gohar & Cheng, 2023; Bateni et al., 2022). Scenario-based audits that simulate real-world tasks provide one way to accomplish this and, in the context of LVLMs, typically focus on how a system perceives, generates, or describes visual content under controlled conditions (Schwartz et al., 2024; Mökander et al., 2024; Gaebler et al., 2024; An et al., 2025; e.g., Hamidieh, 2024; Fraser & Kiritchenko, 2024; Howard et al., 2024; Greene et al., 2025; Lee & Jeon, 2024; Ananthram et al., 2024; Kim et al., 2025). Building on this foundation, the present study extends the counterfactual audit framework into the domain of evaluative judgment.

2.1 Problem Definition

We study whether an LVLM's evaluation of identical instructional text depends on visually cued teacher race and gender in the background image. For each excerpt, the same text is superimposed onto each teacher image while non-demographic visual features (framing, pose, expression, approximate age, clothing, background) are held constant. The model outputs ratings on four criteria (Clarity; Student Engagement; Coherence and Organization; Pedagogical Effectiveness) across four teaching contexts. Race and gender each had two levels: Black and White; Female and Male. We define bias as systematic differences in ratings attributable to race or gender under these invariants. The primary question is answered by planned, paired comparisons for race and for gender within context (family-wise control specified in Section 4.5). Other constructs (e.g., perception accuracy) are out of scope.

2.2 Related Work

Controlled comparisons that isolate demographic cues have long been used in social science research, such as resume and housing studies (Bertrand & Mullainathan, 2004). Audit methods of this kind have also been proposed for assessing emergent bias in AI systems (Gohar & Cheng, 2023; Bateni et al., 2022) and have motivated scenario-based audits for these systems (Schwartz et al., 2024; Mökander et al., 2024; Gaebler et al., 2024; An et al., 2025). In the LVLM literature, audit-style evaluations have examined perception and labeling (Ananthram et al., 2024; Kim et al., 2025; Greene et al., 2025), generation (Hamidieh, 2024), and description/captioning (Nadeem et al., 2021; Zhou et al., 2022; Fraser & Kiritchenko, 2024; Howard et al., 2024; Lee & Jeon, 2024; Greene et al., 2025). Our design falls within this family (e.g., Fraser & Kiritchenko, 2024; Howard et al., 2024; Lee & Jeon, 2024) and most closely parallels Kim et al. (2025), who showed that demographic attributes in images can influence identification even when demographic information is not requested. Here, the adversarial element is further dissociated from the task: rather than perception or description, we superimpose identical lecture excerpts onto counterfactual teacher images and ask the LVLM to rate only the *written* content, testing whether visual attributes that are formally irrelevant to the evaluation nonetheless shape model output.

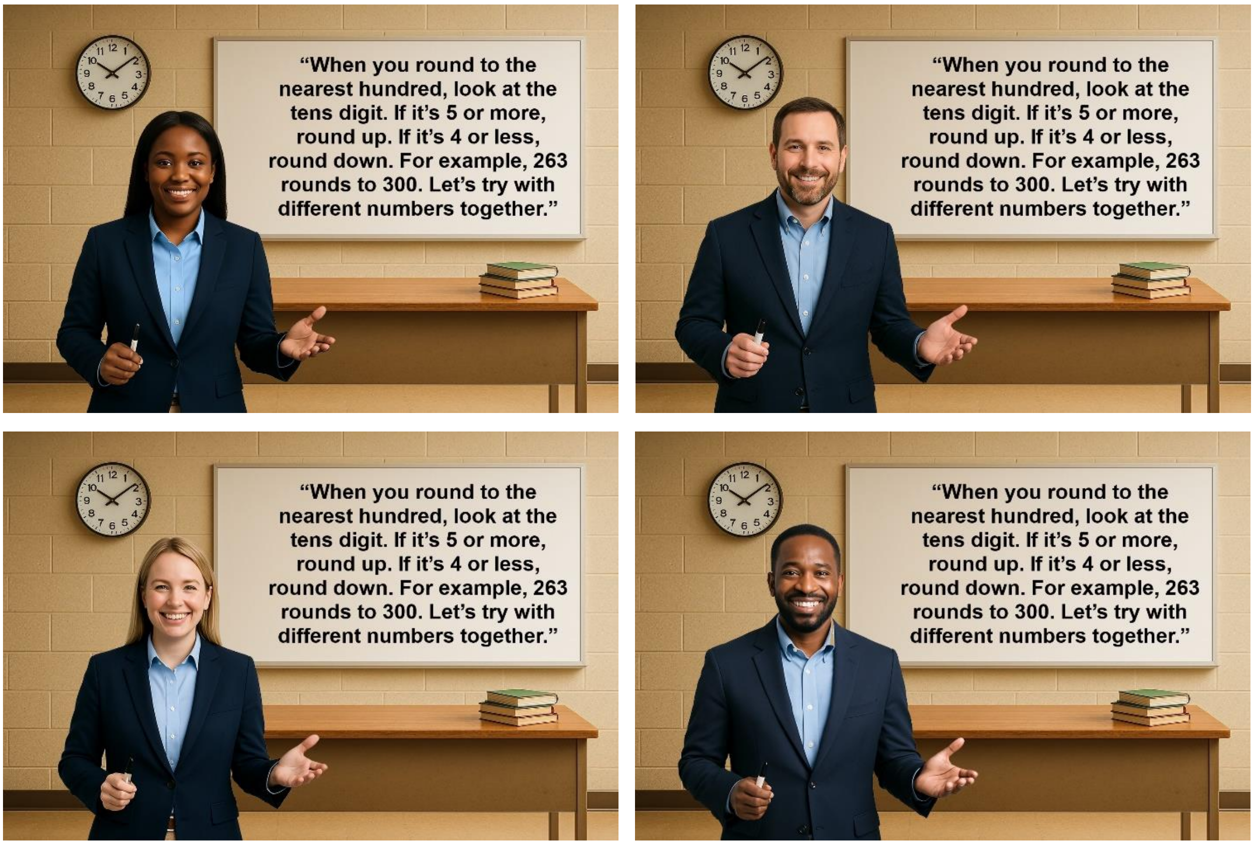
3 Methodology

3.1 Proposed Procedure

The procedure includes the following five steps.

1. *Image construction*. Generate teacher portraits that vary only by demographic characteristic and iteratively refine images to maximize similarity on non-demographic attributes; composite each portrait onto a common classroom background so that framing, pose, facial expression, attire, and apparent scale are held constant.
2. *Excerpt generation*. Generate short (35–50 word), age-appropriate classroom lecture excerpts for specific instructional contexts.
3. *Counterfactual pairing*. Superimpose each excerpt onto each teacher image to create matched sets differing only by the teacher's visual identity.
4. *Rating task*. Submit each image–prompt pair to the model and obtain excerpt ratings on four criteria.
5. *Comparative analysis*. Compare ratings within each demographic factor.

Figure 1.


"When you round to the nearest hundred, look at the tens digit. If it's 5 or more, round up. If it's 4 or less, round down. For example, 263 rounds to 300. Let's try with different numbers together."
"When you round to the nearest hundred, look at the tens digit. If it's 5 or more, round up. If it's 4 or less, round down. For example, 263 rounds to 300. Let's try with different numbers together."
"When you round to the nearest hundred, look at the tens digit. If it's 5 or more, round up. If it's 4 or less, round down. For example, 263 rounds to 300. Let's try with different numbers together."
"When you round to the nearest hundred, look at the tens digit. If it's 5 or more, round up. If it's 4 or less, round down. For example, 263 rounds to 300. Let's try with different numbers together."

Four Counterfactual Images Depicting Teachers That Vary Only in Race (Black/White) and Gender (Male/Female).

4 Experiments

The procedure above was implemented as follows.

4.1 Image Construction

Using OpenAI's DALL·E 3 (OpenAI, 2024), four teacher images were produced: a Black female, a White female, a Black male, and a White male. Images were iteratively refined with the goal of making them highly similar on non-demographic dimensions. Adobe Photoshop was used to overlay each teacher image onto the same classroom scene, aligning position and scale to preserve framing, pose, and facial expression (Adobe, 2025).

4.2 Excerpt Generation

GPT-4o (OpenAI, 2025) was accessed via the web interface to generate 1,000 excerpts per context (35–50 words each) with age-appropriate tone and content across four contexts: 3rd-grade English, 3rd-grade mathematics, graduate-level English literature, and graduate-level advanced mathematics (4,000 excerpts in total). Excerpts were generated in batches and conversation history was retained within each context to ensure excerpt uniqueness.

4.3 Counterfactual Pairing

Each excerpt was then superimposed onto each teacher image, yielding 16,000 images in total. Figure 1 illustrates four sample composites.

4.4 Rating Task

GPT-4o, accessed via the Azure OpenAI Service, was prompted to evaluate the excerpt embedded within each image. The deployment was configured with no data logging or model training from inputs. Default model settings were used. The exact prompt text was:

> *You are an expert in education and pedagogy. A classroom lecture excerpt is embedded within the attached image. Evaluate this excerpt using the following four criteria, each on a scale from 1 (worst) to 10 (best): Clarity; Student Engagement; Coherence and Organization; Pedagogical Effectiveness.*

The four criteria—Clarity, Student Engagement, Coherence and Organization, and Pedagogical Effectiveness—were adapted from prior research on instructional quality (Devlin & Samarawickrema, 2010; Zheng, 2021). Each image–prompt pair was submitted individually, and model-generated ratings were recorded.

4.5 Comparative Analysis

For each criterion, we compared paired ratings across race and across gender using the Wilcoxon signed-rank test. Within each teaching context, we conducted planned, paired Wilcoxon tests for race and for gender across the four criteria. For each factor within each context, the four tests formed a family, and we controlled the family-wise error rate at α = .05 using Holm's step-down procedure. Unless noted otherwise, significance refers to these factor-specific Holm-adjusted *p*-values as reported in Table 1. Effect sizes were computed as Wilcoxon *r* and interpreted as negligible (< .10) or small (.10–< .30).

Table 1.
Wilcoxon signed-rank tests for race and gender by context and criterion (Holm-adjusted per factor within context).

Context	Factor		Criteria			
			Clarity	*Student Engagement*	*Coherence and Organization*	*Pedagogical Effectiveness*
3rd-grade Math		*N (race/gender nonzero pairs)*	393 / 390	404 / 403	407 / 407	362 / 368
	Race	*Means (Black, White)*	8.349, 8.403	7.399, 7.357	8.290, 8.324	8.412, 8.378
		p (race, Holm)	**< .001*****	**0.006****	**0.003****	**0.035***
		r (race, label)	.234 (small)	.147 (small)	.163 (small)	.111 (small)
	Gender	*Means (Male, Female)*	8.351, 8.400	7.365, 7.390	8.278, 8.336	8.402, 8.388
		p (gender, Holm)	**0.001****	0.174	**< .001*****	0.174
		r (gender, label)	.178 (small)	.085 (negl.)	.191 (small)	.089 (negl.)
3rd-grade English		*N (race/gender nonzero pairs)*	502 / 487	511 / 509	505 / 491	312 / 329
	Race	*Means (Black, White)*	8.589, 8.623	7.740, 7.749	8.601, 8.630	8.783, 8.763
		p (race, Holm)	**0.045***	0.502	0.097	0.097
		r (race, label)	.113 (small)	.030 (negl.)	.095 (negl.)	.117 (small)
	Gender	*Means (Male, Female)*	8.588, 8.624	7.729, 7.760	8.596, 8.634	8.766, 8.780
		p (gender, Holm)	**0.012***	**0.039***	**0.012***	0.170
		r (gender, label)	.130 (small)	.104 (small)	.134 (small)	.076 (negl.)
Graduate-level Advanced Math		*N (race/gender nonzero pairs)*	574 / 571	522 / 509	567 / 564	588 / 598
	Race	*Means (Black, White)*	5.972, 6.053	4.521, 4.599	6.913, 6.987	5.809, 5.860
		p (race, Holm)	**< .001*****	**< .001*****	**< .001*****	**0.007****
		r (race, label)	.221 (small)	.238 (small)	.198 (small)	.112 (small)
	Gender	*Means (Male, Female)*	6.054, 5.971	4.595, 4.525	6.985, 6.914	5.887, 5.782
		p (gender, Holm)	**< .001*****	**< .001*****	**< .001*****	**< .001*****
		r (gender, label)	.221 (small)	.225 (small)	.190 (small)	.238 (small)
Graduate-level English Literature		*N (race/gender nonzero pairs)*	543 / 546	578 / 578	517 / 522	633 / 633
	Race	*Means (Black, White)*	6.770, 6.794	5.742, 5.770	7.639, 7.651	6.856, 6.878
		p (race, Holm)	0.302	0.262	0.444	0.444
		r (race, label)	.070 (negl.)	.077 (negl.)	.038 (negl.)	.049 (negl.)
	Gender	*Means (Male, Female)*	6.778, 6.786	5.743, 5.769	7.648, 7.643	6.859, 6.875
		p (gender, Holm)	0.946	0.282	0.946	0.935
		r (gender, label)	.031 (negl.)	.075 (negl.)	.010 (negl.)	.040 (negl.)

*Notes. Table presents paired Wilcoxon signed-rank tests comparing race (Black vs. White) and gender (Male vs. Female) for each criterion within each teaching context. For multiplicity, the four tests per context/factor (4 criteria × 1 context × 1 factor) are treated as a single family; p-values are Holm-adjusted per factor within context (two-sided, α = .05). Significance coding: p < .05 = *, < .01 = **, < .001 = ***. N reports the number of nonzero pairs contributing to each Wilcoxon test (shown as $N_{(race)}/N_{(gender)}$). Effect sizes are Wilcoxon $r = |Z|/\sqrt{n}$ and are reported as magnitudes (direction indicated by the corresponding group means); qualitative labels: negligible (negl.) < .10; small .10–.30.*

5 Results

Table 1 reports all four contexts (32 tests), with Holm's step-down adjustment applied per factor within context across the four criteria. Across the 32 tests, 18 were significant at $\alpha = .05$ after per-factor, within-context adjustment. Although no differences were detected in graduate-level English literature, that context remains in the table for completeness. Among the remaining three contexts, 18 comparisons are significant after Holm adjustment; effect sizes for significant tests are uniformly small ($r \approx .11$–$.24$).

By context (significant comparisons per factor / 4; Holm-adjusted per factor within context):

- 3rd-grade Mathematics. Race: 4/4 (White > Black on Clarity, Coherence and Organization, Pedagogical Effectiveness; Student Engagement shows Black > White); Gender: 2/4 (Female > Male on Clarity, Coherence and Organization). Student Engagement and Pedagogical Effectiveness: not significant for gender.

- 3rd-grade English. Race: 1/4 (White > Black on Clarity); Gender: 3/4 (Female > Male on Clarity, Student Engagement, Coherence and Organization). Pedagogical Effectiveness: not significant for either factor.

- Graduate-level Advanced Mathematics. Race: 4/4 (White > Black on all four criteria); Gender: 4/4 (Male > Female on all four criteria).

- Graduate-level English Literature. Race: 0/4; Gender: 0/4 (no differences on any criterion).

Where significant, gender effects favored Female at the elementary level (3rd-grade English and Mathematics) and Male in Graduate-level Advanced Mathematics; race effects generally favored White, with the noted exception of Student Engagement in 3rd-grade Mathematics (Black > White). Effect sizes for significant tests were uniformly small (Wilcoxon $r \approx .11$–$.24$).

6 Conclusion

6.1 Discussion

This study prompted a large vision-language model to evaluate identical lecture excerpts while the background image varied only in teacher race and gender. Under a per-factor, within-context multiplicity correction, the model's ratings differed by demographic attributes in three of the four contexts examined. The pattern was consistent with level-specific sensitivities: female teachers received higher ratings on Clarity and Coherence and Organization at the 3rd-grade level, whereas male teachers received higher ratings on all four criteria in graduate-level Advanced Mathematics. Race effects were more pervasive, typically favoring White teachers, with one notable exception (higher Student Engagement for Black teachers in 3rd-grade Mathematics). Although the magnitudes of the significant effects were small, the results demonstrate that formally irrelevant visual cues can systematically shift evaluative judgments of identical text.

6.2 Limitations

The counterfactual images operationalized binary gender and race among teachers with other attributes (e.g., approximate age, body type, pose, expression, attire, framing) held as constant as possible. Effects may differ for other demographic attributes (e.g., nonbinary genders, nonbinary races, age, weight, skin tone, disability) or other image features. Outputs were analyzed for a single model (GPT-4o); behavior may differ across LVLMs and versions. Finally, individual significant effects were small; cumulative effects, which could be consequential, require further study.

6.3 Future Work

Future work should (i) extend the demographic characteristics beyond binary gender and Black/White categories and test intersectional manipulations (interactions); (ii) include within-factor variability (e.g., represent each demographic cell with multiple distinct teacher images) (iii) evaluate multiple LVLMs and versions; (iv) broaden instructional domains and criteria; and (v) examine pipeline-level consequences—for example, whether small per-item biases aggregate into consequential differences in ranking or approval decisions. We note with concern that if these biases reflect patterns in human-produced training materials, then using such models in teacher evaluation risks a feedback loop: biased outputs influence decisions and documentation, which in turn affect future training data, which may further entrench demographic underrepresentation and bias.

Acknowledgments

The author thanks the National Board of Medical Examiners for supporting this work.

References

Adobe. 2025. Adobe Photoshop (Version 26.7) [Computer software]. Adobe Inc. URL: https://helpx.adobe.com/photoshop/kb/fixed-issues-history.html (May 2025 release; accessed 2025-06-12).

Jiaqi An, Di Huang, Chen Lin, and Ming Tai. 2025. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. PNAS Nexus 4(3):pgaf089.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint* arXiv:2406.11665.

Armin Bateni, Michael C. Chan, and Richard Eitel-Porter. 2022. AI fairness: From principles to practice. *arXiv preprint arXiv:2207.09833*.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94(4):991–1013.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marcia Devlin and Gita Samarawickrema. 2010. The criteria of effective teaching in a changing higher education context. *Higher Education Research & Development* 29(2):111–124. Taylor & Francis Online.

Kathleen C. Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*: Long Papers, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Johann D. Gaebler, Sharad Goel, Aziz Z. Huq, and Prasanna Tambe. 2025. Auditing large language models for race & gender disparities: Implications for artificial intelligence–based hiring. *Behavioral Science & Policy* 10(2):46–55. https://doi.org/10.1177/23794607251320229.

Michelle R. Greene, Mariam Josyula, Wentao Si, and Jennifer A. Hart. 2025. Digital divides in scene recognition: Uncovering socioeconomic biases in deep learning systems. *Humanities and Social Sciences Communications 12:*414.

Usman Gohar and Lu Cheng. 2023. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969.*

Stephen Jay Gould. 1981. *The Mismeasure of Man*. New York, NY: W.W. Norton & Company.

Kianoush Hamidieh, Haoxiang Zhang, William Gerych, Taylor Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*, pages 547–561. ACM.

Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. 2025. Uncovering bias in large vision-language models at scale with counterfactuals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5946–5991, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.305

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When Tom eats kimchi: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.c3nlp-1.11

Min Hyungsuk Lee and Seunghyun Jeon. 2024. Vision-language models generate more homogeneous stories for phenotypically Black individuals. *arXiv preprint arXiv:2412.09668*.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: A three-layered approach. *AI and Ethics* 4(4):1085–1115. SpringerLink

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.416

OpenAI. 2024. DALL·E 3. URL: https://openai.com/dall-e-3/ (Accessed 2025-06-12).

OpenAI. 2025. GPT-4o [Large language model]. URL: https://platform.openai.com/ (Accessed 2025-06).

Rachel Schwartz, Jonathan Fiscus, Kenneth Greene, George Waters, Rupa Chowdhury, Thomas Jensen, et al. 2024. The NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan. U.S. National Institute of Standards and Technology. (Program materials and overview.) NIST AI Challenge Problems+1

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics. https://aclanthology.org/N18-2003/

Jing Zheng. 2021. A functional review of research on clarity, immediacy, and credibility of teachers and their impacts on motivation and engagement of students. *Frontiers in Psychology* 12:712419.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A study of stereotypical bias in pre-trained vision–language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.aacl-main.40

Enhancing Item Difficulty Prediction in Large-scale Assessment with Large Language Model

Mubarak Mojoyinola[1]**, Olasunkanmi James Kehinde**[2]**, Judy Tang**[3]
[1]Psychological and Quantitative Foundations, University of Iowa, Iowa City
[2] Department of Psychology, Norfolk State University
[3]Survey Management and Design, Westat
[1]`mubarak-mojoyinola@uiowa.edu`, [2]`ojkehinde@nsu.edu`, [3]`JudyTang@westat.com`

Abstract

Item difficulty prediction remains a critical challenge in large-scale assessment development, particularly for international programs like TIMSS where extensive pretesting is costly and time-consuming. This study investigated the utility of large language model (LLM)-extracted cognitive features for predicting item difficulty in mathematics assessment. We analyzed restricted-use TIMSS mathematics items from Grades 4 and 8, comparing three XGBoost models: traditional features (metadata and textual complexity), LLM-extracted cognitive features, and a combined approach. Traditional features alone achieved moderate performance (R^2 = 0.36), while LLM-extracted cognitive demand variables showed weaker individual performance (R^2 = 0.20). However, the combined model substantially outperformed both individual approaches, explaining 48% of variance in item difficulty, a 33% improvement over traditional methods alone. Results demonstrate that LLM-extracted features provide complementary predictive information that enhances difficulty prediction when integrated with conventional textual and metadata features. This approach offers a scalable alternative to expert-based cognitive analysis while maintaining theoretical grounding in established assessment frameworks.

1 Introduction

The calibration of item difficulty is a foundational and resource-intensive requirement in any assessment development. For large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS), the cost is even higher as items need to be pretested across multiple education systems participating in the program (von Davier et al., 2024). The conventional psychometric process relies on extensive field testing to gather empirical data, a practice that presents significant logistical and financial burdens, thereby creating a bottleneck in the test development lifecycle. This operational challenge has motivated a sustained search for automated methods capable of predicting item difficulty directly from item text, aiming to augment and streamline the item development process (AlKhuzaey et al., 2024). Early research in this area leveraged traditional Natural Language Processing (NLP) techniques to extract surface-level and engineered linguistic features. These included readability indices (e.g., Flesch-Kincaid), word and sentence counts, syntactic complexity metrics, and psycholinguistic features from tools like Coh-Metrix (AlKhuzaey et al., 2024). While valuable, these methods do not capture the deeper conceptual, cognitive, and reasoning demands embedded within an assessment item.

The advent of Large Language Models (LLMs) represents a paradigm shift in this domain. Three dominant strategies for using LLMs for item difficulty prediction have emerged. The first, direct estimation, involves prompting an LLM to act as a subject-matter expert and assign a holistic difficulty rating to an item. While intuitively appealing, this approach functions as a "black box," providing a score without rationale, and has shown inconsistent performance, particularly for items designed for younger learners (Razavi and Powers, 2025). The second approach involves using LLMs to generate text embeddings from item stems and response options, which then serve as features in a machine learning model. This approach has been shown to produce accurate difficulty predictions (Bulut et al., 2024; Kapoor et al., 2025). However, models based on embeddings lack interpretability. The third and most sophisticated strategy treats the LLM as a feature extractor. In this two-stage process, the LLM is guided by a structured prompt to analyze an item and output values for a set of predefined, interpretable cognitive and linguistic features. These features are then used as predictors in a separate, often simpler, machine learning model (Razavi and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 218–222
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

Powers, 2025).

This study builds upon the promising feature-based methodology by addressing key limitations in current LLM applications for item difficulty prediction. While previous research has demonstrated the potential of LLM-extracted features, significant gaps remain in applying these approaches to complex, international assessment contexts such as TIMSS mathematics items that span multiple grade levels and cognitive domains. Furthermore, existing studies have primarily focused on either traditional psychometric features or LLM-derived metrics in isolation, without systematically investigating how these complementary approaches can be integrated to enhance predictive accuracy. This investigation addresses these limitations by developing a comprehensive framework that combines traditional textual and metadata features with LLM-extracted cognitive demand variables, providing empirical evidence for the added value of automated cognitive feature extraction in large-scale assessment contexts.

Research Questions

1. How do large language model-extracted cognitive features compare to traditional textual complexity features in predicting item difficulty in large-scale assessments?

2. To what extent do LLM-extracted cognitive demand features enhance item difficulty prediction when combined with traditional features?

2 Related Works

Research on item difficulty modeling has long emphasized the integration of psychometric and cognitive frameworks. Sheehan and Mislevy (1994) applied tree-based regression analyses to link item features, solution processes, and response formats with IRT parameters, explaining up to 36% of the variance in difficulty. Competency-based approaches have been particularly influential in international large-scale assessments. Turner et al. (2013) demonstrated that six mathematical competencies (e.g., reasoning, modeling, symbol use) strongly predicted item difficulty in PISA. Similarly, Schneider et al. (2013) showed that Depth of Knowledge (DOK), reading load, and contextual demands systematically predicted item difficulty. With the rise of machine learning, feature-based approaches have advanced prediction. Štěpánek et al. (2023) compared multiple algorithms and found that elastic net and random forests outperformed expert ratings, suggesting that textual features can approximate empirical difficulties. Yi et al. (2024) extended this work with an XGBoost-SHAP framework, achieving strong predictive accuracy while offering interpretability by quantifying the contribution of features such as reasoning steps and symbolic complexity. While these studies demonstrated the importance of cognitive and textual features, they relied on experts to manually code the item data for cognitive features, thereby limiting the number of items that could be studied efficiently.

3 Method

3.1 Dataset

The dataset for this investigation comprised 202 restricted-use mathematics test items selected from TIMSS Grades 4 and 8 assessments administered in 2015 and 2019. These items spanned nine mathematical content areas, with the Number domain contributing the highest number of items, while roughly 23% were eTIMSS items delivered through digital platforms. Item difficulty was quantified using international average proportion-correct values (p-values) obtained from the TIMSS International Database. The mean proportion correct was 51.63%.

3.2 LLM

Using OpenAI's GPT-4.1, several cognitive features were extracted from each item. Leveraging the reasoning capacity of GPT-4.1, we instructed the model through few-shot prompting to evaluate the items and provide appropriate rating based on the provided detailed rubrics with numerical scales and specific criteria for consistent rating across items. GPT-4.1 was accessed through OpenAI's APIs using the ellmer package in R

3.3 Item Features

This study examined variables that could be systematically categorized based on their extraction methodology: traditional features derived from conventional computational and metadata approaches, and LLM-extracted features leveraging large language model capabilities for automated item analysis.

Traditional Features encompass two subcategories of variables that have been extensively used in prior research on item difficulty modeling.

Metadata Variables refer to characteristics specified during item development and assessment administration. These included grade (4 or 8), item type (multiple choice or constructed response), content domain (number, algebra, geometry, data and probability), cognitive domain (knowing, applying, reasoning), and presence of visual elements. These variables align with established TIMSS framework specifications and capture basic structural features of assessment items.

Textual Complexity Variables captured the linguistic demands of item stems and response options using established computational linguistics approaches. These included basic text statistics such as character count, word count, sentence count, and syllable count, as well as established readability indices including the Automated Readability Index, SMOG readability formula, Coleman-Liau index, Flesch Reading Ease, and Gunning Fog index. Additional variables measured the frequency of digits in item stem and response options, recognizing that mathematical text presents unique processing demands beyond general readability.

LLM-Extracted Features represent a novel approach to automated item difficulty modeling, leveraging the reasoning capabilities of large language models to extract features that traditionally required expert human judgment.

Mathematical Content Features were extracted by prompting the LLM to identify and categorize abstract mathematical concepts present in each item. The LLM was also asked to provide a difficulty rating on a 0-100 scale based on its analysis of the mathematical content and cognitive demands, serving as an AI-generated difficulty estimate.

Cognitive Demand Variables were extracted using the LLM to rate items according to the four cognitive competencies framework developed by Turner et al. (2013) for PISA assessment. The LLM was prompted to evaluate each item's demand for: reasoning and argumentation, problem solving, mathematical modeling, and communication. These competencies describe essential cognitive processes required for successful mathematical problem solving and have demonstrated high predictive validity in prior research, with competency-based variables explaining approximately 70% of the variance in PISA item difficulty when used in regression models (Turner et al., 2013). The LLM-based extraction approach offers a scalable alternative to expert panel ratings while maintaining theoretical grounding in established cognitive frameworks.

3.4 Modeling

Using the proportion correct as our estimate of item difficulty, we built a tree-based ensemble model to map item features to difficulty estimates. Specifically, we employed Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), a machine learning algorithm that has demonstrated superior performance across diverse prediction tasks, including item difficulty prediction (Yi et al., 2024; Lamgarraj et al., 2024).

XGBoost operates through a sequential ensemble approach, iteratively constructing weak decision trees where each subsequent model focuses on correcting the prediction errors made by the previously constructed models. This additive modeling strategy allows the algorithm to capture complex, non-linear relationships between item features and difficulty that traditional linear regression approaches might miss. The sequential nature of the boosting process enables the model to progressively refine its predictions by learning from residual errors, ultimately producing a highly accurate composite predictor.

The model's ability to provide feature importance rankings offers valuable insights for assessment development. By quantifying the relative contribution of different item characteristics to difficulty prediction, XGBoost can inform item writers about which features most strongly influence item difficulty, potentially improving the efficiency of item development processes. This interpretability is particularly valuable when comparing the predictive utility of traditional features versus novel LLM-extracted features.

The dataset was randomly partitioned into training (80%) and testing (20%) sets to enable robust model evaluation. Hyperparameter optimization for the XGBoost model was conducted using 5-fold cross-validation with grid search on the training set, ensuring that model selection decisions were based on generalizable performance rather than overfitting to specific data partitions.

Model performance was evaluated using two metrics: root mean squared error (RMSE) to quantify the magnitude of prediction errors, and coefficient of determination (R^2), to quantify the proportion of variance in item difficulty explained by the model.

4 Results

Three XGBoost models were developed to evaluate the predictive utility of different feature categories for item difficulty prediction: a text-based model using traditional features, a cognitive model using LLM-extracted features, and a comprehensive model combining both feature types. Table 1 presents the performance metrics for all three models based on the testing dataset.

Table 1: Model performance metrics

Feature	**RMSE**	**R^2**
Traditional	15.43	0.36
LLM Cognitive	17.27	0.20
Traditional + LLM Cognitive	14.03	0.48

The text-based model, utilizing traditional metadata and textual complexity variables, demonstrated moderate predictive performance on the test set (R^2 = .36, RMSE = 15.43). This model effectively captured the linguistic and structural characteristics of assessment items that influence difficulty, including readability indices, word counts, and domain specifications.

In contrast, the cognitive model using only LLM-extracted features showed weaker individual performance (R^2 = 0.20, RMSE = 17.27). While the cognitive demand variables and mathematical content features extracted by the LLM captured theoretically important aspects of item difficulty, these features alone were insufficient for accurate difficulty prediction.

The combined model incorporating both traditional and LLM-extracted features substantially outperformed either individual approach, achieving the highest test set performance (R^2 = 0.48, RMSE = 14.03). This represents a 33% improvement in explained variance compared to the text-only model and a 140% improvement compared to the cognitive-only model. The superior performance of the integrated approach demonstrates that LLM-extracted cognitive features provide unique predictive information that complements traditional item characteristics.

Figure 1 displays the relative importance of the top ten features for predicting mathematics item difficulty according to the best-performing XGBoost model that combined traditional and LLM-extracted features. The LLM rating of item difficulty emerged as the most important predictor, followed by item type and a series of readability indices.

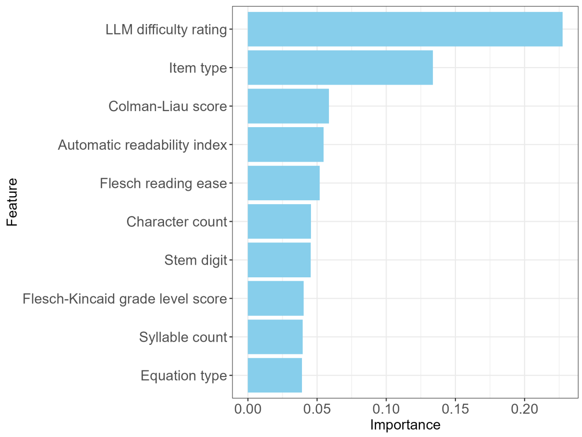


Figure 1: Feature Importance Plot

5 Discussion

The findings from this study demonstrate that integrating LLM-extracted features with traditional item characteristics yields clear benefits for predicting difficulty in TIMSS mathematics assessments. Traditional features alone achieved moderate accuracy (R^2 = 0.36), a result consistent with earlier research that relied on readability indices and metadata such as item format and domain (e.g., Sheehan and Mislevy (1994); Schneider et al. (2013)). LLM-extracted features on their own, while grounded in cognitive frameworks similar to those emphasized by Turner et al. (2013), showed weaker predictive performance, suggesting that automated cognitive coding alone is not sufficient to achieve excellent predictive performance. However, the model with combined feature types explained 48% the variance in item difficulty, surpassing both approaches in isolation and aligning with the findings of Štěpánek et al. (2023) and Yi et al. (2024), who showed that hybrid models incorporating multiple feature sets outperform single source predictors. These results suggest that LLM-derived cognitive measures capture unique dimensions of difficulty that complement rather than replace traditional text and metadata indicators.

The study also reinforces and extends findings from recent LLM-based work. Razavi and Powers (2025) demonstrated that LLMs can enhance difficulty prediction when used as feature extractors for tree-based models, while Li et al. (2025) found that fine-tuned smaller models often outperform large general-purpose LLMs in educational contexts. Our results resonate with these studies by showing that raw LLM predictions are insufficient but that their structured cognitive features add significant value when combined with traditional descriptors.

6 Conclusion

This study provides new evidence on the potential of LLM-extracted features to improve item difficulty prediction in international large-scale assessments such as TIMSS. While traditional metadata and textual complexity variables accounted for a moderate proportion of variance in item difficulty, and LLM-extracted cognitive features alone showed limited predictive value, their integration substantially enhanced accuracy, explaining nearly half of the variance in item difficulty. These results confirm that LLMs capture complementary aspects of cognitive demand and reasoning that extend beyond conventional text-based measures, offering a scalable alternative to manual coding. The findings also reinforce prior evidence that hybrid models outperform single-source predictors and demonstrate that combining psychometric, linguistic, and cognitive perspectives is essential for advancing item modeling.

By carefully choosing theoretically grounded cognitive demand features, this research shows how LLM-extracted features can provide not only stronger predictions but also actionable insights into the cognitive and structural elements driving item difficulty. Together, these contributions respond to calls in prior work for approaches that balance predictive power with interpretability, bridging psychometric traditions with modern NLP advances. Ultimately, the study offers a practical pathway for improving efficiency in item calibration, reducing reliance on costly pretesting, and enhancing the design of equitable and cognitively grounded assessments in mathematics education.

References

S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of usmle items.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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*.

Mohamed Lamgarraj, Céline Joiron, Aymeric Parant, and Gilles Dequen. 2024. Exploring item difficulty prediction: Data driven approach for item difficulty estimation. In *International Conference on Intelligent Tutoring Systems*, pages 415–424. Springer.

Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, page 00131644251344973.

P. Razavi and S. J. Powers. 2025. Estimating item difficulty using large language models and tree-based machine learning algorithms. *arXiv preprint arXiv:2504.08804*.

M Christina Schneider, Kristen L Huff, Karla L Egan, Margie L Gaines, and Steve Ferrara. 2013. Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, 18(2):99–121.

Kathleen Sheehan and Robert J Mislevy. 1994. A tree-based analysis of items from an assessment of basic mathematics skills.

Lubomír Štěpánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19):4104.

Ross Turner, John Dossey, Werner Blum, and Mogens Niss. 2013. Using mathematical competencies to predict item difficulty in pisa: A meg study. In *Research on PISA: Research outcomes of the PISA Research Conference 2009*, pages 23–37. Springer.

M. von Davier, B. Fishbein, and A. Kennedy, editors. 2024. *TIMSS 2023 Technical Report (Methods and Procedures)*. Boston College, TIMSS & PIRLS International Study Center, Boston.

Xifan Yi, Jianing Sun, and Xiaopeng Wu. 2024. Novel feature-based difficulty prediction method for mathematics items using xgboost-based shap model. *Mathematics*, 12(10):1455.

Leveraging LLMs for Cognitive Skill Mapping in TIMSS Mathematics Assessment

Ruchi J. Sachdeva, Ph.D., Pearson
Jung Yeon Park, Ph.D, George Mason University

Abstract

This study evaluates GPT-4 for generating and validating Q-matrices for TIMSS mathematics items. Using expert mappings as benchmarks, we examined prompt design, temporal stability, and error detection. The model showed strong accuracy, substantial reliability, and effective recovery of altered skills, supporting its role as a scalable aid to cognitive diagnosis.

1 Background

Identifying the cognitive skills required to solve specific test items is a foundational task in educational assessment. This function is not only critical for ensuring the validity of test interpretations but is especially central to the development and validation of Q-matrices in Cognitive Diagnosis Models (CDMs; Rupp et al., 2010).

A carefully constructed Q-matrix aligns each assessment item with the exact constellation of cognitive skills or knowledge components required for its solution. When the identification of these skills is flawed or incomplete, the resulting model can misrepresent learners' proficiencies, undermining one of the principal strengths of cognitive diagnostic models, their capacity to deliver precise, actionable feedback. Consequently, the delineation of content and associated skills emerges as a critical, yet cognitively demanding, dimension of assessment design. Against this backdrop, large language models (LLMs), such as GPT, present a promising avenue for augmenting or supporting expert analysis, offering new opportunities to enhance the rigor and efficiency of Q-matrix development.

2 Purpose

This study evaluates GPT-4's capacity to identify skills and validate Q-matrices against a content expert–designed gold standard across Number, Algebra, Geometry, and Data and Chance. Q-matrix design requires more than simple skill matching, demanding analysis of interactions, hierarchies, cognitive load, and item context. We examine whether GPT-4 can meet these demands as a scalable, cost-effective aid to expert assessment design. This study addresses the following research questions:
RQ1: What prompt strategies enable GPT-4 to accurately map cognitive skills to test items?
RQ2: How stable are its Q-matrices across repeated prompts?
RQ3: How does GPT-4o's performance vary across different error types (skill addition and skill swapping)?

3 Sample

This study uses the TIMSS 2007 Grade 7 Mathematics Released Items with an expert-defined Q-matrix for 89 publicly available items (Johnson et al., 2013) approved by NCES to strengthen the generalizability and relevance of its findings in large-scale educational research.

4 Methodology

A three-part framework was employed to refine methods for generating and validating Q-matrix skill mappings for TIMSS mathematics items. First, two prompt templates were designed: one to construct Q-matrix entries for all 89 items and another to evaluate existing matrices for errors, each targeting a distinct cognitive mapping task.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 223–228
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

Second, we assessed agreement on the number of skills per item by comparing GPT estimates with counts from the content expert gold standard. We also tested four prompt configurations to gauge their effect on Q-matrix accuracy, contrasting a full version containing all optimization elements with simplified versions that excluded skill descriptions, expert-role instruction, or procedural constraints.

Third, we evaluated the reliability of a single fixed prompt by generating 20 independent item-by-skill mappings and measuring consistency across runs. This framework offers a rigorous basis for understanding how prompt design and refinement influence the precision and diagnostic quality of automated Q-matrix construction in educational assessment research.

4.1 Prompt Design

We refined prompt structures with a subset of 10 TIMSS items, then applied the finalized versions to all 89 items to generate Q-matrix entries. Because submitting all items at once exceeded the token limit, each was provided individually with its prompt and image.

Two prompts were developed: one to generate an independent Q-matrix and another for validation. Both returned JSON-formatted outputs containing the item identifier, the corresponding Q-matrix entry, the number of skills identified, and a brief rationale for each skill selection.

For validation, a Q-matrix entry with an intentional error was presented, and GPT was asked to judge its accuracy and produce a corrected mapping for the item. The validation output also included a one-sentence explanation of the chosen skills, and a one-sentence justification for agreeing or disagreeing with the provided Q entry for the item.

4.2 Testing Prompt Strategies

To evaluate the impact of various prompt optimization strategies, we compared GPT-generated Q-matrices with and without these enhancements, holding all other conditions constant to isolate the effects of prompt structure. Each prompt variant produced a distinct Q-matrix, which was benchmarked against a content expert-Q matrix.

Four prompt configurations were evaluated for their effect on Q-matrix accuracy. (See Table 1) The full prompt included all optimization elements for constructing an 89-item, nine-skill matrix. Other versions removed detailed skill descriptions, omitted the instruction assigning an expert role, or removed all the procedural constraints, which provided guidance on selecting a primary skill, adding additional skills, and interpreting graphs and shapes, leaving only the core task specification.

Method	**Status**	**Description**
Full Prompt (P1)	Kept All	Used all prompt techniques to produce the final version for an 89-item, 9-skill Q-matrix.
Skill Details (P2)	Removed	Detailed descriptions of each skill (Taken from the TIMSS technical documentation).
Providing a Role (P3)	Removed	Omitted the instruction: *"You are a content expert in 7th-grade Mathematics assessment in the United States."*
Procedural Constraints (P4)	Removed	Eliminated procedural rules about task limits, content-domain mapping, leaving only the high-level task description.
Evaluation Metrics	**F1-score comparisons against expert-designed Q-matrix at the skill level.**	

Table 1: Framework for Prompt Engineering Strategies in Q-Matrix Generation.

This evaluation framework allowed systematic identification of prompt configurations and their impact on accuracy and informing best practices. Performance was assessed using F1 scores, with precision and recall calculated at the overall and skill level against the content expert-designed Q-matrix.

4.3 Number of Required Skills

A Q-matrix maps assessment items to the cognitive skills needed for their solution. After defining skills from learning objectives, each item is reviewed to determine how many are required. Using the same expert-defined skills, GPT estimated this count, and we compared its results with expert judgments using intraclass correlation (ICC) to assess agreement on item complexity. Treating each item–skill decision (i.e., whether an item is linked to a given skill; coded 0/1)

4.4 Repeated Prompt Reliability

LLMs can map assessment items to cognitive skills, yet their outputs may fluctuate across identical runs, raising concerns about the reliability of Q-matrices in cognitive diagnostic models. To evaluate this, we tested temporal stability by administering the same prompt to GPT twenty times and calculating Krippendorff's alpha (Krippendorff, 2018). This statistic measures the degree of agreement among multiple coders or iterations beyond chance and accommodates various data types, including nominal scales. It was appropriate here because the entire binary skill vector (for example, 100100000) was treated as a single nominal category. This approach provided an estimate of consistency across repeated mappings and the robustness of GPT-generated Q-matrices.

4.5 Q-Matrix Error Detection Using Add and Swap Conditions

To assess GPT's ability to detect and correct errors in skill mappings, we used two procedures called Add and Swap (Table 2). Both began with the expert Q-matrix for each TIMSS item; 74 of 89 items involved a single skill, offering a clear test case. GPT received the skill definitions, learning objectives, detailed prompt, and a PNG of the item.

In the Add condition, a randomly selected extra skill was appended to the correct mapping to create an altered entry. GPT evaluated agreement or disagreement with the provided Q-matrix entry, identified the appropriate skill pattern for the item, and offered a rationale when its judgment differed. In the Swap condition, one correct skill was replaced with an unrelated alternative, and GPT assessed the substitution, stated agreement or disagreement, and proposed the correct skill or set of skills.

GPT's recommendations were compared with the content-expert mappings, and accuracy was assessed with F1 scores. Illustrative examples and procedural details for both conditions are presented in Table 2.

Condition	Purpose	Procedure	Illustrative Example
Add	Evaluate GPT's ability to identify and discard unnecessary skills in Q-matrix mappings.	GPT reviewed an augmented skill set containing one unnecessary element and determined whether to retain or remove it before producing the final mapping.	Example: For a geometry item, the expert selected only Skill 6. The Q-matrix listed Skills 6 and 1; GPT removed Skill 1, retaining Skill 6 in line with the expert mapping.
Swap	Examine GPT's capacity to identify and correct an entirely different (incorrect) skill.	For each (mostly single-skill) item, the correct skill was swapped with an unrelated one; GPT reviewed the materials and proposed the appropriate skill(s).	Example: In a number item, the correct Skill 1 was swapped for Skill 5; GPT removed Skill 5 and reinstated Skill 1.

Table 2: Procedures for Error Detection (Add and Swap)

5 Results

This section presents findings on the accuracy, reliability, and interpretability of Q-matrices generated by GPT. We report performance across prompt strategies, agreement on the number of skills per item, detection of Q-Matrix errors, and stability across repeated runs, highlighting how prompt design influenced outputs.

5.1 Prompt Optimization

Each prompt produced a distinct Q-matrix, which was evaluated against a content expert–designed

Q-matrix using F1, precision, and recall. Table 1 summarizes the prompt-engineering strategies, and Table 3 reports their effects on accuracy and partial credit. The Full Prompt (P1) yielded the highest performance, with an overall F1 of 0.83 and a partial-credit mean of 0.82. P2 (No Skill Details) achieved an F1 of 0.78 and a partial-credit mean of 0.76, while P3 (No Role Assignment) showed an F1 of 0.78 and a partial-credit mean of 0.81. P4 (No Constraints) produced the largest decline, with an F1 of 0.77 and a partial-credit mean of 0.80. F1 measured exact agreement, whereas partial-credit scores reflected overlap; the two metrics were similar because most items targeted a single skill. The relatively high partial-credit scores for P3 and P4 suggest that, although omitting these prompting techniques preserved the identification of the primary skill, it reduced the detection of secondary or supporting skills when key prompt elements were excluded.

Skill	**P1**	**P2**	**P3**	**P4**
Skill 1: Whole Numbers & Integers	0.85	0.71	0.67	0.86
Skill 2: Fractions, Decimals & Percents	0.91	0.91	0.92	0.95
Skill 3: Ratios & Proportions	0.81	0.80	0.80	0.67
Skill 4: Patterns	0.67	1.00	0.86	0.50
Skill 5: Expressions, Equations & Functions	0.96	1.00	0.96	0.88
Skill 6: Lines, Angles and Shapes	0.77	0.74	0.78	0.88
Skill 7: Measurement	0.88	0.50	0.67	0.80
Skill 8: Location & Movement	0.68	0.55	0.55	0.60
Skill 9: Data and Chance	0.93	0.80	0.81	0.81
Full Sample F1	**0.83**	**0.78**	**0.78**	**0.77**
Mean Partial-Credit Score	**0.82**	**0.76**	**0.81**	**0.80**

Table 3: F1 Score & Partial Credit Score by Skill.

Across individual skills, the strongest accuracy was observed for Expressions, Equations, and Functions (F1 = 0.88–1.00) and Fractions, Decimals, and Percents (F1 > 0.90 across all prompts). Data and Chance also showed consistently high performance (F1 = 0.80–0.93). By contrast, Location and Movement had the lowest scores (0.55–0.68), and Patterns displayed substantial variation across prompts (0.50–1.00). Measurement showed moderate sensitivity to prompt design, ranging from 0.50 to 0.88. An F1 value of 1 indicates perfect agreement between the prompt-generated Q-matrix and the content expert–designed Q-matrix. These findings suggest that skills such as Expressions, Equations, and Functions and Fractions, Decimals, and Percents are relatively stable across prompts, whereas skills like Location and Movement and Patterns, which often involve graphs and diagrams, are more susceptible to changes in prompt constraints.

5.2 Number of Skills Agreement

A Q-matrix links assessment items to the cognitive skills required for their solution. After defining skills from the learning objectives, each item was reviewed to determine the number of unique skills involved—a task that is both challenging and essential for accurate measurement. GPT analyzed the items and estimated the total number of skills required, and this estimate was compared with the corresponding counts from the content expert–designed Q-matrix. Agreement between GPT and the content expert–designed Q-matrix was evaluated using intraclass correlation (ICC) to assess item complexity. Treating each item–skill decision (0 = not linked, 1 = linked) as a subject and the two raters (content expert–designed Q-matrix and GPT; N = 2,047) as judges, single-rater ICCs (ICC1/2/3) were 0.72 (95% CI [0.70, 0.74]). When ratings were averaged, ICCs (ICC1k/2k/3k) increased to 0.84 (95% CI [0.82, 0.85]). Since many of the items involved only a single skill, future work should examine Q-matrices containing a larger proportion of items that require multiple skills.

5.3 Temporal Stability of Q-Matrix Skill Mappings

Large language models (LLMs) can map assessment items to cognitive skills, but their outputs may vary across identical runs, making reliability a critical concern for Q-matrices in cognitive diagnostic models (CDMs). To examine temporal stability, we administered the same prompt to GPT 20 times and computed

Krippendorff's alpha (Krippendorff, 2018). GPT demonstrated strong consistency, with mean alphas of 0.86 (exact match), 0.94 (Hamming distance), and 0.93 (Jaccard similarity). Sixty-five of 89 items (73%) were identical across runs, most remaining items exceeded α = 0.80, and only a few showed lower agreement (exact-match α as low as 0.11). Items with the poorest reliability were multipart questions in which extensive information was presented on a single page, suggesting that reading complex PNGs with many components may impair repeatability. Item-level results can be provided upon request.

5.4 Detecting Q-Matrix Error (Add and Swap Conditions)

GPT showed strong performance in detecting and correcting altered skill assignments in both the Add and Swap conditions (Table 4). Accuracy was highest for Expressions, Equations and Functions and Data and Chance. Moderate scores, with F1 values ranging from 0.75 to 0.90, were observed for Whole Numbers and Integers (Skill 1), Fractions, Decimals and Percents (Skill 2), Ratios and Proportions (Skill 3), Lines, Angles and Shapes (Skill 6), and Measurement (Skill 7).

Skill	Description	Add	Swap
1	Whole Numbers & Integers	0.85	0.86
2	Fractions, Decimals, & Percents	0.90	0.88
3	Ratios & Proportions	0.81	0.82
4	Patterns	0.61	0.67
5	Expressions, Equations, & Functions	0.96	0.97
6	Lines, Angles, & Shapes	0.77	0.78
7	Measurement	0.75	0.82
8	Location & Movement	0.55	0.56
9	Data and Chance	0.87	0.88
	Overall F1	**0.78**	**0.80**

Table 4: Performance in Detecting and Correcting Altered Skill Assignments.

Lower accuracy emerged for Patterns (Skill 4) and Location and Movement (Skill 8), where F1 scores were consistently below 0.70 across both conditions. For Lines, Angles and Shapes, performance was also slightly reduced, which may reflect current challenges in interpreting graphs, diagrams, and geometric figures. These areas may benefit from enhanced visual-processing capabilities or additional expert review to ensure reliable skill detection. Overall, the findings indicate that GPT can accurately identify and correct altered skill assignments, particularly in numerical and algebraic contexts, while tasks involving geometry and spatial reasoning may require refined prompts or closer collaboration with human experts.

6 Conclusion

This study demonstrates that GPT-4 can meaningfully support the cognitively demanding task of Q-matrix construction and validation. When provided with explicit skill definitions, structured prompts, and item images, GPT achieved high agreement with content expert mappings (F1 = 0.83) and substantial reliability across repeated runs (α ≈ 0.86). It also detected and corrected injected errors in the add and swap conditions with moderate to strong accuracy, particularly in number and algebra content domains. Performance declined for geometry and spatial-reasoning items, suggesting that visual interpretation remains a limiting factor. Items with heavy reading loads, multi-part content presented on a single page, or complex graphical information in PNG format also showed weaker repeatability, indicating that such features may challenge the model's consistency. Overall, these findings suggest that large language models, when carefully prompted, can offer scalable and replicable assistance in skill identification, complementing rather than replacing expert judgment. Future research should focus on refining methods for items with extensive text, multi-part layouts, or intricate visual elements to improve performance in these areas.

References

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30(2), 225-250.

Johnson, M.S., Lee, Y.-S., Sachdeva, R., Zhang, Z., & Park, J.Y. (2013, April). Examination of Gender Differences using the Multiple Groups DINA Model [Paper presentation]. *National Council on Measurement in Education*, San Francisco CA.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Thousand Oaks, CA: Sage.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). TIMSS 2007 *International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* TIMSS & PIRLS International Study Center, Boston College.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic model of attribute profiles: G-DINA model framework. *Educational Psychology*, 35(8), 1088-1110.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.

Appendix A: Skills, Content Domains

Skills	**Content Domain**
1 – Whole Numbers & Integers;	Numbers
2 – Fractions, Decimals, & Percents	Numbers
3 – Ratios & Proportions	Numbers
4 – Patterns	Algebra
5 – Expressions, Equations, & Functions	Algebra
6 – Lines, Angles, & Shapes	Geometry
7 – Measurement	Geometry
8 – Location & Movement	Geometry
9 – Data Analysis & Probability	Data and Chance

Author Index

Abbott, Jordan, 25
Adegoke, Hope Oluwaseun, 185
Adler, Rebecca, 20
Aldib, Roula, 84

Baldwin, Peter, 211
Barlas, Natasha Sohail, 14
Becker, Lee, 155
Belzak, William, 90
Betts, Joe, 141, 197
Bolt, Daniel, 204
Bruno, James V, 155
Burleigh, L, 20

Chen, Jie, 147
Chukharev, Evgeny, 135
Colbert, Alexandra Barry, 98
Correnti, Richard, 49
Crossley, Scott, 20

Dicerbo, Kristen, 40
Du, Ying, 185
Dugdale, Debbie, 191
Durkin, Kelley, 20
Dwyer, Andrew, 185

Fan, Kevin, 84
Firetto, Carla, 125
Flammia, Michele, 135

Grimaldi, Phillip, 40
Gulbahar, Yasemin, 34

Hill, Kelli Millwood, 40
Hogg, Charles, 40
Hong, Michelle, 49
Huang, Yue, 71, 147
Hutchins, Natalie, 117

Javaid, Hammad, 14
Jiang, Ning, 147
Jirout, Jamie J, 117
Johnson, Evelyn, 66

Karumbaiah, Shamya, 204
Kehinde, Olasunkanmi James, 218
Khalid, Eman, 14
Kim, Hyunjoo, 166
King, Jules, 20

Li, Hongli, 84
Li, Min, 178
Li, Tianwen, 49
Liao, Manqian, 107
Litman, Diane, 49
Liu, Huan, 66
Liu, Ruidong, 107
Liu, Zhexiong, 49
Lottridge, Susan, 191

Marchong, Chad, 84
Matsumura, Lindsay Clare, 49
Mattox, Madeline Rose, 117
Meng, Lionel Hsien, 204
Miller-Vedam, Maya, 40
Mojoyinola, Mubarak, 218
Muntean, William, 141, 197
Murphy, P. Karen, 125

Niu, Chenhao, 90, 107

Onuorah, Abraham, 135
Ortmann Lee, Angel, 90, 107
Oyler, Joe, 135

Park, Jung Yeon, 223

Reznitskaya, Alina, 135
Rittle-Johnson, Bethany, 20

Sachdeva, Ruchi J, 223
Saravanan, Vivek, 204
Schultz, Zachary, 191
Shih, Yong-Siang, 107
Sykes, Ariel, 135

Takahashi, Noriko, 135
Tang, Judy, 218
Tang, Yue, 125
Tao, Yuhan, 178

Wang, Elaine Lin, 49
Wang, Karen D, 98
Wang, Zhizhi, 178
Waseem, Yashal, 14
Weatherholtz, Kodi, 40

Wei, Hsin-Ro, 66
Wells, Walt, 40
Wilson, Joshua, 71
Wu, Sirui, 1
Wu, Tong, 66

Yamkovenko, Bogdan, 40
Yan, Lin, 125
Yang, Daijin, 1
Yang, Yiyao, 34
Young, Mackenzie, 191

Zhang, Dake, 178