End-to-End Automated Item Generation and Scoring for Adaptive English Writing Assessment with Large Language Models

Kamel Nebhi, Amrita Panesar, Hans Bantilan
Education First
Selnaustrasse 30
8001 Zürich, Switzerland
{kamel.nebhi, 09panesara, hansbantilan}@gmail.com

Abstract

Automated item generation (AIG) is a key enabler for scaling language proficiency assessments. We present an end-to-end methodology for automated generation, annotation, and integration of adaptive writing items for the EF Standard English Test (EFSET), leveraging recent advances in large language models (LLMs). Our pipeline uses few-shot prompting with state-of-the-art LLMs to generate diverse, proficiency-aligned prompts, rigorously validated by expert reviewers. For robust scoring, we construct a synthetic response dataset via majority-vote LLM annotation and fine-tune a LLaMA 3.1 (8B) model. For each writing item, a range of proficiency-aligned synthetic responses, designed to emulate authentic student work, are produced for model training and evaluation. These results demonstrate substantial gains in scalability and validity, offering a replicable framework for next-generation adaptive language testing.

1 Introduction

The demand for scalable, authentic, and adaptive English proficiency assessments has grown rapidly in recent years, as language learning expands across global and digital platforms. This surge has compelled test developers to explore advanced Natural Language Processing (NLP) and Machine Learning (ML) solutions that can deliver reliable and fair measurement at scale. The EF Standard English Test (EFSET)[1] exemplifies recent innovation in this space, having introduced performance-based Writing and Speaking tasks that leverage state-of-the-art NLP and ML methods for both test delivery and automated scoring (Nebhi and Szaszák, 2023; Williams et al., 2022).

Despite these advances, item generation remains a major challenge for adaptive assessment. Creating high-quality prompts that are valid across a range of topics, calibrated for all proficiency levels, and secure from test exposure is both resource-intensive and psychometrically complex (Zhang et al., 2022; Brown, 2023; Gierl and Haladyna, 2012). As the development and deployment of adaptive language tests like EFSET increases, scalable and robust methods for generating, validating, and securing writing assessment items are crucial for the advancement of fair and accurate proficiency measurement.

To address this issue, we present a novel pipeline for generating and incorporating new items into the EFSET writing assessment scoring process. Our method uses Large Language Models in Automatic Item Generation (AIG) and Synthetic Data Generation for Student Responses for scalable adaptive writing assessment. First, we generate new assessment items using a few-shot learning strategy applied to LLMs, systematically exploring multiple prompting combinations. Human evaluators then verify item quality, ensuring appropriate difficulty, clarity, and topic relevance.

In order to then integrate these validated, newly generated items into our existing automated assessment pipeline, we fine-tune a LlaMa-3.1 8B model via ORPO (Optimized Reward Preference Optimization) (Hong et al., 2024) to generate realistic student-like responses for these items across different proficiency levels. The fine-tuning relies on real test data combined with systematically generated synthetic annotations obtained via consensus annotation (majority vote) from three distinct LLM annotators. These item-response pairs then allow use to train our existing RoBERTa-based Transformer model for proficiency scoring on these new writing prompts. This synthetic annotation approach ensures scalable yet reliable response-label assignment without intensive human labor.

A summary of the main contributions of this paper is as follows: (1) we introduce an automated item generation (AIG) pipeline for adaptive writ-

[1] https://www.efset.org/en/

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 968–977
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

ing assessment that leverages state-of-the-art large language models and few-shot prompting to create high-quality, proficiency-aligned prompts; (2) we propose and validate a synthetic data augmentation process based on fine-tuned LLMs and consensus annotation via majority voting, resulting in robust and reliable datasets for model training; and (3) we develop and empirically evaluate a fully automated scoring framework based on Transformer models (RoBERTa), demonstrating significant gains in accuracy and consistency through extensive testing on EFSET items and a carefully calibrated validation set.

In the following sections, we first review the state of the art in automated writing assessment and item generation. We then detail our methodology for prompt generation, dataset construction, and automated evaluation. Next, we present empirical results illustrating the validity and reliability of our approach on both the EFSET validation set and a dedicated calibration dataset. Finally, we discuss the implications and potential extensions of this framework for scalable, adaptive language proficiency assessment.

2 Related Work

This section synthesizes two key developments in recent research on automated language assessment. First, we review state-of-the-art approaches to item generation that leverage large language models (LLMs), prompt engineering, and few-shot learning to efficiently produce diverse and high-quality assessment prompts. Second, we examine emerging methods for synthetic data annotation, with a particular focus on the use of LLMs to simulate candidate responses and facilitate reliable labeling at scale for proficiency scoring tasks.

2.1 Automated Item Generation with LLMs and Prompting

The automated generation of test items, especially for language assessment, has evolved considerably in recent years. Early systems used template-based approaches, in which test developers designed fixed "item shells" and populated them with variable linguistic elements—such as word lists or grammatical forms—to produce items at scale (Bejar et al., 2003). For example, thousands of cloze items, exercises where words are removed from a passage for the student to fill in the gaps, could be created programmatically by instantiating such templates with preselected vocabularies and distractors, providing structural consistency and psychometric control. However, content diversity and authenticity remained limited by the template bank, and extensive manual authoring was needed to cover new topics or scales. These constraints have since led to the exploration of more flexible, data-driven methods, most notably involving large language models (LLMs).

The advent of large pre-trained language models (LLMs) has fundamentally shifted automated item generation toward more data-driven, scalable, and flexible paradigms. Models such as GPT-3 and GPT-4 have been shown to generate diverse assessment items—including reading, writing, and cloze tasks—by leveraging few-shot prompting, where only a handful of examples guide the model's output (Brown et al., 2020). Educational evaluation shows that LLM-generated items are closely aligned to human-authored items, with Zhang et al. (2022) reporting that over 80% of reading comprehension questions automatically generated by GPT-3 were rated as valid by expert reviewers, and prompt appropriateness and difficulty levels closely aligned to human-authored items. Similarly, Kurdi (2023) found that LLMs could create contextually relevant language assessment prompts, achieving human-likeness scores above 4/5 on standard rubrics. Research by Brown (2023) and Zhai et al. (2023) supports that such approaches not only accelerate item production and reduce costs, but also enable rapid adaptation to new topics and test formats, with acceptance rates for LLM-generated prompts ranging from 60–95% after light expert editing.

However, even high-performing LLM-generated items require careful evaluation and annotation before they can be reliably used in machine learning-based assessment pipelines to ensure that they are well-calibrated and capable of distinguishing student ability. Recent work has shown that using synthetic annotation, consensus labeling strategies (e.g., majority voting among multiple LLMs), or semi-automatic calibration processes significantly improves dataset consistency and psychometric validity (Liu, 2023; Mai, 2022; Clark, 2021; Yao et al., 2024).

2.2 Synthetic Data Annotation for Language Assessment

A persistent challenge in automated educational assessment is the limited availability of high-quality annotated data required to train and validate modern NLP models for predicting student proficiency. Recent advances have addressed this by not only generating novel assessment prompts, but also leveraging LLMs to simulate candidate responses and assign linguistic proficiency or accuracy labels at scale (Yao et al., 2024; Wang et al., 2024; Brown, 2023).

For instance, Clark (2021) showed that the scarcity of labeled data for language tasks can be mitigated by generating synthetic examples, improving model robustness and generalization. Moreover, ensemble annotation methods—where multiple LLMs independently label each sample and a majority vote is used—have demonstrated increased labeling reliability, especially when compared to standard human annotation benchmarks (Liu, 2023). These synthetic annotation strategies make it possible to rapidly construct large, diverse, and reliable datasets matched to newly generated items.

Integrating both automated item generation and synthetic annotation creates a complete pipeline for adaptively expanding and enhancing machine learning-based scoring systems. Not only does this combination facilitate the inclusion of new item types without costly manual labeling, but it also supports the continual improvement of model accuracy, as shown by transformer-based scoring systems trained on such enriched datasets (Mayfield and Black, 2020; Mai, 2022). This integrated approach forms the basis for recent innovations in fully automated and scalable assessment frameworks.

3 Proposed Approach

This section presents our integrated pipeline for adaptive writing assessment, encompassing automated item generation, synthetic training data creation, LLM-based response simulation, and transformer-based scoring. Our approach is designed to efficiently generate, validate, and psychometrically calibrate novel test items, ensuring both robustness and scalability for deployment in real-world language proficiency testing environments.

3.1 Overview of the Integrated Pipeline Approach

Figure 3.1 presents the end-to-end pipeline developed for adaptive writing assessment. The process begins with the generation of new writing prompts, using prompt engineering and few-shot learning with a single LLM (GPT-4o) to produce candidate items focused on specific topics and proficiency levels. All generated prompts undergo human expert review, where only validated items are retained for integration into the assessment bank.

We fine-tune a LLaMA 3.1 (8B) model—leveraging ORPO optimization—on a custom instruction dataset to generate synthetic student responses for each validated item reflecting varying proficiency levels. This instruction dataset is created by collecting real candidate answers, prompting several LLMs to annotate each response for accuracy using the few-shot paradigm, and applying a majority voting scheme to select the final label. Only samples with strong inter-model agreement are retained, ensuring high label reliability and calibration.

The resulting synthetic dataset, containing approximately 200 responses per new item, is then used to further train and fine-tune a RoBERTa-based transformer scoring model. This updated scoring engine is evaluated both on existing and new items to ensure seamless integration and consistency. Throughout the pipeline, quality is maintained through a combination of automated filtering and targeted human-in-the-loop validation, enabling scalable, reliable item generation and robust scoring for real-world proficiency assessment.

3.2 Phase 1: Automated Creation and Validation of Writing Items

The aim of this first phase is to automate the generation of writing prompts intended for students to write an essay about, each of which is targeted at a specific proficiency level and topic. We leveraged a large language model (LLM)–specifically, GPT-4o (OpenAI, 2024)– to achieve this. To ensure that the generated writing prompts were tailored to adaptive assessments, we provided the LLM with a set of representative triplets of writing prompt, proficiency level, and topic. We then guided the LLM to generate new writing items with the intended form, content scope, and level-appropriateness by few-shot prompting the LLM with examples of the target structure and the level of complexity required.

In order to guide the LLM, we first carefully curated a small set of $\approx$ 20 examples explicitly designed to match the communicative demands of English writing proficiency tests. Each writing example consisted of a proficiency level and a suc-

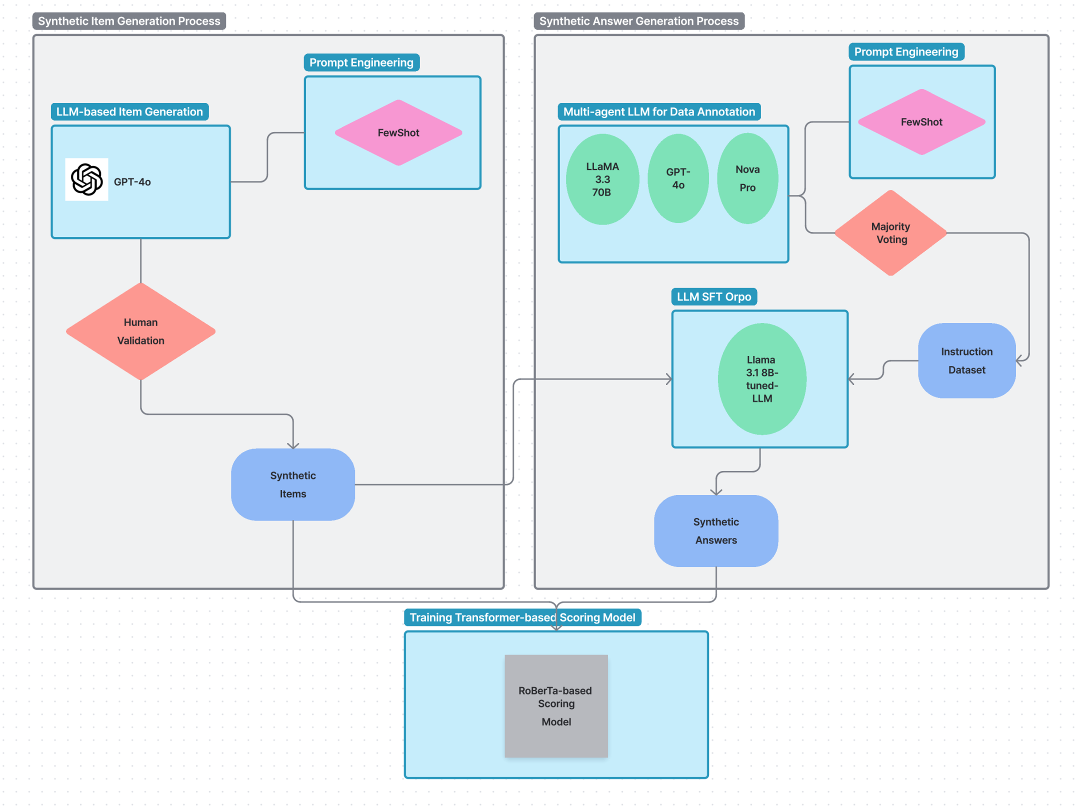

Figure 1: Pipeline for automated item and synthetic answer generation in the adaptive writing assessment

cinct writing prompt (no more than 25 words). To augment these examples with the topic as further context in the few-shot prompt, GPT-4o was applied to generate the topic for these hand-picked examples. The resulting triplets of example writing prompt, proficiency level and topic constituted our set of few-shot examples.

To generate a new writing item, we specified the desired topic(s) and proficiency level as part of the input for the LLM. A few semantically similar but diverse set of examples were chosen from our manually curated collection using LangChain's MaxMarginalRelevanceExampleSelector (LangChain, 2025). These chosen examples were then passed into the LLM as a part of the input prompt to generate a new item based on the requested topic and proficiency level, as shown in Figure 2.

We ensured a wide coverage of content and linguistic complexity across all proficiency bands by systematically generating new items across multiple combinations of topic and proficiency level.

Five language assessment experts were asked to judge a sample of 100 prompts based on appropriateness for assessment regarding the following metrics: clarity, curriculum fit, and difficulty level. In addition, to further evaluate item quality, we compared expert annotations on both difficulty level and topic with GPT-4o's predictions, finding a correlation of nearly 0.9. This high level of agreement suggests that GPT-4o is able to closely approximate expert judgment in these qualitative aspects.

The use of LLMs and enabled rapid and scalable item generation, whilst retaining strict quality control through expert review, allowing the assessment to expand to new topics and levels efficiently and reliably, as recommended in recent work on few-shot prompting in language assessment contexts (insert citation).

3.3 Phase 2: Synthetic Generation of Training Data

In this phase, the aim was to generate a high-quality training set to generate responses for the new items produced in Phase 1. The use of ORPO in the next stage requires pairs of good and bad student responses for each item, and hence we require a way to assess the quality of generated responses to produce these pairs of examples. To do so, we first evaluated several available LLMs of different architectures and sizes for its ability to rate student responses. Each model was assessed for its consistency and reliability in assigning grammatical

```
Prompt:
Imagine you are a language teacher writing essay question prompts for an English
level written test. Given the level and topic, write a short prompt (max 25 words)
for the student. The prompt should be succinct and appropriate.

Examples:
Topic: Daily life
Level: 4
Prompt: Describe your daily routine.

Topic: Work, Company policies
Level: 6
Prompt: Your boss has asked for your help with the office dress code policy.
What rules do you suggest?
...
```

Figure 2: Illustration of a few-shot prompting template used for automated writing prompt generation.

accuracy scores to (item, response) pairs using a calibration set drawn from real test data. We measured agreement between each model and expert human annotations using Cohen's Kappa statistic. Based on these preliminary experiments, we then selected the three LLMs that demonstrated the highest inter-annotator agreement with human raters as well as amongst themselves to perform a majority vote over the quality of the synthetic student response. This enabled us to then produce a larger dataset of pairs of student responses that can be used in the next phase of response generation.

For the annotation process, each selected LLM was first provided with a few-shot prompt comprising the grammatical accuracy scale (0–4) and multiple labeled examples. Each model independently assigned an accuracy score to every response, leveraging the internalized patterns from the few-shot instruction. Majority voting was then applied to the three scores produced for each sample, retaining the class most frequently assigned as the final label.

To ensure the highest possible data quality, we filtered the resulting dataset to retain only the samples where annotator agreement was strongest—either full consensus or clear majority among the three LLMs. This approach allowed us to construct a robust, reliable, and well-calibrated instruction dataset for producing realistic student responses via subsequent model fine-tuning and evaluation.

To evaluate the quality and reliability of the annotation process, we created an evaluation set (gold standard) consisting of approximately 200 (item, response) pairs. Each of these samples was independently annotated for grammatical accuracy, on a scale from 0 to 4, by five expert human raters. Only those samples with an inter-annotator agreement above 70% were retained, ensuring a high level of reliability in the ground truth annotations.

This gold standard dataset was then used to benchmark each candidate LLM's annotation performance. For the comparison, we calculated the Cohen's Kappa score between the accuracy levels assigned by each LLM and the ground truth established by human annotations. The LLMs evaluated in this process included Llama 3.3 70B (Touvron et al., 2024), Nova Pro and Nova Small (AWS proprietary models[2]), Mistral Small and Mistral Large (Jiang et al., 2023), Claude Opus (Anthropic, 2024), and GPT-4o (OpenAI, 2024).

This systematic comparison enabled us to identify the models with the highest alignment to expert human judgments, guiding the selection of annotators for the synthetic data generation pipeline.

LLM Candidate	**CK**
LLaMA 33 70B	0.86
GPT-4o	0.87
Nova Pro	0.78
Mistral Small	0.18
Mistral Large	-0.01
Claude Opus	-0.08
Ensemble	0.89

Table 1: Cohen's Kappa agreement between each LLM and gold standard human annotation

We selected the three best-performing LLMs, LLaMA 33 70B, GPT-4o, and Nova Pro as shown in Table 1, to form an ensemble for the majority voting procedure. Notably, this combination achieved a Cohen's Kappa of 0.89 with the human-annotated gold standard, outperforming any individual model. This result demonstrates that major-

[2] Technical details available in the AWS Bedrock documentation: https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-nova.html.

ity voting among top-performing models further increases annotation reliability and brings machine annotation closer to human-level agreement.

Finally, we generated our synthetic dataset of 20,000 pairs of responses, employing our ensemble of LLMs for the rating process to ensure that each pair consisted of one higher accuracy response and one lower accuracy response with the associated grammatical accuracy and proficiency level.

3.4 Phase 3: Fine-tuning of LLM-Based Response Generator

In this phase, we focused on enhancing the quality and proficiency alignment of synthetically generated responses by fine-tuning a large language model. We selected LLaMA 3.1 (8B parameters) as the base model for fine-tuning, utilizing the Optimized Reward Preference Optimization (ORPO) technique. The training data comprised approximately 20,000 synthetic samples generated during Phase 2, each annotated for grammatical accuracy and proficiency level.

Fine-tuning was conducted using a distributed training setup on the SageMaker infrastructure, employing the following configuration:

- **Instance type:** ml.g5.12xlarge
- **Environment:** PyTorch 2.5.1, GPU, CUDA 12.4, Ubuntu 22.04
- **Batch size:** 8 per device
- **Gradient accumulation steps:** 1
- **Learning rate:** 2×10^{-4}
- **Number of epochs:** 3
- **LoRA settings:** $r = 8$, $\alpha = 16$, dropout=0.1
- **Seed:** 42 (for reproducibility)

Model training was orchestrated with distributed computing support (Torchrun) to fully leverage available GPU resources, and checkpointing mechanisms were in place to ensure reliability.

Through this fine-tuning process, the LLaMA 3.1 model was adapted to generate candidate responses at specific proficiency levels, closely mimicking real student outputs in both accuracy and variety. The resulting model serves as a robust response generator for subsequent scoring model development and evaluation within the adaptive writing assessment pipeline.

3.5 Phase 4: Training Adaptive Transformer-Based Scoring Model

In the final phase, we aimed to robustly integrate the newly generated writing items into our automated scoring pipeline. To achieve this, we focused on the domain of education, and manually composed approximately thirty new writing prompts covering a broad range of proficiency levels, as generated during Phase 1.

For each prompt, the fine-tuned LLaMA 3.1 (8B) response generator was used to synthesize approximately 200 sample answers at varying proficiency levels. This resulted in a substantial and well-stratified dataset representing a wide spectrum of student abilities.

We then fine-tuned our RoBERTa-based transformer scoring model, training it on a combination of both initial (pre-existing) and newly generated items and responses. This approach was designed to ensure a smooth integration of new items into the scoring system while maintaining performance on established items.

The model was trained using the following hyperparameters with the `TrainingArguments` setup:

- **Evaluation and save strategy:** every 200 steps
- **Batch size:** 16 per device
- **Learning rate:** 2×10^{-5}
- **Warmup ratio:** 0.1
- **Epochs:** 6
- **Weight decay:** 0.01
- **Learning rate scheduler:** linear
- **Mixed precision training:** enabled (fp16)

This fine-tuning procedure enables the scoring model to generalize to new adaptive items and proficiency levels while ensuring reliable and consistent automated assessment performance.

4 Experiment

4.1 Psychometric Analysis

4.1.1 Dataset

We constructed a dataset containing 500 responses to approximately 50 different writing prompts, with

Level	Prompt	Synthetic Response	Label (proficiency)
2	What activities do you do at school?	I am studying english at university. I love english and talk with my friend, we share knowledge.	1
3	Describe your teaching style.	I like to teach students about english and use example and video for help them learn easy.	1
3	What do you think makes a good teacher?	A good teacher possesses patience, empathy, and effective communication skills. They foster a supportive environment, encourage critical thinking, and adapt teaching methods to cater to diverse learning styles, promoting academic growth and personal development in their students.	4
5	Explain the importance of extracurricular activities in a student's overall development.	Ggghhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh	0
5	Describe a typical school day, including your classes, activities, and any special events you participate in.	My school day starts at 8am with class. I join clubs after and do volunteer work. Math is my favorite.	3

Table 2: Examples of synthetically generated items and responses on the topic of education, illustrating different levels of language proficiency.

an equal split between human-generated and automatically generated answers (250 each). Each response is scored on an ordinal scale from 0 to 4, providing a rich basis for psychometric analysis. This dual-sourced dataset enables us to directly compare human performance and large language model (LLM) behavior under similar assessment conditions.

4.1.2 Evaluation

To better understand the scoring dynamics and the comparability between synthetic and authentic responses, we conducted an Item Response Theory (IRT) analysis. Figure 4.1.2 presents average Item Characteristic Curves (ICCs) derived from the dataset, shown separately for human and LLM-generated responses. Each curve reflects the probability of achieving at least a given score threshold as a function of modeled proficiency, averaged across all items.

The ICCs for both human and synthetic responses reveal similar shapes and threshold spacing, indicating that LLM-generated answers closely emulate the probability distributions observed in real student performance. This suggests that synthetic responses can serve as effective proxies for actual learner data in calibrating and evaluating automated scoring models.

4.2 Pre-piloting Study

To validate the integration and quality of the newly generated items, a pre-piloting study was conducted with approximately 250 participants in Rwanda. The main objective of this phase was to compare the performance and acceptability of the newly generated items. Participants completed a test composed of a balanced mix of traditional (previously validated) items and automatically generated new items. The distribution of items was designed to ensure diversity in both content and difficulty levels.

4.2.1 Comparative Analysis

To assess the effectiveness of the automatically generated items, we conducted a comparative analysis of success rates between old and newly generated items using statistical significance testing. For each item, we computed the difference in success rates and tested for significance using a z-test for proportions. To ensure a sufficient number of items per analysis group, we grouped the original 16 difficulty levels into 6 broader categories, thereby increasing the number of items per test group for more robust statistical analysis.

Table 3: Statistical Comparison: z-test and p-value for Each Item Level

Item-Level Group	z-score	p-value
1	4.35	0.001*
2	1.33	0.182
3	0.99	0.323
4	1.08	0.285
5	-1.96	0.056
6	-1.86	0.060

* Statistically significant at $p < 0.05$.

The results presented in Table 3 show that although most items did not show significant differences, the only statistically significant effects ($p < 0.05$) were observed between items in dif-

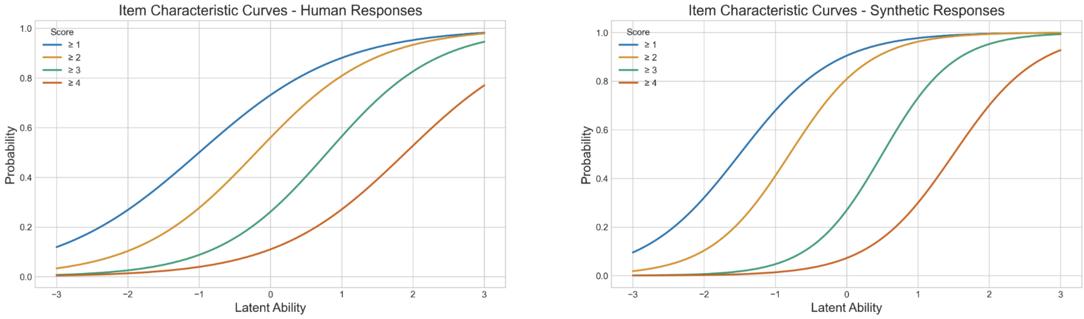

Figure 3: Average Item Characteristic Curves from IRT analysis of human and synthetic responses.

ficulty group 1 (the easiest items). For items belonging to difficulty group 5, p-values were close to the significance threshold, indicating borderline significance. These observations underscore the importance of careful quality control, particularly at both extremes of item difficulty, when integrating automatically generated items into assessments.

4.2.2 Item Characteristic Curve Analysis

To further assess the psychometric properties of both traditional and automatically generated items, we performed an Item Characteristic Curve (ICC) analysis using the Two-Parameter Logistic (2PL) model from Item Response Theory (IRT) (Baker and Kim, 2004). The 2PL model estimates two main parameters for each item: the difficulty parameter (indicating the level of ability required for a 50% probability of a correct response) and the discrimination parameter (reflecting how well the item differentiates between participants of differing ability levels).

For each item, we fitted the 2PL model using the participants' response data. T

Figure 4 displays the ICCs for three representative synthetic items extracted from the assessment. Each curve presents the probability of a correct response ($P(\theta)$) as a function of participant ability (θ), and the items were chosen to illustrate a range of difficulty and discrimination parameters.

- **Easy item:** This item is answered correctly by participants even at lower ability levels. The steep, less rounded shape of the ICC indicates a high discrimination parameter, meaning the item sharply differentiates between participants just below and just above its difficulty threshold.

- **Medium item:** This item requires a higher ability level for a 50% probability of a correct answer, suggesting moderate difficulty. It also exhibits high discrimination, as seen in the sharp transition.

- **Difficult item:** This item is considerably harder and is only likely to be answered correctly by participants with the highest abilities. The more gradual slope of its ICC suggests a lower discrimination parameter compared to the other items.

These three examples demonstrate both the range of difficulty present in the test and the variation in item discrimination. Such diversity ensures that the assessment can reliably differentiate participants across a broad spectrum of ability levels.

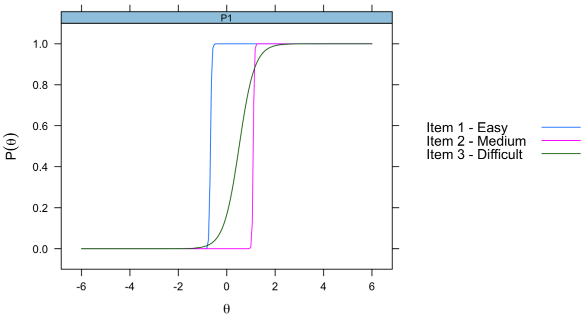

Figure 4: Item Characteristic Curves for 3 synthetic items.

4.3 Scoring Model Evaluation

4.3.1 Dataset

To assess the impact of synthetic items on model training, we conducted controlled experiments using both authentic and synthetic data. The evaluation was performed on a fixed test set of 800 samples that included a balanced selection of prompts and responses, covering all proficiency levels and a wide range of topics.

4.3.2 Evaluation

We evaluated the proficiency classification task using three distinct transformer-based models, each fine-tuned on our training data. First, we included two widely used traditional encoders: **BERT-base-uncased** (Devlin et al., 2018) and **RoBERTa-base** (Liu et al., 2019). Both are pre-trained bidirectional transformers and have served as robust baselines for a range of NLP classification tasks. Second, we fine-tuned **Flan-T5 Base** (Longpre et al., 2023), an instruction-tuned sequence-to-sequence model with strong generalization abilities for text-to-text tasks, adapting it specifically for multi-class classification by framing the label prediction as sequence generation.

Table 4 summarizes the precision, recall, and F1-scores obtained by each model, macro-averaged across proficiency levels. RoBERTa shows the strongest overall performance (macro F1-score of 0.82), illustrating the benefits of its more advanced pre-training. BERT achieves good results but slightly lower than RoBERTa, consistent with prior findings in classification tasks. Notably, Flan-T5 Base also provides competitive performance (macro F1-score of 0.80), demonstrating the viability of adapting generative models to classification through prompt engineering and sequence-based fine-tuning.

Model	**Precision**	**Recall**	**F1-score**
BERT-base-uncased	0.80	0.77	0.78
Flan-T5 Base	0.81	0.80	0.80
RoBERTa-base	0.83	0.81	0.82

Table 4: Macro-averaged precision, recall, and F1-score for each fine-tuned model on proficiency classification.

5 Conclusion

In this work, we introduced an end-to-end automated pipeline for adaptive English writing assessment, leveraging recent advancements in large language models for both item generation and synthetic data annotation. Our methodology utilizes few-shot prompting, robust majority-vote labeling, and transformer-based scoring to efficiently generate, calibrate, and evaluate new writing tasks within a psychometrically-sound framework. Extensive experiments demonstrate that the proposed system achieves high agreement with expert evaluations, ensuring both the validity and scalability required for operational proficiency testing. We anticipate that this approach will provide a solid foundation for future research on data-driven adaptive assessment and the broader application of LLMs in language testing.

Limitations

Automated scoring models risk perpetuating biases, particularly across demographic groups, language proficiencies, or socio-cultural contexts. The use of synthetic data and automated generation may also introduce or reinforce unintended patterns, potentially affecting educational fairness. To mitigate these risks, it is vital to incorporate diverse training data, implement human-in-the-loop evaluations, and regularly audit system performance. We regularly monitor test quality through ongoing psychometric analyses and expert human evaluation. This process ensures that both automated item generation and scoring maintain high standards of validity and reliability over time.

Furthermore, the introduction of a substantial number of new items into the assessment pool needs large-scale psychometric analysis to fully evaluate their functioning and impact. We acknowledge this as an essential next step, and plan to conduct comprehensive studies to further validate the psychometric properties of these newly introduced items across diverse populations and contexts.

References

Anthropic. 2024. Introducing claude 3. https://www.anthropic.com/news/claude-3-family.

Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*. CRC press.

Isaac I. Bejar and 1 others. 2003. A template-based approach to the generation of test items. In *Principles and Practice in Automated Item Generation*.

Tom B. Brown and 1 others. 2020. Language models are few-shot learners. *NeurIPS*.

Tom V. et al. Brown. 2023. The duolingo english test interactive writing task. In *Proc. BEA Workshop at ACL 2023*.

Christopher et al. Clark. 2021. Generating synthetic data for improved accuracy in educational nlp tasks. *arXiv preprint arXiv:2106.05071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mark J. Gierl and Thomas M. Haladyna. 2012. *Item Generation for Test Development*. Routledge.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Zheng-Xin Jiang and 1 others. 2023. Mistral: A simple yet effective baseline for instruction-tuned large language models. *arXiv preprint arXiv:2310.06825*.

Maryam et al. Kurdi. 2023. Controlled generation of assessment items for educational applications using language models. In *Proc. BEA Workshop at ACL 2023*.

LangChain. 2025. Few shot prompt template. https://python.langchain.com/api_reference/core/prompts/langchain_core.prompts.few_shot.FewShotPromptTemplate.html.

Ximing et al. Liu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Christopher Patil, Abhay Rao, Albert Webson, Le Hou, Pengfei Liu, and 1 others. 2023. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Zeyi et al. Mai. 2022. Towards generalizable essay scoring with language models and augmented data. In *NAACL 2022*.

Elizabeth Mayfield and Alan W. Black. 2020. Automated scoring of written essays with transformer models. In *Proc. BEA at ACL 2020*.

Kamel Nebhi and György Szaszák. 2023. Automatic assessment of spoken english proficiency based on multimodal and multitask transformers. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 769–776.

OpenAI. 2024. Gpt-4o technical report. *arXiv preprint arXiv:2405.16408*.

Hugo Touvron, Thibaut Lavril, and 1 others. 2024. Llama 3: Open foundation and instruction models. *arXiv preprint arXiv:2404.14219*.

Yizhong Wang, Zhiwei Zhang, Zexuan Zhong, Zhe Gan, Jingjing Liu, and Noah A. Smith. 2024. Simulating candidates in educational assessment with large language models. *arXiv preprint arXiv:2401.07043*.

Luke Williams and 1 others. 2022. The duolingo english test: 2022 technical report. In *arXiv preprint arXiv:2206.01056*.

Xuechen Yao, Ankur P. Parikh, Noah Constant, Dwi Susanti, and Heng Ji. 2024. Leveraging llm-respondents for item evaluation: a psychometric analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1602–1628.

Xu Zhai, Xiang Wan, Qian Jin, and Eduard Hovy. 2023. Automatic generation of language assessment tasks using large language models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 340–350.

Diyi Zhang and 1 others. 2022. Automatic generation of factual reading comprehension questions with large language models. In *Proceedings of ACL 2022*.