Effects of Generation Model on Detecting AI-generated Essays in a Writing Test

Jiyun Zu and **Michael Fauss** and **Chen Li**
Educational Testing Service, Princeton, NJ
Correspondence: jzu@ets.org

Abstract

Various detectors have been developed to detect AI-generated essays using labeled datasets of human-written and AI-generated essays, with many reporting high detection accuracy. In real-world settings, essays may be generated by models different from those used to train the detectors. This study examined the effects of generation model on detector performance. We focused on two generation models – GPT-3.5 and GPT-4 – and used writing items from a standardized English proficiency test. Eight detectors were built and evaluated. Six were trained on three training sets (human-written essays combined with either GPT-3.5-generated essays, or GPT-4-generated essays, or both) using two training approaches (feature-based machine learning and fine-tuning RoBERTa), and the remaining two were ensembled detectors. Results showed that a) fine-tuned detectors outperformed feature-based machine learning detectors on all studied metrics; b) detectors trained with essays generated from only one model were more likely to misclassify essays generated by the other model as human-written essays (false negatives), but did not misclassify more human-written essays as AI-generated (false positives); c) the ensembled fine-tuned RoBERTa detector had fewer false positives, but slightly more false negatives than detectors trained with essays generated by both models.

1 Introduction

Generative artificial intelligence (AI) tools, such as ChatGPT, Copilot, and Gemini, have become increasingly capable at generating human-like text and are now more accessible. In education, AI has great potential at enhancing teaching, learning, and assessments (U.S. Department of Education, Office of Educational Technology, 2023). At the same time, there are also concerns about the misuse of AI in writing tasks (Lund et al., 2025). Writing assignments are routinely given in both K-12 and higher education. There are also many standardized writing tests designed to measure test takers writing proficiency, such as the ACT writing test, the Graduate Record Examinations (GRE) writing test, and the Writing assessment program (WrAP) for grades 3-12 students. These tests require test takers to write essays independently. If some test takers use generative AI tools to write essays and use these essays as their own, the validity and fairness of the writing assessment are compromised.

To address concerns about AI-generated text, many detectors have been developed to identify such content. For example, Grammarly (Grammarly Inc., 2025), Scribbr (Scribbr, 2025), and GPTZero (GPTZero, 2025) provide online tools that allow users to enter text and then output an estimated percentage of the text being AI-generated, although documentation on how they were trained is generally unpublished. Several research studies reported the training and evaluation of custom-built AI-generated essay detectors. For example, Yan et al. (2023) generated essays using GPT-3 for four writing items from a large-scale assessment. Using these essays and real human test takers' essays, the authors trained two detectors: one using supervised machine learning (ML) approach and the other by fine-tuning the pre-trained language model RoBERTa (Liu et al., 2019). The detection accuracy on a holdout test set was respectively 96% and 99.75% for these two detectors. Jiang et al. (2024) studied the accuracy and potential bias in detecting ChatGPT-generated essays. Using 10,000 essays generated by ChatGPT and 10,000 essays written by real test takers for 50 GRE writing items, the authors trained detectors using supervised ML with linguistic features extracted by e-rater (Attali and Burstein, 2006) and GPT-2-based perplexity features. Detection accuracy of the best performing detector was nearly 100% on a holdout test set, and showed no evidence of bias against non-native English speakers.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 92–98
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

When AI-generated essay detectors are applied in real-world settings, several factors may affect their performances. For example, users may use a different generation model than the models used to generate the essays used for training the detectors. They may also use a different sampling temperature, different prompts, instruct the AI tool to paraphrase the generated essays to disguise its being AI-generated, or revise the generated output essays manually (themselves or other human).

Given the growing number of generative AI tools and the rapid release of newer AI models, understanding how different generation models affect the detection of AI-generated essay is an important research question. The study by Zhong et al. (2024) provides insights into this issue. The authors generated 200 essays using each of 10 different large language models (LLMs) and compared the essays in terms of linguistic features, textual similarities, and scores. They also trained a detector for each LLM using a feature-based ML approach relying on human-written essays and the 200 essays generated by that specific LLM. They found that while the detection accuracies for identifiying essays trained by the same LLM were higher than .9, when the detectors were applied to essays generated by the different LLMs, detection accuracy could be as low as .5. These findings showed the challenge of generalizing AI detectors to different generative models.

In this study, we investigate the effects of generation model on detecting AI-generated essays, expanding prior research to detectors trained on essays generated by more than one LLM as well as using a fine-tuning LLM approach. Specifically, we focus on two widely used generation models – GPT-3.5 and GPT-4 – and use writing items from a large-scale standardized English proficiency test for detector training and evaluation.

2 Method

2.1 Writing Items

We used 20 writing items from a standardized English proficiency test. The majority of the test takers are young adults. Each item asks test takers to write an essay expressing their opinion on a given topic with supporting details, with at least 100 words written within a 10-minute time limit. Essays were typed on a computer.

2.2 Data

Human-written essays We collected all test takers' responses to these 20 items when each item was administered for the first time and in test centers. The number of essays per item ranged from 192 to 6,438. For items with more than 300 essays, we randomly sampled them down to 300. The resulting total number of human-written essays used in this study was 5,745.

AI-generated essays We used GPT-3.5 turbo (version 0613) and GPT-4 (version 0613) to generate essays via the Azure OpenAI API. To generate a diverse sample of essays and match the length of human-written essays, we used 15 prompts per item – covering 5 levels of content (i.e., varying the amount of detail in the item stem or the direction of opinion to be expressed in the essay) and 3 levels for word count targets (100 words, 110 words, and 120 words). 20 essays were generated per prompt. Sampling temperature was set to 1.2 to balance text variance and text quality. Artificial typos were added to an average of 3.5% of the words using the python package *typo* (Kumar, 2022). 6,000 essays were generated using each generation model, resulting from 20 items × 15 prompts × 20 essays.

Training and test sets The 5,745 human-written essays were given a label of 0 (i.e., not AI-generated), and the 6,000 GPT-3.5-generated and 6,000 GPT-4-generated essays were given a label of 1. These essays and labels are referred to as the total dataset. Among them, we randomly selected 1,000 human-written, 1,000 GPT-3.5-generated, and 1,000 GPT-4-generated essays as the test sets for evaluating detectors' performances. From the remaining 4,745 human-written, 5,000 GPT-3.5-generated, and 5,000 GPT-4-generated essays, we created three training sets to build AI-generated essay detectors. All three training sets contain the same 4,745 human-written essays and differ by the AI-generated essays: respectively 5,000 GPT-3.5-generated, 5,000 GPT-4-generated, and a combination of randomly selected 2,500 GPT-3.5-generated and 2,500 GPT-4-generated essays. These training sets are named as Human + GPT-3.5, Human + GPT-4, and Human + GPT-3.5 + GPT-4.

2.3 Detector Training

We trained detectors for AI-generated-essay using a combination of two training approaches crossing the three training sets described in the previous section. The two training approaches are feature-based machine learning (ML) approach and fine-tuning RoBERTa. Two additional detectors were ensembled from the detectors trained on Human + GPT-3.5 and Human + GPT-4, respectively for the ML and fine-tuning approach.

Feature-based ML approach Eleven features – 10 high-level linguistic features and the logarithm of GPT-2-based perplexity of an essay - were used in the ML approach. The 10 high-level linguistic features were extracted using e-rater (Attali and Burstein, 2006). These features represent grammatical errors, usage errors, mechanics errors, organization, development, word length, word frequency, collocation and preposition, sentence variety, and discourse coherence aspects of the essays. Perplexity (i.e., exponential of the cross-entropy loss) reflects how uncertain a language model is with predicting the next token given previous tokens. A higher value indicates the text sequence is less likely to be generated by the language model. It has been found to be contributing features for detecting AI-generated essays in previous research (see e.g., Yan et al., 2023; Jiang et al., 2024). Because the essays were relatively short, we only used perplexity for the entire essays. Although essays were generated by GPT-3.5 and GPT-4, those two models were proprietary and perplexity were not available via the API. Thus, we calculated the open-source GPT-2 perplexity using the transformers library (Wolf et al., 2020).

Four type of ML classifiers – random forest, gradient boosting, support vector machine and multilayer perception - were employed. Five-fold cross validation was used on the training set for hyperparameter tuning. The best classifier with hyperparameters that led to the highest cross-validation accuracy were selected to train the final models on the entire training set. Analyses were conducted using the Scikit-learn package (Pedregosa et al., 2011).

Fine-tuning approach We fine-tuned the base version of the pretrained language model RoBERTa (Liu et al., 2019) for classification. Batch size was fixed at 16. Five-fold cross validation was used on the training set for tuning hyperparameters, including the learning rate (in the range of $5e-4$, $1e-4$, $5e-5$, $1e-5$, $5e-6$, $1e-6$) and the number of epochs (from 2 to 5). Again, hyperparameters that led to the highest cross-validation accuracy were selected to train the final models on the entire training set. All fine-tuning was conducted using the transformers (Wolf et al., 2020) and PyTorch libraries(Paszke et al., 2019).

Ensemble Using the combined Human + GPT-3.5 + GPT-4 training set is one way to help detectors learn that essays generated by either GPT-3.5 or GPT-4 are AI-generated. An alternative way is to ensemble the detectors trained separately on Human + GPT-3.5 and Human + GPT-4 training sets. We obtained two additional detectors – one for the ML approach and another for the fine-tuning approach – by ensembling the predictions from the respective GPT-3.5 and GPT-4 detectors. Note that ensembling happened at inference time, without additional model training. For each essays to be classified, we averaged the predicted probabilities from the GPT-3.5 and GPT-4 detectors. If the resulting ensemble probability was higher than 0.5, the essays was classified as AI-generated (label = 1).

2.4 Detector Evaluation

A total of eight detectors were applied to the test sets for evaluation. We can organize these detectors into four conditions, each comprising two detectors trained using either a feature-based ML approach or a fine-tuned RoBERTa model. In the first two conditions, detectors were trained on AI-generated essays produced by only one LLM, either GPT-3.5 or GPT-4. The third condition used the combined Human + GPT-3.5 + GPT-4 training set. The fourth condition involved the ensembled detectors.

We used the number of correctly and falsely classified essays in each of the 1,000 human-written, GPT-3.5-generated, and GPT-4-generated essays as evaluation metrics. Given the goal of detecting AI-generated essays, the number of human-written essays that were misclassified as AI-generated essays are false positives, and the number of AI-generated essays misclassified as human-written essays are false negatives. We used frequencies as evaluation metrics instead of accuracy, precision and recall, which are affected by the ratio between human-written essays and AI-generated essays in the test set. This is because in our test set, the composition of human-written and AI-generated essays is 1:2, which is unlikely in real settings.

3 Results

3.1 Essay Similarities

We first examined pairwise text similarities among essays for each item, because the extent of similarities among human-written, GPT-3.5-generated, and GPT-4 generated essays can affect detector performances. Per item, within the same generation source (i.e., human-written, GPT-3.5, and GPT-4), the number of pairs was $n_k(n_k - 1)/2$, where n_k is the number of essays in source k for this item. Across sources, the number of pairs was $n_i n_j/2$, where n_i and n_j are the number of essays for sources i and j. The number of pairs for GPT-3.5-generated, and GPT-4-generated essays was 897,000; for human-written essays was 834,074, for GPT-3.5-generated and GPT-4-generated essays for 1,800,000, for GPT-3.5/4-generated and human-written essays was 1,723,500. For each pair of essays, we calculated the cosine similarity of trigram term frequency-inverse document frequency (TF-IDF) vectors, and the edit similarity (Navarro, 2001). Both similarity measures are within 0 to 1, with a higher number indicating higher similarities. Box plots of pairwise similarities of essays for the same items within and between sources are provided in Figure 1. Sources with higher median similarities are located higher on the y-axis.

Essay similarity results revealed differences between GPT-3.5- and GPT-4-generated essays. Within the same source, essays generated by GPT-3.5 were the most similar as each others, while GPT-4 was able to generate essays with higher text variability, but not as diverse as human-written essays. Across sources, essays generated by the two LLMs were more similar to each other than with human-written essays. Human-written essays were more similar with GPT-3.5-generated essays than with GPT-4-generated essays.

3.2 Detector Performances

Eight detectors were applied to the three test sets consisting of respectively 1,000 human-written, GPT-3.5-generated and GPT-4-generated essays. The number of essays correctly and wrongly classified as human-written or AI-generated on the test sets by each detector are reported in Table 1. Among all the studied ML classifiers, SVM yielded the highest cross-validation accuracy in the three training sets. Thus, detectors obtained using SVM were used to represent the ML approach. When building detectors by fine-tuning RoBERTa, the following hyperparameters led to the highest cross-validation accuracy respectively for the three training sets, $lr = 5e - 5$ and $epoch = 4$, $lr = 5e - 5$ and $epoch = 5$, and $lr = 1e - 5$ and $epoch = 4$.

First focus on detectors trained with AI essays generated by only one LLM (i.e., conditions 1 and 2 in Table 1). In the columns for human-written essays, we see that fine-tuned RoBERTa misclassified fewer number of human-written essays as AI-generated essays than SVM. While SVM misclassified 22 and 32 human-written essays as AI-generated essays, detectors based on fine-tuned RoBERTa misclassified fewer than 5 essays. For AI-generated essays that were generated by the same LLM as used in the training set (i.e., column GPT-3.5-generated for condition 1, and column GPT-4-generated for condition 2), detectors based on fine-tuned RoBERTa correctly classified all AI generated essays, while SVM missed 24 and 15 AI-generated essays. However, when detectors trained with AI-generated essays by one LLM were applied to essays generated by the other LLM (i.e., column GPT-4-generated for condition 1, and column GPT-3.5-generated for condition 2), the number of false negative cases increased. Fine-tuned RoBERTa trained with GPT-3-generated essays failed to identify 84 GPT-4-generated essays and fine-tuned RoBERTa trained with GPT-4 generated essays GPT-4 missed 192 GPT-3.5-generated essays. Performances of SVM detectors were worse. They failed to identify respectively 522 and 370 essays generated by the other LLM.

When both GPT-3.5- and GPT-4-generated essays were included in the training set (i.e., condition 3 in Table 1), the resulting detectors had lower number of false negatives cases for the combination of 1,000 GPT-3-generated and 1,000 GPT-4-generated essays. Fine-tuned RoBERTa identified all GPT-3.5 generated essays and only missed one GPT-4 generated essay, while SVM missed 29 GPT-3.5-generated and 26 GPT-4 generated. In terms of false positives, fine-tuned RoBERTa misclassified 9 out of the 1000 human-written essays (0.9%) as AI-generated essays, while SVM misclassified 41 (4%). These number of false positives were slightly higher than those for detectors trained with AI essays generated using only one model (i.e., conditions 1 and 2).

When GPT-3.5 detector and GPT-4 detector were ensembled (condition 4 in Table 1), the ensembled fine-tuned RoBERTa detector only misclassified 1 human-written essays as AI-generated essays and

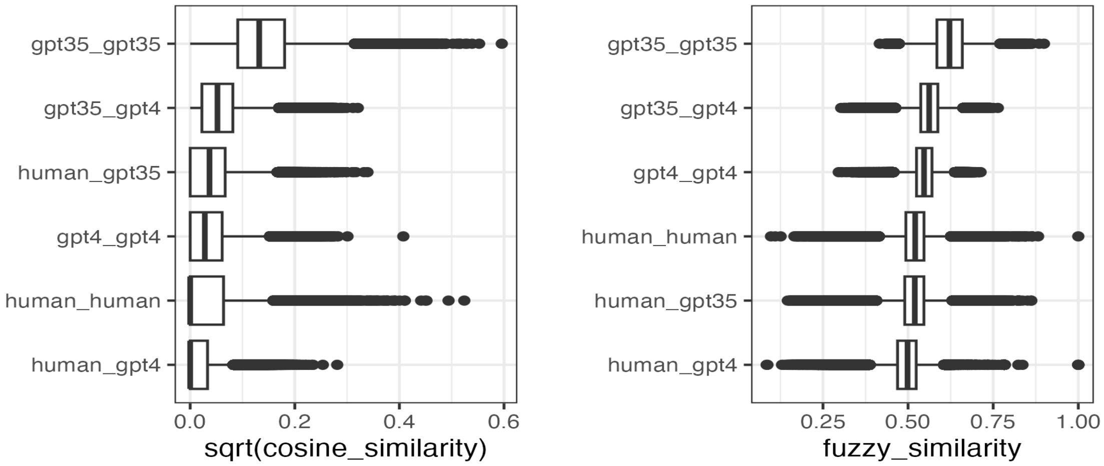

Figure 1: Box plots of the square root of cosine similarity and edit similarity among essays for the same items within and between sources.

Table 1: Number of Correctly and Falsely Labeled Essays in Test Sets by Different Detectors

Training condition	Approach	Human-written (n=1000)		GPT-3.5-generated (n=1000)		GPT-4-generated (n=1000)	
		Correct	Wrong	Correct	Wrong	Correct	Wrong
1. Human + GPT-3.5	SVM	978	22	976	24	478	522
	RoBERTa	996	4	1000	0	916	84
2. Human + GPT-4	SVM	968	32	630	370	985	15
	RoBERTa	997	3	808	192	1000	0
3. Human + GPT-3.5 + GPT-4	SVM	959	41	971	29	974	26
	RoBERTa	991	9	1000	0	999	1
4. Ensemble	SVM	986	14	894	106	928	72
	RoBERTa	999	1	993	7	993	7

failed to identify 7 GPT-3.5-generated and also 7 GPT-4 generated essays. Ensembled SVM misclassfied 14 human-written essays as AI-generated essays, but missed 106 GPT-3.5-generated and 72 GPT-4 generated. Comparing the conditions 3 and 4, in which detectors were given the information that both GPT-3.5- and GPT-4-generated essays are AI-generated, the ensembled detectors had lower number of false positives and higher number false negatives.

4 Discussion

In this study, we investigated the effects of generation model on performances of detectors for AI-generated essays. We studied two generation models (GPT-3.5 and GPT-4), two training approaches (feature-based ML and fine-tuning), and two ways of providing information from both generation models (including essays generated by both LLMs in the training set and ensembling detectors trained with only one LLM for essay generation). We found that a) fine-tuned detectors outperformed feature-based ML detectors on all studied metrics; b) compared to detectors trained with essays generated from both models, those trained with essays generated from only one model did not misclassify more human-written essays as AI-generated (false positives), but did misclassify more essays generated by the other model as human-written essays (false negatives); c) the ensembled fine-tuned RoBERTa detector had fewer false positives, but slightly more false negatives comparing to detectors trained with essays generated by both GPT-3.5 and GPT-4.

Fine-tuning pre-trained large language models has been found to be effective for many classification tasks, including natural language inference (Devlin et al., 2019), automated essay scoring (Fernandez et al., 2023), and AI-generated essay detection (Kaggle Community, 2025). Our findings are inline with these previous findings, suggesting superior performances of the fine-tuning approach comparing to the feature-based ML approach for AI-generated essay detection. However, the complexity of LLMs makes it difficult to explain the predicted results from fine-tuned LLMs. This posts challenges of using fine-tuned detectors in high-stakes situations, where false accusations against individuals can have serious consequences. Research to identify tokens or phrases that affecting the fine-tuned detectors' decisions, or the effects of adversarial inputs can be important future directions.

To detect essays generated by a wide range of AI models, the natural choice is to train a detector using essays generated by a diverse number of AI models. However, it can be resource-intensive to re-train the detector each time a new AI model is released. If the number of human-written essays don't increase, creating a balanced training set may mean not include all the AI-generated essays from previous AI models for training. This is the scenario we studied. Even though we generated 5,000 GPT-3.5-generated essays in conditions 1, and 5,000 GPT-4-generated essays in condition 2, we only used 2,500 from each generation model in condition 3. We found ensembling fine-tuned RoBERTa can be an effective alternative. It allows the use of the same number of AI-generated essays for each generation model as the number of human-written essays. Once detector is built for each generation model, one can flexibly adjust the contribution from each detector at inference, if there is evidence on the likelihood of essays from each generation model. Ensemble also allows easy adjustment of threshold. For example, if reducing false positives is more important, one may adjust the threshold to higher than .5.

5 Limitations

In this study, we generated essays using two AI models, built detectors with balanced sets of human-written and AI-generated essays, and studied detector performance in terms of detection accuracy. Results need to be generalized with caution beyond these conditions. As noted in the introduction, in the real-world, AI can be used in creating essays in many different ways. Other models that are more distant from the models in the OpenAI family, such as LLaMA or DeepSeek-R1, may produce more different essays, thus affect the detection performance. Essays may also be created by both humans and AI, with only a portion of the text generated by AI or humans revise AI-drafted essays. Moreover, fairness in detection across demographic groups is also an importance metric for evaluating detector performance. For future work, we plan to expand the study by including a broader range of generation models and also varying the proportion of AI generated text within essays.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment*, 4(3):3–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. 2023. Automated scoring for reading comprehension via in-context bert tuning. *Preprint*, arXiv:2205.09864.

GPTZero. 2025. Gptzero - ai detector - the original ai checker for chatgpt & more. Accessed: 2025-06-18.

Grammarly Inc. 2025. Grammarly ai detector. Accessed: 2025-06-18.

Yang Jiang, Jiangang Hao, Michael Fauss, and Chen Li. 2024. Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, 217:105070.

Kaggle Community. 2025. LLM - Detect AI Generated Text: Discussion Post. `https://www.kaggle.com/competitions/llm-detect-ai-generated-text/discussion/473295`. Accessed: 2025-06-20.

Ranvijay Kumar. 2022. A python package to simulate typographical errors in english language. `https://github.com/ranvijaykumar/typo`. Accessed: 2025-06-20.

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. *Preprint*, arXiv:1907.11692.

Brady D. Lund, Tae Hee Lee, Nishith Reddy Mannuru, and Nikhila Arutla. 2025. Ai and academic integrity: Exploring student perceptions and implications for higher education. *Journal of Academic Ethics*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Scribbr. 2025. Ai detector - trusted ai checker for chatgpt, copilot & gemini. Accessed: 2025-06-18.

U.S. Department of Education, Office of Educational Technology. 2023. Artificial intelligence and the future of teaching and learning: Insights and recommendations.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Duanli Yan, Michael Fauss, Jiangang Hao, and Wenju Cui. 2023. Detection of AI-generated essays in writing assessment. *Psychological Testing and Assessment Modeling*, 65(2):125–144.

Yang Zhong, Jiangang Hao, Michael Fauss, Chen Li, and Yuan Wang. 2024. Evaluating ai-generated essays with gre analytical writing assessment. *Preprint*, arXiv:2410.17439.