

Automatic Prompt Engineering for Generative AI–Based Essay Scoring

Yue Huang

Measurement Incorporated
yueh@udel.edu

Abstract

This study examined *automatic prompt engineering* (APE) for holistic scoring of one independent argumentative writing item from the PERSUADE 2.0 dataset. The APE approach optimized scoring instructions and examples for a GPT-4–family model under different configurations. The finalized configuration, selected through validation accuracy, was compared with a research-informed, zero-shot Baseline prompting approach on the held-out test set. Quadratic-weighted kappa (QWK) increased from .646 to .812, alongside improved exact agreement, adjacent agreement, and mean-score alignment. Gains were concentrated at higher score levels. Gender gaps narrowed descriptively, while English language learner subgroup patterns varied across metrics. These findings support prompt optimization for *automated essay scoring* (AES) while highlighting the importance of rubric review and subgroup evaluation.

1 Introduction

Automated essay scoring (AES) uses computational methods to assign essay scores that approximate those awarded by human raters. Its development has progressed from models based on engineered linguistic features to neural approaches that learn representations of writing, and more recently to generative language models (Huang et al., 2025). Rapid advances in generative artificial intelligence (AI) have renewed interest in using these models for AES. Their ability to interpret natural-language instructions offers a way to specify scoring tasks without training a separate scoring model for each application.

Recent research suggests that generative AI can approximate human essay ratings through prompt engineering, the deliberate design of model inputs to guide the content and format of its scores (Huang and Wilson, 2025; Huang et al., 2025, 2026). In

AES, these inputs may specify a rater role, the writing task, scoring criteria, and examples of human-scored essays. Much of the prompt-based research reviewed by Huang et al. (2025) relied on researcher-designed prompts. Their synthesis identified useful design strategies, including clear instructions and rubric guidance, while also showing that the benefits of individual components varied across studies. Such guidance supports prompt construction, but leaves researchers to determine how components should be worded and combined for a particular scoring task.

Automatic prompt engineering (APE), a form of prompt optimization, uses algorithms to generate and evaluate alternative instructions and examples. Huang et al. (2025) recommended investigating DSPy, a framework for specifying language-model tasks and optimizing their prompts, for AES. Here, prompt optimization refers to searching for a prompt that improves agreement with human ratings; it does not update the language model's parameters. Examining both prompt content and scoring performance is important for understanding what this process contributes to AES.

2 Related Work

Prior research identifies two connected questions in prompt development: which information to provide and how to select an effective formulation. The AES studies synthesized by Huang et al. (2025) addressed the first through variations in rubric guidance, context, and examples. Huang et al. (2026) further showed that contextual information and scored examples influenced GPT-4o scoring of elementary students' essays. Together, these studies provide evidence about which information should be included in scoring prompts. However, they offer less guidance on how prompt formulations can be systematically generated and refined for a new scoring task beyond comparing a small set of

manually specified alternatives.

Research on prompt optimization addresses this problem by treating prompts as candidates whose performance can guide further search. Zhou et al. (2023) generated and evaluated multiple instruction candidates, with an iterative variant further exploring alternative formulations of high-performing candidates. Later approaches made revision more explicitly feedback-driven: ProTeGi (Prompt Optimization with Textual Gradients) used natural-language analyses of model errors to identify weaknesses in the current prompt and guide its iterative refinement, whereas OPRO (Optimization by PROMPTing) used histories of candidate prompts and their scores (Pryzant et al., 2023; Yang et al., 2024). These methods differ in how they propose revisions, but share the principle that prompt development can be guided by observed performance rather than researcher judgment alone.

This principle extends beyond instruction wording to demonstrations, examples included in a prompt to illustrate the task. Wan et al. (2024) showed that exemplar selection can be as consequential as instruction optimization in the tasks they examined, motivating attention to what a model is shown as well as what it is told. Extending automatic prompt optimization to multi-stage language-model programs, Opsahl-Ong et al. (2024) brought these elements together in MIPRO (Multi-prompt Instruction PROposal Optimizer), which jointly optimizes instructions and demonstrations across the connected LLM calls. For AES, this creates an opportunity to adapt both the description of the task and the examples illustrating its application, although the effectiveness of the resulting combination still requires evaluation in the target assessment context.

Educational applications of APE provide promising but distinct precedents. Xue et al. (2025) applied APE to 11 science constructed-response items, evaluating optimized scoring prompts in zero-shot and few-shot settings and examining how prompt changes affected scoring accuracy and agreement with human scores. Harada et al. (2026) extended prompt optimization to essay-scoring rubrics, iteratively refining them based on discrepancies between model-generated and human scores and the model's scoring rationales. These studies establish that prompt optimization is relevant to educational scoring, while addressing different parts of the scoring prompt. Xue et al. (2025) retained fixed examples within their few-shot conditions,

and Harada et al. (2026) focused on rubric revision. They therefore leave room to examine joint instruction and demonstration optimization for holistic essay scoring.

The writing task also gives this extension substantive importance. Xue et al.'s (2025) short science responses were evaluated against item-specific content criteria. In the independent argumentative essays examined in this study, students develop a position through their own reasons and examples, and holistic scoring integrates the quality of development, organization, and language across an extended text. Effective responses can follow different argumentative paths rather than reproduce a single expected answer. This distinction does not imply that short-response scoring is simple; it means that findings from content-focused responses cannot establish how optimized prompts will interpret broader writing-quality criteria. Harada et al.'s (2026) essay study makes this a question about the optimization approach within AES, rather than an entirely unexamined application.

Evaluation also needs to consider whom a scoring approach serves. Huang et al. (2026) found that overall scoring performance and subgroup patterns did not necessarily move together. A brief comparison by gender and English language learner (ELL) status in this present study can therefore complement aggregate results by identifying whether agreement patterns differ across groups. Such comparisons provide preliminary evidence about the consistency of scoring performance, rather than a comprehensive test of fairness.

3 The Present Study

This study examines whether an APE approach improves the generative AI-based holistic scoring of independent argumentative essays relative to a researcher-developed zero-shot prompt informed by prior research on effective prompt-based AES strategies, hereafter termed the Baseline prompting approach. It extends prior educational prompt optimization research by jointly optimizing instructions and demonstrations (examples) and evaluating prompt content and agreement with human ratings. An exploratory analysis also examines performance across gender and ELL groups. The findings can inform scalable, rubric-aligned scoring systems by evaluating a systematic alternative to manual prompt development. Three research questions guide the study:

RQ1. How does the prompt generated by the APE approach differ in content from that used by the Baseline prompting approach?

RQ2. How do the APE approach and Baseline prompting approach compare in agreement with human ratings and score-level prediction patterns?

RQ3. How does scoring performance for the two approaches vary across gender and ELL groups?

4 Methods

4.1 Sample

This study used PERSUADE 2.0, a publicly available dataset of argumentative essays written by US students in Grades 6–12 (Crossley et al., 2024). The analysis was restricted to one independent-writing assignment asking students to take a position on an argumentative topic and support it with reasons and examples. Responses were retained when essay and assignment text were nonempty and the human holistic score was valid. The resulting 2,148 essays came from Grades 9–12, with 12, 191, 1,548, and 397 responses, respectively. Female students contributed 48.8% of essays, and ELL students contributed 27.9%. Dataset-reported essay lengths ranged from 159 to 1,584 words. Human raters scored the essays using the six-point holistic rubric for independent writing provided with the dataset, with 1 indicating the lowest performance level and 6 the highest.

Following the competition split recorded in the dataset, I allocated 1,492 essays (69.5%) to a training set for prompt development and 656 essays (30.5%) to a held-out test set for evaluating scoring performance. The training set was further divided into separate subsets for developing candidate prompts and validating their performance, as detailed in Section 4.3. The test essays were excluded from both prompt optimization and configuration selection and used only to evaluate and compare the APE approach and Baseline prompting approach.

4.2 Prompting Approaches

Both approaches scored essays through the GPT application programming interface (API). Scoring refers to applying a specified prompt to an essay and requesting one holistic score within the rubric’s six-point scale. The APE approach used the MIPROv2 optimizer within DSPy framework to jointly optimize scoring instructions and essay–score demonstrations for agreement with hu-

man scores (Opsahl-Ong et al., 2024; DSPy, 2026).

The Baseline prompting approach used a researcher-written zero-shot prompt. Drawing on Huang and Wilson (2025) and Huang et al. (2026), it included an expert-rater role, the scoring task, assignment and student context, the original six-level holistic rubric, and a requirement to return one integer from 1 to 6 to reflect the writing quality following the rubric. This instruction also served as the starting point for the APE approach. During optimization, DSPy could rewrite the instruction and select essay–score demonstrations. The comparison therefore evaluated the complete APE approach against the Baseline prompting approach, rather than isolating individual prompt components.

4.3 Prompt Optimization and Configuration

To develop the APE approach, I implemented MIPROv2 in DSPy version 3.2.1 (DSPy, 2026). MIPROv2 is an updated implementation of the MIPRO method described by Opsahl-Ong et al. (2024), retaining its core strategy of jointly optimizing instructions and demonstrations through Bayesian search. Within each configuration, training essays were divided into candidate-prompt-development and validation subsets. Using the Baseline instruction and candidate-prompt-development data, the optimizer proposed alternative instructions and demonstration sets. Each search trial evaluated a candidate prompt by scoring the full validation subset and calculating exact agreement with human ratings. Completed trials guided subsequent search, and the candidate prompt with the highest validation agreement was retained.

Validation-set size and search budget varied across optimization configurations. Larger validation subsets provided more evidence for evaluating prompts, while additional trials explored more prompt combinations; both increased the number of scoring requests. These settings explored practical development choices rather than demonstrating cost savings or establishing a universally optimal configuration.

Table 1 summarizes the configurations. Configuration 2 increased validation-set size relative to Configuration 1, whereas Configuration 3 increased the search budget. Configurations 1 and 3 shared the same 100 validation essays, included within Configuration 2’s 200-essay subset. Each configuration was optimized separately. GPT-4.1-mini proposed prompting instructions at an initial tem-

perature of 1.0. Configuration 3 had the highest validation exact agreement (.540) and was selected for the APE approach reporting. Its optimized instruction and six demonstrations were finalized before held-out evaluation.

Config.	Candidate Development (N)	Validation (N)	Search Trials (N)	Validation Exact Agreement
1	1,392	100	12	.510
2	1,292	200	12	.525
3	1,392	100	20	.540

Table 1: Configurations for the APE approach.

4.4 Essay Scoring Procedure

After finalizing the prompting approaches, I applied each to all held-out essays using GPT-4.1-mini through the API, with scoring temperature set to 0 to reduce sampling variability. Each request supplied one target essay and requested a single holistic score. The Baseline prompting approach used the zero-shot researcher-written instruction; the APE approach used the finalized optimized instruction and demonstrations.

4.5 RQ1 and RQ2: Prompt Content and Scoring Performance Evaluation

For RQ1, I compared the final prompts in instruction content, rubric presentation, and demonstrations. For RQ2, I evaluated human-machine agreement using quadratic-weighted kappa (QWK), exact agreement, adjacent agreement, the proportion within one point; and standardized mean difference (SMD). $QWK \geq 0.70$ and $|SMD| \leq 0.15$ served as reference criteria from Williamson et al. (2012).

Both approaches used the same valid human-machine score pairs. Differences were calculated as APE minus Baseline. I also estimated 95% percentile confidence intervals from 10,000 paired bootstrap resamples.

Finally, confusion matrices displayed counts and percentages within each human score level to examine where the approaches agreed with human ratings and where they under- or over-scored.

4.6 RQ3: Subgroup Comparisons

For RQ3, I calculated the four metrics separately by gender and ELL status and descriptively compared subgroup performance under the Baseline prompting approach and the APE approach. Gender and ELL contrasts were calculated as male minus female and non-ELL minus ELL, respectively, for

QWK, exact agreement, and adjacent agreement. For SMD, I reported absolute values within groups and absolute differences between groups' absolute SMD values.

5 Results

5.1 RQ1: Comparison of Prompt Content

The APE approach's final instruction retained the Baseline prompting approach's expert-rater role, task context, six-point scale, and numeric-score requirement. It condensed the instruction from over 700 to just over 100 words, excluding demonstrations and the target essay. The Baseline prompt reproduced the assignment and all six rubric descriptors. The exact original rubric was not presented in the APE prompt. Instead, the APE prompt summarized dimensions from the rubric including argument development, critical thinking, organization, language, and mechanics.

The APE approach added six examples in its demonstrations with scores of 5, 3, 5, 5, 4, and 5, including one bootstrapped and five labeled examples. These did not cover the full scale. The shorter instruction did not imply a shorter complete prompt because example essays added text. The comparison therefore combined changes in wording, rubric presentation, and demonstrations.

5.2 RQ2: Comparison of Scoring Performance

The APE approach showed higher overall agreement with human ratings than the baseline prompting approach (Table 2). QWK increased from .646, below the .70 reference threshold, to .812, above it—a difference of .166. Exact agreement increased from 39.9% to 54.6%, a gain of 14.6 percentage points; adjacent agreement increased from 90.4% to 98.8%, a gain of 8.4 points. The paired intervals for all three differences excluded zero.

The human mean was 4.209, compared with 3.651 for the Baseline prompting approach and 4.331 for the APE approach. The Baseline prompting approach under-scored on average ($SMD = -.494$), whereas the APE approach slightly over-scored (.108). Absolute SMD therefore decreased. The APE approach's absolute SMD met the .15 criteria. The signed difference (.602) reflects the shift from under-scoring to over-scoring.

Metric	Baseline	APE	Difference
QWK	0.646	0.812	+0.166 [0.137, 0.196]
Exact agreement	39.9%	54.6%	+14.6 pp [9.0, 20.4]
Adjacent agreement	90.4%	98.8%	+8.4 pp [6.1, 10.7]
SMD	-0.494	0.108	+0.602 [0.556, 0.651]

Table 2: Human-machine agreement and paired differences on the test set.

Note. $N = 656$. Baseline = the Baseline prompting approach; APE = the APE approach. Difference = APE - Baseline. Brackets contain 95% confidence intervals. Agreement differences are percentage points (pp). SMD values and their difference are signed.

5.3 RQ2: Comparison of Predictions Across Score Levels

Figure 1 shows that improvement varied across score levels. Of 86 human-rated score-6 essays, the Baseline prompting approach assigned 6 to one (1%), whereas the APE approach assigned 6 to 54 (63%). The APE approach substantially reduced upper-end under-scoring, while also assigning 6 to 63 essays rated 5 and three rated 4 by human raters. Consequently, 54 of its 120 score-6 predictions (45%) matched the human rating.

At human score 1, the Baseline prompting approach matched two of three essays, whereas the APE approach assigned 2 to all three. Thus, the APE predictions remained within one point, although exact agreement decreased. Only three essays supported this comparison, limiting conclusions at the lowest score level; the score-6 pattern had a stronger descriptive basis ($N = 86$).

For human scores 2–5, exact agreement was 55%, 53%, 42%, and 42% for the Baseline prompting approach, compared with 55%, 50%, 62%, and 47% for the APE approach. Gains were concentrated at scores 4–6. Overall, under-scored essays decreased from 347 to 110, while over-scored essays increased from 47 to 188. The APE approach reduced the Baseline’s downward scoring tendency without improving every score level.

5.4 RQ3: Comparison Across Student Groups

The APE approach showed higher agreement and lower absolute SMD in every subgroup (Table 3). Gender gaps were descriptively smaller with the APE approach than with the Baseline prompting approach across all four metrics. The APE approach achieved QWK above .80 for both genders, with

absolute SMD of .102 for male and .116 for female students. These point estimates met both reference thresholds.

Across ELL groups, the QWK gap decreased from .118 under the Baseline prompting approach to .029 under the APE approach, although neither group’s QWK reached .70 under the APE approach. Under the Baseline approach, exact agreement was higher for non-ELL students; under the APE approach, it was higher for ELL students. With the APE approach, adjacent agreement was nearly 99% for both groups. Although absolute SMD decreased for both groups under the APE approach, the difference between the groups increased from .097 under the Baseline approach to .176 under the APE approach. The APE approach met the .15 criterion for ELL students (.016), while some over-scoring remained for non-ELL students (.192). These comparisons describe observed subgroup patterns rather than statistically established differences.

6 Discussion

This study demonstrates the promise of automatic prompt engineering (APE) for improving generative AI-based essay scoring. By connecting prompt-content analysis with held-out performance and subgroup comparisons, it shows how prompt optimization can strengthen agreement with human ratings while identifying priorities for further refinement. The findings support a systematic approach to prompt development that combines research-informed design with empirical evaluation.

An important finding is that the APE approach produced a condensed instruction paired with scored demonstrations, suggesting that scoring expectations can be communicated through complementary forms of guidance: instructions identify relevant qualities, while examples illustrate their application. This connects with [Xue et al.’s \(2025\)](#) observation that APE can reformulate scoring instructions, raising questions about whether optimization changes the intended scoring criteria. In the present study, stronger human-machine agreement suggests that concise instructions combined with examples can effectively support scoring, although agreement alone does not establish that all rubric criteria were preserved. The key distinction is therefore between refining how criteria are communicated and changing what is assessed. Expert

review against the original rubric could help maintain this distinction, while comparisons of full and summarized rubrics, with and without demonstrations, could clarify the contributions of instruction changes and examples.

Another key finding is the APE approach's substantial improvement in scoring performance over an already well-structured baseline prompting approach. Its rater role, assignment context, complete rubric, and output requirements reflected prior research guidance. The gains therefore demonstrate the added value of automatic optimization beyond a carefully designed zero-shot starting point. [Huang et al. \(2026\)](#) highlighted the benefits of contextual information and scored examples for prompt-based essay scoring. Building on this evidence, the present study shows how APE can use human-scored data to refine scoring instructions and select essay–score demonstrations, translating general prompting guidance into a configuration tailored to the scoring task. This finding supports treating research-informed prompt design as a foundation for further empirical optimization and positions APE as a systematic means of improving agreement with human scores.

A particularly relevant comparison comes from [Plasencia-Calaña \(2025\)](#), who evaluated the same assignment within PERSUADE 2.0. Reported QWK ranged from .48 for zero-shot prompting and .60 for chain-of-thought with few-shot examples to .77 for the best-performing feature-based histogram gradient boosting classifier. The present APE approach's estimate of .812 is numerically higher, providing encouraging context for prompt optimization without model fine-tuning. However, differences in models and evaluation procedures preclude attributing this cross-study difference solely to APE. Within the present comparison, improved recognition of highly rated essays further demonstrates its practical value, while additional highest-score assignments identify opportunities for calibration.

Subgroup findings also highlight the importance of examining how gains are distributed. The APE approach produced similar performance across gender groups, whereas ELL patterns differed by metric. [Wilson and Huang \(2024\)](#) found that both human and automated scores were less predictive of state writing-test performance for ELL students, situating these concerns within broader writing assessment. In the present study, differences in human-score distributions between ELL and non-

ELL students may also contribute to the observed metric differences, particularly because the APE approach's performance varied across score levels.

Overall, this pilot study extends educational prompt optimization research by examining the APE approach for scoring full-length independent argumentative essays. The findings show that jointly optimizing instructions and demonstrations can strengthen research-informed scoring prompts and improve agreement with human ratings. Integrating such data-guided refinement with rubric review and subgroup evaluation provides a constructive foundation for developing scalable AES systems and strengthening the evidence supporting their use.

6.1 Limitations and Future Research

First, the study was limited to one assignment, one generative AI model, and one researcher-written baseline prompt. Future work should examine generalizability across writing tasks, models, and baseline prompts, and assess optimization stability using different seeds and validation samples.

Second, configuration comparisons used partly overlapping validation subsets of different sizes and did not establish an optimal search budget. More controlled comparisons, alongside direct measurements of token use and processing time, would clarify the balance between scoring performance and computational demands.

Third, sparse lowest-score observations and the descriptive nature of subgroup analyses limit conclusions about score-level and demographic differences. Larger samples at underrepresented score levels and confirmatory subgroup analyses would strengthen these comparisons.

Lastly, the Baseline prompting approach was research-informed but zero-shot, so gains from the APE approach cannot be attributed separately to instruction optimization and added demonstrations. Comparisons with [Plasencia-Calaña \(2025\)](#) provide encouraging context, but differing evaluation procedures limit direct comparisons. Comparing the APE approach with a research-informed few-shot approach would clarify its added value.

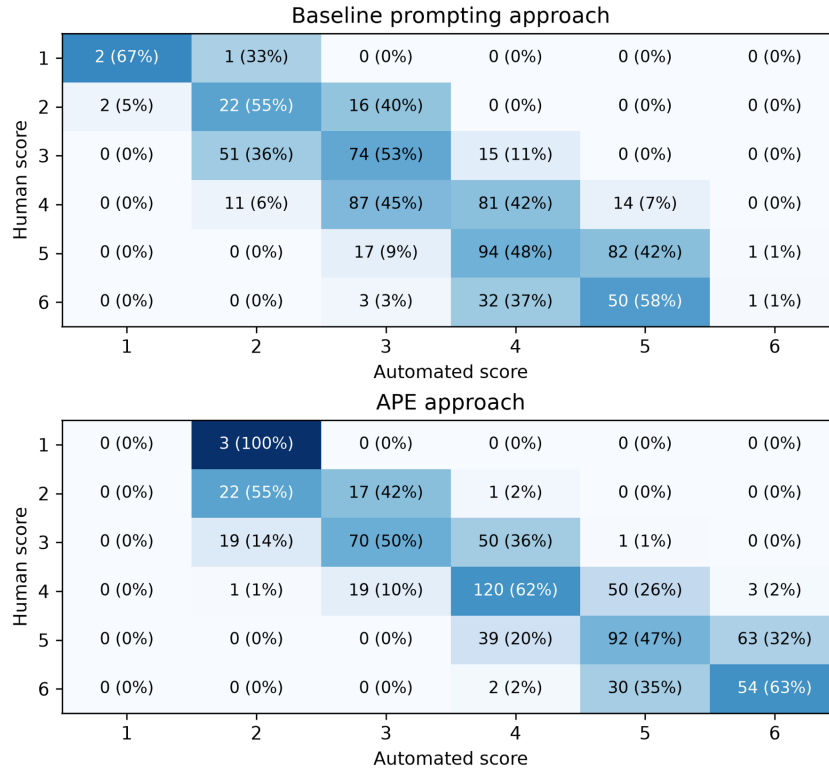


Figure 1: Confusion matrices for the Baseline prompting approach and the APE approach.

Note. Rows indicate human scores and columns automated scores. Cells show essay counts and percentages within each human-score row. Human-score counts for levels 1–6 were 3, 40, 140, 193, 194, and 86, respectively.

Group/Contrast	<i>N</i>	Approach	QWK	Exact Agreement	Adjacent Agreement	Absolute SMD*
Male	358	Baseline	.659	41.6%	91.9%	.458
Female	298	Baseline	.630	37.9%	88.6%	.538
Male – Female	—	Baseline Δ	.029	3.7	3.3	.080
Male	358	APE	.815	55.6%	98.9%	.102
Female	298	APE	.806	53.4%	98.7%	.116
Male – Female	—	APE Δ	.009	2.2	0.2	.014
Non-ELL	429	Baseline	.475	41.7%	89.5%	.590
ELL	227	Baseline	.357	36.6%	92.1%	.687
Non-ELL – ELL	—	Baseline Δ	.118	5.2	–2.6	.097
Non-ELL	429	APE	.695	52.2%	98.6%	.192
ELL	227	APE	.666	59.0%	99.1%	.016
Non-ELL – ELL	—	APE Δ	.029	–6.8	–0.5	.176

Table 3: Scoring performance by gender and ELL status.

Note. Baseline = the Baseline prompting approach; APE = the APE approach. Exact and adjacent agreement differences are expressed in percentage points. *Group rows report absolute SMD; contrast rows report the magnitude of the between-group difference in absolute SMD.

References

- Scott Andrew Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Meg Benner, and Ulrich Boser. 2024. [A large-scale corpus for assessing written argumentation: PERSUADE 2.0](#). *Assessing Writing*, 61:100865.
- DSPy. 2026. [DSPy source code \(version 3.2.1\)](#). GitHub. Version 3.2.1, released May 5, 2026.
- Keno Harada, Lui Yoshida, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. [Automated refinement of essay scoring rubrics for language models via reflect-and-revise](#). In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pages 771–789, San Diego, California, USA. Association for Computational Linguistics.
- Yue Huang, Corey Palermo, Ruitao Liu, and Yong He. 2025. [An early review of generative language mod-](#)

- els 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):5.
- Yue Huang, Corey Palermo, and Joshua Wilson. 2026. Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*, 69:101047.
- Yue Huang and Joshua Wilson. 2025. Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.
- Yenisel Plasencia-Calaña. 2025. Operationalizing automated essay scoring: A human-aware approach. *arXiv*, 2506.21603. Preprint.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with “gradient descent” and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.
- Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Serkan Ö. Arik. 2024. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 58174–58244. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Joshua Wilson and Yue Huang. 2024. Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing*, 60:100815.
- Mingfeng Xue, Yunting Liu, Xingyao Xiao, and Mark Wilson. 2025. Automatic prompt engineering for automatic scoring. *Journal of Educational Measurement*, 62(4):559–587.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.