MAGIC: Multi-Agent Argumentation and Grammar Integrated Critiquer

Joaquín Jordán*, Xavier Yin*, Melissa Fabros*, Gireeja Ranade, Narges Norouzi

UC Berkeley
{jjordanoc, nzxyin, mfabros, gireeja, norouzi}@berkeley.edu

Abstract

Automated Essay Scoring (AES) and Automatic Essay Feedback (AEF) systems aim to reduce the workload of human raters in educational assessment. However, most existing systems prioritize numerical scoring accuracy over feedback quality and are primarily evaluated on pre-secondary school level writing. This paper presents Multi-Agent Argumentation and Grammar Integrated Critiquer (MAGIC), a framework using five specialized agents to evaluate prompt adherence, persuasiveness, organization, vocabulary, and grammar for both holistic scoring and detailed feedback generation. To support evaluation at the college level, we collated a dataset of Graduate Record Examination (GRE) practice essays with expert-evaluated scores and feedback. MAGIC achieves substantial to near-perfect scoring agreement with humans on the GRE data, outperforming baseline LLM models while providing enhanced interpretability through its multi-agent approach. We also compare MAGIC's feedback generation capabilities against ground truth human feedback and baseline models, finding that MAGIC achieves strong feedback quality and naturalness.

Supplement — https://github.com/magic-aes/MAGIC

1 Introduction

Automated Essay Scoring (AES) and Automated Essay Feedback (AEF) are gaining importance in educational assessment, aiming to replicate human evaluation of written work based on content, coherence, grammar, and style (Dikli 2006). While AES systems have achieved notable success in predicting human-assigned numerical scores, generating meaningful, personalized essay feedback at scale remains an open problem (Behzad, Kashefi, and Somasundaran 2024).

The limitations of commonly used datasets for multi-trait scoring and feedback evaluation have created a gap in AES and AEF research. The NLP research community has invested in compiling essay datasets such as TOEFL11 (Blanchard et al. 2013), ASAP++ (Mathias and Bhattacharyya 2018), and ICLE++ (Li and Ng 2024). However, most of these datasets do not provide ground truth per-trait scores, feedback, or exist behind a paywall. Furthermore, many of the commonly used datasets are collected from writers as part of English as a second language (L2) exams or come from high-school populations of native speakers of English (L1), such as ASAP (Hewlett Foundation 2012) and PERSUADE 2.0 (Crossley et al. 2024). Feedback to L2 English learners will focus on different qualities than feedback for learners with native English who are still developing their writing and critical thinking skills (Pan, Reppen, and Biber 2016). As for essays written by L1 speakers, the reliance on ASAP-based evaluation limits robust exploration of how to improve AES and feedback for L1 English writers above the 10th grade level (Li and Ng 2024).

AES systems are currently most useful in large-scale testing environments where thousands of writing samples must be scored quickly. In classrooms, where students are learning to produce effective essays, numerical scores or grades alone do not improve learning (Guskey 2019). Writing and argument are central to both educational development and intellectual growth. Riddell (2015) therefore advocates for frequent feedback with increased writing opportunities as a recipe for higher learning outcomes. However, scaling instructors' feedback capacity without sacrificing quality remains an ongoing challenge (Page 1966). Since our educational aims should prioritize teaching "intelligent humans" over intelligent tutoring systems (Baker 2016), AI feedback on writing tasks must be integrated thoughtfully and carefully to enhance rather than diminish learners' cognitive engagement.

When using generative AI for feedback generation, like Favero et al. (2025), we found frontier large language models such as OpenAI's ChatGPT and Anthropic's Claude capable of discerning nuances of argument and grammar. However, these enterprise-level API-based models are expensive for school systems to support and difficult to guarantee privacy and accessibility. Therefore, smaller open-sourced models are advantageous for their computational efficiency and open-weights. Educators and administrators can deploy these systems locally to ensure student privacy, model observability, and prediction explainability.

We introduce Multi-Agent Argumentation and Grammar Integrated Critiquer (MAGIC), a generic framework for zero-shot multi-trait AES using independent small LLM agents to grade and provide feedback for each writing dimension of a rubric.

*These authors contributed equally.

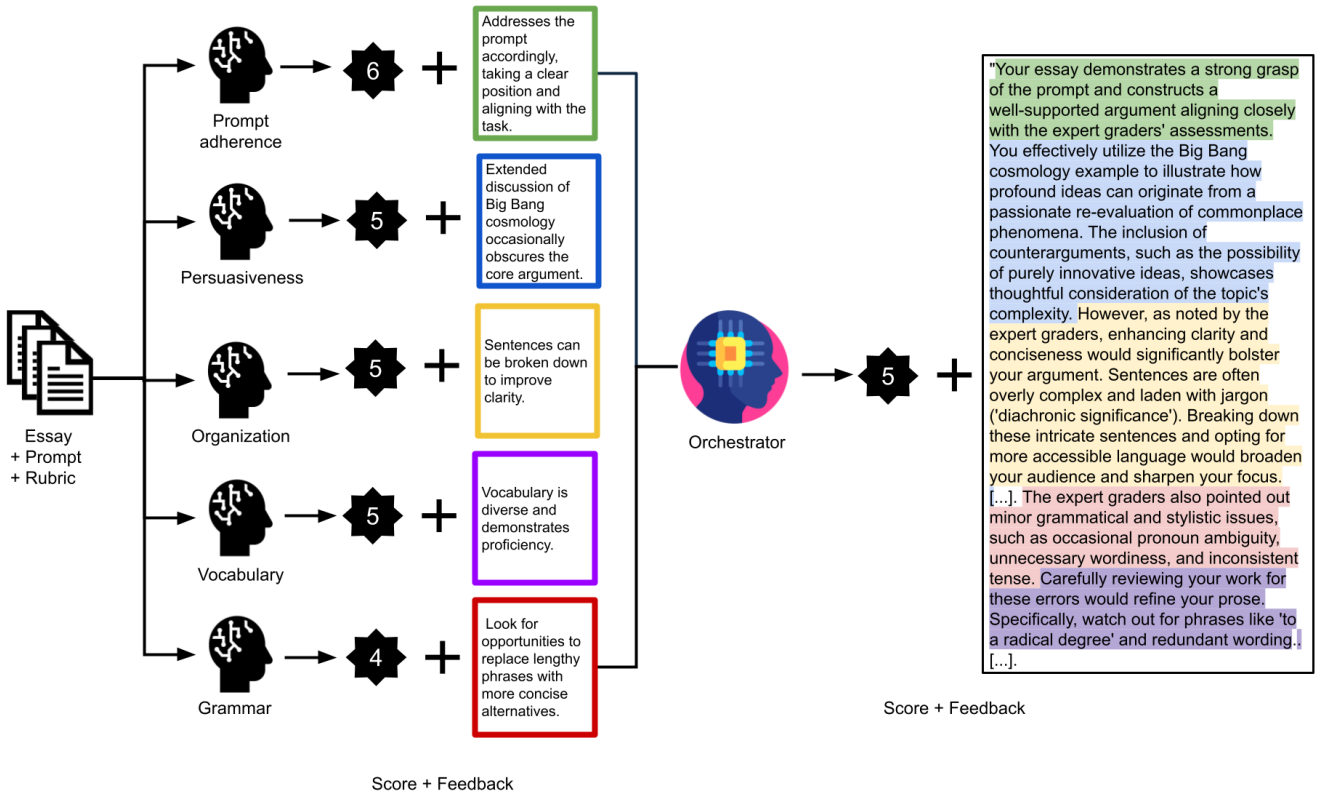

Figure 1: **MAGIC AES Feedback and Scoring Pipeline.** Each agent (prompt adherence, persuasiveness, organization, vocabulary, and grammar) scores the essay separately and provides feedback for their assigned trait. The orchestrator merges the agents' results into a holistic score and combined feedback.

Our experimentation yielded the following insights and contributions to AES and AEF:

1. We propose the use of a compiled Graduate Record Examination (GRE) dataset for assessing AES and AEF model performance in college-level argumentative writing, exhibiting high-quality argumentation, and including feedback ground truths.

2. We present MAGIC, a zero-shot multi-agent framework for AES and AEF, generalizing to various argumentative essay prompts, responses, and rubrics without the need for fine-tuning or training. MAGIC outperformed our single agent baseline in scoring, showing improvements on both small- and medium-sized open-weight LLMs.

3. We provide a systematic analysis of MAGIC's feedback using various metrics to help assess feedback specificity and relevance with the source essay.

All complete prompts and rubrics mentioned in this work, as well as dataset sources, are available in our online supplement.

2 Prior Work

The fields of Automated Essay Scoring (AES) and Automated Essay Feedback (AEF) have evolved from statistical modeling approaches (Page 1967) to modern LLM-based systems, revealing persistent challenges in balancing scoring accuracy with meaningful feedback generation.

2.1 Early Approaches

Earlier NLP studies employed Naive Bayes, Support Vector Machines, and Decision Trees for scoring, with Latent Semantic Analysis for feedback generation (Liu et al. 2017). While handcrafted features offered some interpretability, they remain costly to create and generalize poorly across different prompts and essay types (Misgna et al. 2025). The transition to deep learning marked significant improvements, with transformer-based approaches like R^2BERT outperforming most traditional models (Yang et al. 2020).

2.2 Datasets

The ASAP corpus (Hewlett Foundation 2012) became the de facto AES benchmark, containing essays from US students in grades 7-10, with extensions like ASAP++ (Mathias and Bhattacharyya 2018) adding multi-trait annotations. Other

corpora covering different demographics, TOEFL11 (Blanchard et al. 2013), CLC-FCE (Yannakoudakis, Briscoe, and Medlock 2011), and ICLE++ (Li and Ng 2024), remain underutilized. Critically, research in feedback generation remains limited compared to AES (Misgna et al. 2025; Wang, Lee, and Park 2022), with existing datasets focusing primarily on lower-grade or second-language writing, creating a significant gap for college-level argumentative writing evaluation.

2.3 Large Language Models and Zero-shot AES

LLMs initially showed weak correlation with human evaluations on the ASAP dataset, though prompt engineering with few-shot examples improved alignment (Kundu and Barbosa 2024). Naismith, Mulcaire, and Burstein (2023) achieved QWK nearing 0.80 with GPT-4 using task instructions and rubrics, while Seßler et al. (2024) found OpenAI o1 outperformed other LLMs across multiple traits. However, GPT-4 performance still does not surpass modern AWE methods (Yancey et al. 2023).

Recent zero-shot approaches address the accessibility limitations of few-shot strategies. Lee et al. (2024) introduced Multi Trait Specialization (MTS), decomposing writing proficiency into distinct traits with trait-specific evaluation, achieving QWK gains of 0.437 on TOEFL11. Shibata and Miyamura (2025) proposed comparative scoring using pairwise judgments to address bias issues in direct scoring methods.

2.4 Feedback Generation Challenges

Early feedback work achieved only 50% appropriate response rates with seq2seq models (Hanawa, Nagata, and Inui 2021). Villalon et al. (2008) developed Glosser, an LSA-based model for topic-specific feedback, though it exhibited a bias toward longer sentences despite lacking coherence. Recent studies found GPT-4 to be error-prone in Grammar Error Explanation with frequent hallucinations (Song et al. 2023). Joint scoring and essay feedback generation using Chain-of-Thought achieved only 0.533 QWK (Stahl et al. 2024), far below state-of-the-art scoring baselines of 0.79 (Yang et al. 2020). While some studies show promise for EFL learners (Han et al. 2024) and comparisons with forum feedback (Behzad, Kashefi, and Somasundaran 2024), fundamental challenges remain in providing interpretable, high-quality feedback that matches human expertise.

This background motivates the need for transparent, multi-agent frameworks that provide both accurate scoring and interpretable feedback.

3 Dataset

We collated exam preparation material published for free by Educational Testing Services (ETS) for Graduate Record Examination (GRE) self-study and exam transparency. The GRE exam consists of multiple choice questions and an essay response, and test-takers are most often students writing at the post-secondary or university level who have more experience and facility with the argumentative essay genre, consisting of a mix of both L1 and L2 English writers.

This ground truth evaluation set is comprised of "Analytical Writing Sample Essays with Commentaries" from 48 GRE essays in eight essay prompts, each with holistic scores between 1–6 and associated human qualitative feedback based on a provided rubric (Educational Testing Service 2023).

Furthermore, in order to have corresponding ground truth not only for holistic scores but also at the trait level, we created five trait-based sub-rubrics (T) based on the holistic rubric, where each feature can be scored between 1–6:

T1. Quality of the response to the prompt instructions

T2. Considering the complexities of the issue

T3. Organizing, developing, and expressing ideas

T4. Vocabulary and sentence variety

T5. Grammar and mechanics

One of the authors with previous experience in GRE essay grading annotated each essay with a 1–6 score for every trait.

We argue that this dataset is more representative of the typical workloads teachers might encounter in the higher-education classroom, being unable to train large neural AES models because of the scarce data and having to rely on zero-shot approaches.

4 Methodology

4.1 MAGIC: A Multi-Agent Approach

MAGIC (Multi-Agent Argumentation and Grammar Integrated Critiquer) is an adaptable framework for zero-shot multi-trait AES. MAGIC decomposes holistic assessment into specialized LLM agents, each evaluating a distinct writing dimension specified by a rubric before an orchestrator agent produces the final score and feedback. Our proposed framework adapts to different essay types and rubrics by using modular prompts for each agent.

Unlike previous work using LLMs for holistic scoring through single prompting strategies, our approach isolates each rubric trait in a separate agent interaction. This method, similar to (Lee et al. 2024), enables deeper model reasoning for individual traits. We improve upon this premise by introducing an orchestrator agent that integrates trait-level assessment into holistic scores and feedback, improving scoring agreement with human raters. Our summarized approach is shown in Figure 1.

For this work, we instantiate MAGIC with the GRE argumentative writing evaluation task, as described in the Dataset section. We employ five specialized agents corresponding to our rubric traits T1–T5. Agents T1–T3 evaluate the argumentative qualities of the essay, agent T4 assesses vocabulary, and agent T5 focuses on grammar. The orchestrator then receives these assessments and synthesizes the final feedback and score. All agents are aware of the given essay's content and prompt, but each specialized agent only has access to their trait's rubric to ensure isolation.

While our evaluation focuses on argumentative essays, our framework can be extended to other types of essays (e.g. narrative essays) by replacing the model's prompts and providing new rubrics for the desired task.

4.2 Evaluation

We built MAGIC with small open-source instruction-tuned LLMs in mind, given their accessibility in educational environments and competitive performance in AEF-related tasks (Favero et al. 2025). The LLMs used for our evaluation were Llama 3.1 8B, Gemma 3 12B, and Gemma 3 27B. Our baseline model consists of a single LLM with a CoT prompt producing a holistic score and feedback.

Aligning with standard AES research, we apply Quadratic Weighted Kappa (QWK) to measure score agreement, which penalizes larger score differences more heavily than smaller ones (Table 1).

QWK score	Agreement
≤ 0	None
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Near-Perfect
1.00	Perfect

Table 1: **Explanation of QWK score ranges.** The interpretation of different QWK score ranges used in our evaluation, as presented in Landis and Koch (1977).

To demonstrate its capabilities, we evaluated MAGIC against our baseline results against the collated GRE essay set for both holistic scoring and feedback generation. We report QWK and RMSE results for both baseline and MAGIC variants. Beyond holistic scoring, we also measure per-trait QWK of the independent agents in MAGIC against human-annotated per-trait ground truth scores.

Moreover, to assess the quality of the feedback generated by our models, we annotated human–LLM feedback pairs. We used the following evaluation criteria (C) to assess feedback quality (Behzad, Kashefi, and Somasundaran 2024):

C1. Which is more relevant to the essay content?

C2. Which is better at highlighting weaknesses?

C3. Which is better at highlighting strengths?

C4. Which is more specific and actionable?

C5. Which is more helpful for a student overall?

Additionally, we assessed MAGIC's feedback in the absence of human annotators using an LLM judge (Zheng et al. 2023), and we obtained aligned feedback judgments. For this judge agent, we used OpenAI o4-mini reasoning model, via OpenAI's API, with the "medium" reasoning level for all experiments.

4.3 Experiment Infrastructure

We used one NVIDIA A100 GPU with 80GB of VRAM, for 10 hours to run the scoring and feedback generation experiments. We used vLLM for GPU-efficient LLM inference.

5 Results

5.1 Scoring Agreement Against Humans

We compared QWK scores and feedback quality across all our tested LLM models for baseline and MAGIC. Our results in Table 2 show that all three of the models saw an increase in QWK between baseline and MAGIC configurations. The largest increase was observed in Gemma 3 12B, from moderate (0.607) to substantial (0.762). Gemma 3 27B had the highest QWK for both base and MAGIC configurations, 0.613 and 0.766 respectively. As for RMSE, Gemma 3 12B had the largest improvement from baseline to MAGIC, decreasing RMSE by 0.269. Gemma 3 12B MAGIC outperforms the Gemma 3 27B MAGIC marginally on RMSE despite the 27B model having better baseline RMSE. These results highlight MAGIC's effectiveness for AES tasks.

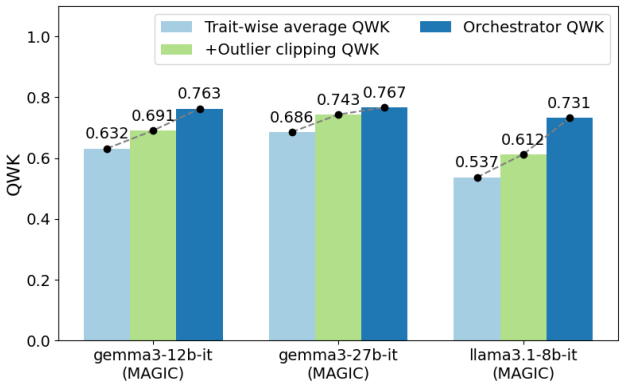

Figure 2: **Comparison of holistic score QWK** between taking the average across traits (Trait-wise QWK), adding an additional outlier clipping and scaling stage (+Outlier clipping QWK), as shown in Lee et al. (2024), and using an orchestrator agent (Orchestrator QWK).

We designed an orchestrator agent to produce a holistic score, as opposed to using a simpler aggregation scheme on the per-trait scores. We compared its score against human scores across different aggregation strategies: simple averaging over the trait scores, averaging with outlier clipping and scaling as shown in Lee et al. (2024). We observed that the holistic score readout provided by these other strategies yields lower concordance with human scores. Instead, having the orchestrator consider all the agents' scores and feedback and then produce its own holistic score shows improved concordance, as shown in Figure 2. This might suggest unequal weighting of the traits by the human graders, which the orchestrator can more faithfully capture.

Moreover, on the agent scoring level, we see that the agents' scores tend to avoid extremes but roughly follow the human score distribution (Figure 3). We also observe some of the grade inflation tendencies associated with LLM-based AES systems. Our experiments with generating feedback and scores holistically and on a feature-trait level corroborate recent findings where LLM graders were more likely to assign scores near the average, and humans assigned a wider range of scores (Smith and Crossley 2025).

Model	QWK base ↑	QWK MAGIC ↑	ΔQWK ↑	RMSE base ↓	RMSE MAGIC ↓	ΔRMSE ↓
gemma3-12b-it	0.607	0.762	**0.155**	1.263	**0.994**	**−0.269**
gemma3-27b-it	**0.613**	**0.766**	0.153	**1.250**	1.010	−0.240
llama3.1-8b-it	0.605	0.731	0.126	1.307	1.163	−0.144

Table 2: **AES performance comparison.** Quadratic Weighted Kappa (QWK) and Root Mean Squared Error (RMSE) are measured against the ground truth GRE scores. The QWK and RMSE results for both baseline and MAGIC as well as the change in each metric (Δ) are shown for the three chosen LLMs. The best result for each column has been bolded.

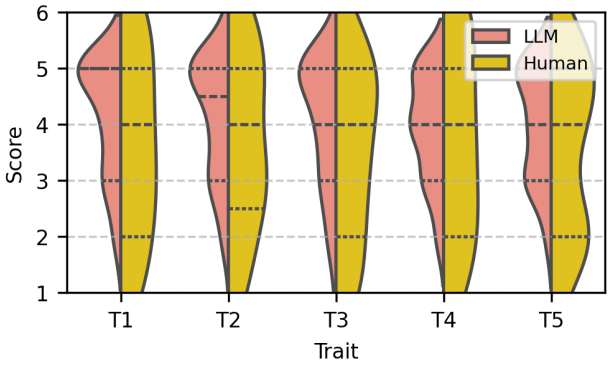

Figure 3: **Per-trait score distributions** for Gemma 3 27B LLM (left) and human (right) annotated ground-truth on the GRE dataset. Quartiles highlighted as dotted lines.

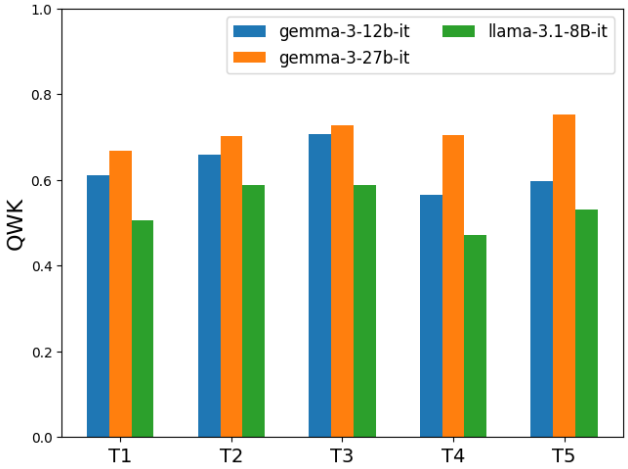

Figure 4: **Per-trait QWK of MAGIC independent agents across different base LLMs.** Our writing dimension traits (T1–T5) are as described in our methodology.

Further breaking down the QWK score, Figure 4 shows the per-trait QWK between LLMs and human ground truth. We evaluated the per-trait scoring capabilities of each of our agents against per-trait human ground truths, reaching moderate to substantial agreement between MAGIC scores and human scores on each of the traits. We notice an LLM trend to score higher on argumentative-related criteria (T1–T3) than on vocabulary (T4) and grammar (T5). Moreover, Gemma 3 27B offers the highest agreement with ground truth scores, consistent with our results on holistic score QWK.

5.2 Comparison Against Human Feedback

For subsequent experiments in AEF, we selected Gemma 3 27B baseline and Gemma 3 27B with MAGIC as our LLM models.

To compare the quality of the generated feedback, we paired models in an A–B test "battle," similar to Chatbot Arena (Chiang et al. 2024), using an LLM judge following the previous feedback assessment criteria (C1–C5) as explained in the Methodology Section. We compute the average majority win-rate over all 5 criteria. Results are shown in Figure 5. Due to previous works such as (Zheng et al. 2023) demonstrating that strong LLMs are capable of having a high level of agreement with humans, we believe that this judging strategy provides a scalable way to evaluate essay feedback. At the same time, the authors note that strong LLMs tend to also prefer LLM-generated responses (Zheng et al. 2023). Further analysis of the Judge model is in the "Judging Feedback" section.

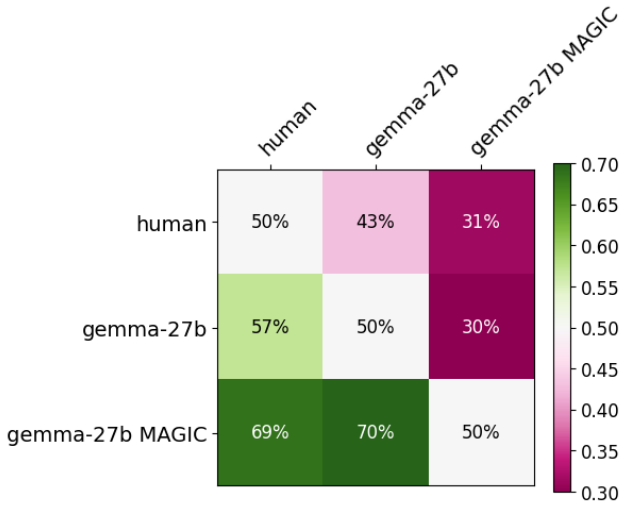

Figure 5: **Head-to-head Model Feedback Win-rates as Rated by a Judge LLM (o4-mini).** Value at row i and column j denotes the average win-rate of row i over column j across all 5 criteria (C1–C5).

Gemma 3 27B without MAGIC wins marginally more often than losing with a 57% winrate against the ground-truth human-written GRE feedback. With MAGIC, the

winrate jumps to 69% against human feedback and 70% against baseline Gemma 3 27B. However, LLM-judges have been shown to have biases towards positive evaluations (Koutcheme et al. 2024), so we further inspect and analyze the outputs of our best performing model: Gemma 3 27B with MAGIC.

5.3 Feedback Characteristics

We observe that MAGIC feedback was statistically on par with the human feedback on the word count axis. MAGIC's feedback length saw only 14% difference. Human feedback was on average longer, however, MAGIC feedback remained more consistent across score points. MAGIC averaged up to twice as many words than human evaluators for the lowest scoring samples.

While LLM-generated feedback demonstrates comparable lexical diversity (MATTR: 0.76-0.78 vs 0.71), human feedback exhibits a 60% larger vocabulary (1,705 vs 1,050-1,127 unique words), higher entropy (8.52 vs 8.11-8.25), and greater use of unique expressions (.534 vs .49-.51 hapax legomena). This suggests that while LLMs maintain consistent linguistic variety within individual feedback instances, human experts draw from a richer and less predictable lexical repertoire across the corpus (Table 3).

Metric	Human	MAGIC	Baseline
Vocabulary Size	1,705	1,127	1,050
MATTR	0.7081	0.7764	0.7620
Repetition Rate (%)	90.3	91.6	91.2
Hapax Ratio (%)	53.4	49.2	51.0
Entropy	8.5220	8.2505	8.1075

Table 3: **Corpus wide Human-LLM lexical diversity** LLM feedback shows good text diversity but reveals lower overall vocabulary richness against human feedback

The individual agent feedback is centered on how to improve the specific essay and provides specific and actionable recommendations for this essay sample in a single turn alongside a numerical score. Peer research such as Lee et al. (2024) required a second conversation turn to collect essay quotes to justify the trait score. MAGIC appears to avoid generic assessments by tailoring its feedback on how to improve the essay at hand along a specific trait. We also observe MAGIC producing feedback that spent more time highlighting the essay sample's strengths (C3) over the human feedback, which appeared to better highlight the essay's weaknesses (C2). It also tends to reference the other agents as expert graders or aspect evaluators as part of its rationale for its own commentary.

We measure MAGIC's ability to "personalize" or target its feedback by looking at Jaccard Similarity between the source essay and the feedback text. LLMs are known to provide generic non-specific advice, and we test MAGIC's feedback specificity by comparing its Jaccard scores against an unrelated and a random text. Jaccard Similarity measures direct overlap between words in the feedback and essay to which it is responding. We see this metric as a proxy for textual references and quotations that the feedback text uses to engage with the specificity of the essay.

As seen in Figure 6, all types of feedback are clear improvements over random text. While the orchestrated feedback closes the gap toward the human level specificity, it is a statistically moderate improvement from the unrelated text. The baseline text offers a low statistical advantage over valid unrelated text.

We also evaluate alignment between LLM and human feedback using text similarity metrics (Table 4). Overall, LLM feedback aligns well with human feedback, as reflected in moderately high BERTScore F1 values—a context-aware metric believed to correlate with human judgments (Rehman et al. 2025). Differences among MAGIC, baseline, or other LLM outputs are not significant, suggesting our prompts already elicit human-like assessments. Meanwhile, ROUGE-L, an LCS-based metric, remains low, indicating that LLM wording choices diverge perceptibly yet consistently, likely due to correct adherence to the shared prompt.

Model	ROUGE-L	BERTScore
gemma-27b	$\mathbf{0.175 \pm 0.007}$	0.851 ± 0.003
gemma-27b MAGIC	0.159 ± 0.005	0.844 ± 0.002
gemma-12b	0.174 ± 0.008	$\mathbf{0.853 \pm 0.002}$
gemma-12b MAGIC	0.162 ± 0.006	0.845 ± 0.002
llama-8b	0.163 ± 0.009	0.852 ± 0.003
llama-8b MAGIC	0.166 ± 0.018	0.849 ± 0.006

Table 4: **Human-LLM feedback alignment** between LLM feedback outputs and ground truth feedback. F1 scores are reported for both as mean with 95% confidence intervals. Best results per-column are highlighted in bold.

5.4 Judging Feedback

Criteria	κ_{IAA}	κ_{AJA}
C1	0.208	0.211
C2	0.556	0.476
C3	0.468	0.583
C4	0.287	0.139
C5	0.395	0.236
Overall	0.427	0.382

Table 5: **Inter-annotator Agreement Table.** Inter-Annotator Agreement κ_{IAA} and Adjudicator–Judge agreement κ_{AJA}, both calculated using the Cohen's Kappa statistic. Rows C1–C5 represent agreement for the specified feedback criteria, and row "Overall" represents the agreement over all the criteria.

To evaluate the quality of the LLM-as-a-judge for feedback evaluation, two of this paper's authors annotated all 48 MAGIC–human feedback pairs for each of the five criteria (C1–C5) with a label of "LLM" or "Human" to decide a winner for each criterion. A third author then adjudicated any split votes to determine the "ground truth" winner, so

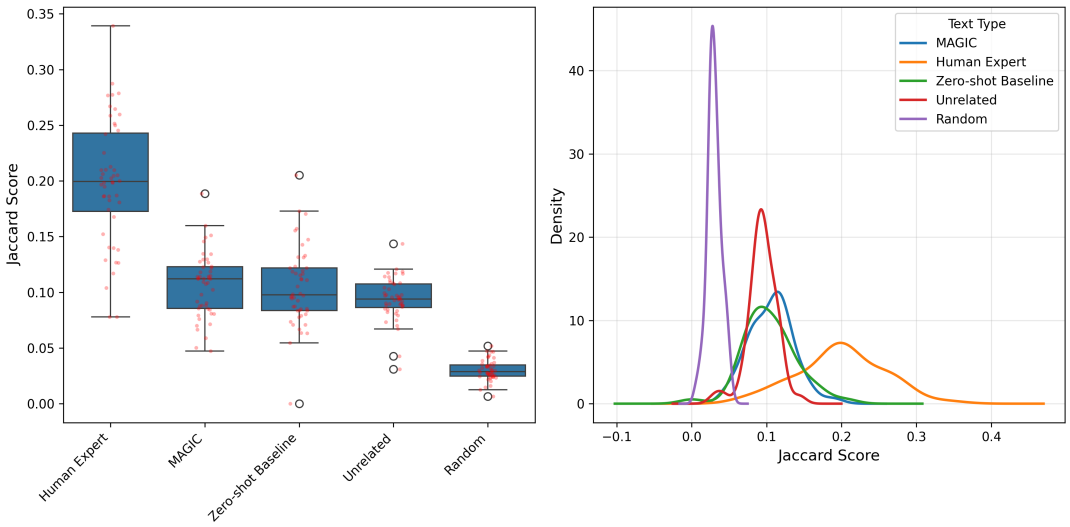

Figure 6: **Jaccard Similarity between different feedback conditions and essay text.** Left: Red dots represent essays, y-axis is Jaccard Similarity. Right: Distribution of Jaccard scores (KDE). Unrelated and random are baselines with an unrelated text passage and nonsensical random text.

each criterion winner had at least 2 human votes. We then computed the Cohen's Kappa between the two annotators and between the adjudicator and the judge LLM as shown in Table 5. The overall (across C1–C5) Inter-Annotator Agreement was moderate ($\kappa_{\text{IAA}} = 0.427$) while the overall Adjudicator–Judge agreement was fair ($\kappa_{\text{AJA}} = 0.382$).

Our analysis of LLM-as-a-judge shows considerable agreement between human-adjudicated preferences and LLM (o4-mini) preferences for C2 and C3, indicating that the LLM judge can reliably assess which feedback has better analysis of strengths and weaknesses. Meanwhile C1, C4, and C5 had slight to fair agreement. This might be due to these criteria being more subjective and to the feedback quality being roughly similar between human and MAGIC. Finally, while the Judge LLM prefers the MAGIC feedback 69% of the time, the adjudicator only gave MAGIC preference 52% of the time, averaged across C1–C5, which could be due to the positive character of the prompt used for our study (Koutcheme et al. 2024).

6 Conclusion

We introduced MAGIC, a zero-shot framework for multi-trait AES and AEF. We evaluated MAGIC on a set of collated GRE essays with feedback ground truths and demonstrated the effectiveness of our approach. To the best of our knowledge, the utility of LLMs for both zero-shot AES and AEF in L1 and L2 college-level writing had not been fully demonstrated prior to this work, breaking the long-standing over-reliance on ASAP and variants.

We find that MAGIC's orchestrated multi-agent design improves scoring agreement relative to a single agent baseline. To accomplish this, we developed an enhanced version of the GRE dataset where each of the five traits was scored for every essay. We then assessed both overall agreement with GRE holistic scores and per-agent reliability using annotated per-trait scores, demonstrating consistent performance on both levels.

Furthermore, we assess MAGIC's feedback quality using different methods. We propose an LLM judge to automatically evaluate the feedback at scale, and then evaluate its agreement with human preferences by manually annotating feedback pairs. We observe an overall fair agreement between judge and human decisions, and conclude that MAGIC generated feedback has quality comparable to humans.

The dataset's high quality enabled us to formally characterize feedback with respect to its source essay. In particular, we developed a metric for feedback specificity, a novel measurement for AEF research. Based on the Jaccard similarity scores between a ground truth essay and its associated feedback, we pose a threshold where generated feedback should outperform a random, yet valid, text, by at least one standard deviation, to affirm that its advice is targeted rather generic. MAGIC exceeds baseline feedback under this measure, but its outputs could be improved with clearer and guided prompting. Finally, we find that MAGIC's resulting feedback is comparable to human feedback in length, semantic relevance, and lexical diversity measures, further strengthening our claim that the generated feedback is human-aligned.

7 Future Work

The availability of a large corpus containing a mixture of L1 and L2 speech, ground truth scores per-trait and feedback remains to be seen, and is of utmost importance for future work in AES and AEF. Quality feedback and essays are already rare, and essays with feedback for languages outside of English are even more so. The MAGIC framework can be tested on more diverse samples to demonstrate true generalizability.

Ethical Statement

Publicly available material of sample essays, scores, and feedback was collated from publicly distributed legacy ETS study material to build the ground truth evaluations. The corpus does not contain personally identifiable or sensitive information.

All the texts and scores collated as ground truth are accessible under Fair Use guidelines and have been made freely available to the public. However, ETS holds formal copyright of GRE material and has not yet approved publication of a standalone data set.

The authors offer this work as a bridge to deepen participation and discussion between students and instructors to motivate the development of critical thinking and writing skills. However, this type of work has the potential to be abused by bad actors to disrupt standardized testing environments by providing unapproved feedback or false scores.

Acknowledgments

The authors appreciate the insight, guidance, and support of Narges Norouzi and Gireeja Ranade, who piloted AI for Education courses at UC Berkeley's Computer Science department for the 2024-2025 academic year. We also thank Nelson Lojo for providing feedback and polish to the drafts. Modal (modal.com) generously supplied research credits to access server infrastructure to run exploratory and test experiments.

References

Baker, R. S. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*, 26(2): 600–614.

Behzad, S.; Kashefi, O.; and Somasundaran, S. 2024. Assessing Online Writing Feedback Resources: Generative AI vs. Good Samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1638–1644.

Blanchard, D.; Tetreault, J.; Higgins, D.; Cahill, A.; and Chodorow, M. 2013. TOEFL11: A CORPUS OF NON-NATIVE ENGLISH. *ETS Research Report Series*, 2013: i–15.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132.

Crossley, S.; Tian, Y.; Baffour, P.; Franklin, A.; Benner, M.; and Boser, U. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61: 100865.

Dikli, S. 2006. An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Educational Testing Service. 2023. GRE Practice Test 1 Writing Responses (18-point Large Print): Analytical Writing Sample Essays and Commentaries. Technical Report 835790 Large Print Edition, Educational Testing Service. Accessed: 29 April 2025.

Favero, L.; Pérez-Ortiz, J. A.; Käser, T.; and Oliver, N. 2025. Leveraging Small LLMs for Argument Mining in Education: Argument Component Identification, Classification, and Assessment. arXiv:2502.14389 [cs].

Guskey, T. R. 2019. Grades versus comments: Research on student feedback. *Phi Delta Kappan*, 101(3): 42–47.

Han, J.; Yoo, H.; Myung, J.; Kim, M.; Lim, H.; Kim, Y.; Lee, T. Y.; Hong, H.; Kim, J.; Ahn, S.-Y.; et al. 2024. LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, 284–293.

Hanawa, K.; Nagata, R.; and Inui, K. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9719–9730.

Hewlett Foundation. 2012. Automated Student Assessment Prize (ASAP) Automated Essay Scoring Dataset. https://www.kaggle.com/competitions/asap-aes/data. Accessed: 29 April 2025.

Koutcheme, C.; Dainese, N.; Hellas, A.; Sarsa, S.; Leinonen, J.; Ashraf, S.; and Denny, P. 2024. Evaluating Language Models for Generating and Judging Programming Feedback. arXiv:2407.04873 [cs].

Kundu, A.; and Barbosa, D. 2024. Are Large Language Models Good Essay Graders? arXiv:2409.13120.

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

Lee, S.; Cai, Y.; Meng, D.; Wang, Z.; and Wu, Y. 2024. Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. arxiv:2404.04941 [cs].

Li, S.; and Ng, V. 2024. ICLE++: Modeling Fine-Grained Traits for Holistic Essay Scoring. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8465–8486. Mexico City, Mexico: Association for Computational Linguistics.

Liu, M.; Li, Y.; Xu, W.; and Liu, L. 2017. Automated Essay Feedback Generation and Its Impact on Revision. *IEEE Transactions on Learning Technologies*, 10(4): 502–513.

Mathias, S.; and Bhattacharyya, P. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. In Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; and Tokunaga, T., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Misgna, H.; On, B.-W.; Lee, I.; and Choi, G. S. 2025. A survey on deep learning-based automated essay scoring and

feedback generation. *Artificial Intelligence Review*, 58(2): 1–40.

Naismith, B.; Mulcaire, P.; and Burstein, J. 2023. Automated evaluation of written discourse coherence using GPT-4. In Kochmar, E.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Madnani, N.; Tack, A.; Yaneva, V.; Yuan, Z.; and Zesch, T., eds., *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 394–403. Toronto, Canada: Association for Computational Linguistics.

Page, E. 1966. The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, 47(5): 238–243.

Page, E. 1967. COLING '67: Proceedings of the 1967 conference on Computational linguistics.

Pan, F.; Reppen, R.; and Biber, D. 2016. Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, 21: 60–71.

Rehman, T.; Ghosh, S.; Das, K.; Bhattacharjee, S.; Sanyal, D. K.; and Chattopadhyay, S. 2025. Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets. *arXiv preprint arXiv:2502.19339*.

Riddell, J. 2015. Performance, Feedback, and Revision: Metacognitive Approaches to Undergraduate Essay Writing. *Collected Essays on Learning and Teaching*, 8: 79.

Seßler, K.; Fürstenberg, M.; Bühler, B.; and Kasneci, E. 2024. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring1. arXiv:2411.16337.

Shibata, T.; and Miyamura, Y. 2025. LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models. arXiv:2505.08498 [cs].

Smith, K.; and Crossley, S. 2025. Identifying Limitations and Bias in ChatGPT Essay Scores: Insights from Benchmark Data. *The Cutting Ed.*

Song, Y.; Krishna, K.; Bhatt, R.; Gimpel, K.; and Iyyer, M. 2023. Gee! grammar error explanation with large language models. *arXiv preprint arXiv:2311.09517*.

Stahl, M.; Biermann, L.; Nehring, A.; and Wachsmuth, H. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint arXiv:2404.15845*.

Villalon, J.; Kearney, P.; Calvo, R.; and Reimann, P. 2008. Glosser: Enhanced Feedback for Student Writing Tasks. 454 – 458. ISBN 978-0-7695-3167-0.

Wang, X.; Lee, Y.; and Park, J. 2022. Automated evaluation for student argumentative writing: A survey. *arXiv preprint arXiv:2205.04083*.

Yancey, K. P.; Laflair, G.; Verardi, A.; and Burstein, J. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 576–584.

Yang, R.; Cao, J.; Wen, Z.; Wu, Y.; and He, X. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1560–1569.

Yannakoudakis, H.; Briscoe, T.; and Medlock, B. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189. Portland, Oregon, USA: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.