MAGIC: Multi-Agent Argumentation and Grammar Integrated Critiquer

Anonymous EMNLP submission

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 numeric scoring accuracy over the quality of feedback. This paper presents Multi-Agent Argumentation and Grammar Integrated Critiquer (MAGIC), a framework that uses multiple specialized agents to evaluate distinct writing aspects to both predict holistic scores and produce detailed, rubric-aligned feedback. To support evaluation, we curated a novel dataset of past GRE practice test essays with expert-evaluated scores and feedback. MAGIC outperforms baseline models in both essay scoring , as measured by Quadratic Weighted Kappa (QWK). We find that despite the improvement in QWK, there are opportunities for future work in aligning LLM-generated feedback to human preferences.

1 Introduction

Automated Essay Scoring (AES) and Automated Essay Feedback (AEF) have become important tools in educational assessment, aiming to replicate human judgment in evaluating 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 et al., 2024).

Effective feedback helps students improve their writing, deepens subject-matter understanding, and fosters continuous learning. Graff (2003) opens his book *Clueless In Academe* with the claim that the ability to "listen closely to others, summarize them in a recognizable way, and make your own relevant argument" is central to the education project. Menary (2007) goes further, posing that "writing is thinking in action," and that "Creating and manipulating written sentences are not merely outputs from neural processes but, just as crucially, they shape the cycle of processing that constitutes a mental act."

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. Since our educational aims should prioritize training "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. For example, Liu et al. (2017) tested AI-supported feedback processes around "grammar, spelling, sentence diversity, structure, organization, supporting ideas, coherence, and conclusion," finding that such feedback helped English-as-a-Second-Language (ESL) students revise their work more effectively.

Possible strategies to integrate Generative AI tooling could combine known working strategies such as: Instructor feedback frameworks or grading rubrics (Norton and Norton, 2001), peer editing or social writing opportunities (Kerman et al., 2024), or increased reflection practices through self-revision (Riddell, 2015). Automated Essay Feedback (AEF) systems offer the promise of improving instructors' workloads while supporting students' personalized instruction and learning outcomes at scale (Dikli, 2006).

However, key challenges remain. Subjectivity and bias can arise when human evaluators' preferences or societal stereotypes are learned and perpetuated by models, leading to unfair outcomes (Smith and Crossley, 2025). AES systems also often struggle to generalize across writing prompts

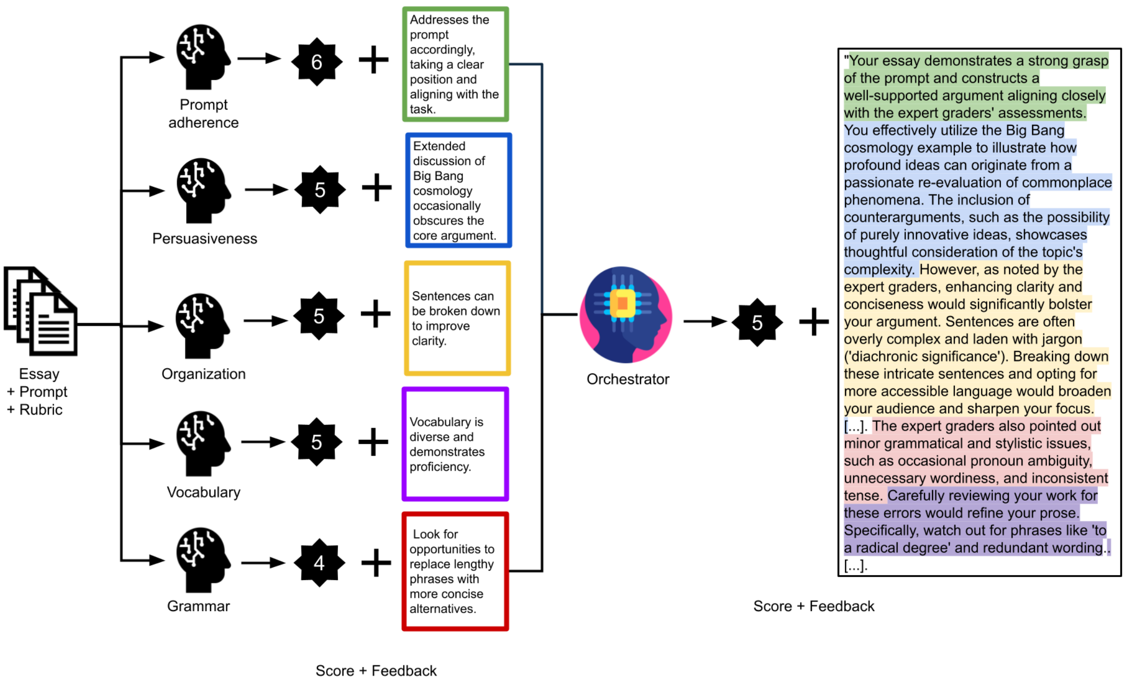

Figure 1: **MAGIC AES Feedback and Scoring Pipeline.** Each agent (prompt adherence, persuasiveness, organization, vocabulary, and grammar) scores the essay separately and provides feedback for their assigned trait. The orchestrator merges the agent's results into a holistic score and combined feedback.

or domains, limiting their usefulness in diverse educational settings (Li and Ng, 2024a). The interpretability of many models is limited, making it difficult for educators and students to understand or trust how scores are determined. (Li et al., 2014).

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. 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.

In this work, we aim to investigate the following research questions:

RQ1. Can a zero-shot, multi-agent approach to AES improve scoring agreement with human graders for educational applications?

RQ2. Can per-trait feedback reasoning with individual agents result in higher quality feedback, greater interpretability, or explainability than human feedback?

RQ3. Can we reliably use small open-source models to score essays and generate feedback in natural language?

To answer these questions, we introduce Multi-Agent Argumentation and Grammar Integrated Critiquer (MAGIC), a modular, agent-based framework that uses multiple LLM-powered agents, each focused on a specific component of argumentative writing, e.g. argument structure, grammar, vocabulary, and comprehension. Each agent provides targeted scores and detailed feedback for its assigned dimension. An orchestrator model integrates these outputs to produce a holistic score and synthesized feedback, simulating the nuanced reasoning of human evaluators. This architecture offers greater transparency, flexibility, and extensibility than monolithic AES and feedback systems. We evaluate our framework against existing ground truth essays that have numeric and qualitative as a baseline, and we analyze feedback quality using LLM-as-a-judge protocols. Results show improvements in scoring robustness and in the clarity and usefulness of feedback provided to students.

2 Prior Work

Earlier Natural Language Processing (NLP) studies on AES and AEF strategies focused on Naive Bayes, Support Vector Machines, and Decision Trees for scoring and Latent Semantic Analysis (LSA) for generation feedback (Liu et al., 2017). While AES and providing useful student feedback have long been of interest to the NLP community, the early methods proved to be brittle compared to the quality of human feedback. Villalon et al. (2008) developed Glosser, an LSA-based model, to provide feedback assistance to students on topic clusters. Specifically, Glosser's LSA technique created a vector space model representation of a text, and then a singular value decomposition (SVD) technique was applied to the created matrix. Model bias favored longer sentences despite lack of coherence, and other failure modes were seen with shifts between an essay's title and its first paragraph.

Released via a 2012 Kaggle competition, the Automated Student Assessment Prize (ASAP) corpus (Hewlett Foundation, 2012) rose to become the de facto AES benchmark, containing thousands of essays per prompt written by US students in grades 7-10. Extensions like ASAP++ (Mathias and Bhattacharyya, 2018) add multi-trait annotations, such as organization, and coherence. Other corpora which cover different demographics, e.g. TOEFL11 (Blanchard et al., 2013), CLC-FCE (Yannakoudakis et al., 2011), and ICLE++ (Li and Ng, 2024b) are underutilized in the literature.

Previous research shows that LLMs exhibit weak correlation with human evaluations on ASAP. Prompt engineering via few-shot examples was found to improve the alignment between human and LLM ratings (Kundu and Barbosa, 2024). Stahl et al. (2024) show that joint scoring and feedback using LLMs improves the model's AES ability without affecting the quality of feedback.

Naismith et al. (2023) found that GPT-4, when prompted with task instructions, rubric criteria, guidelines, and few-shot examples, achieves significantly higher agreement with human scores for discourse coherence than a baseline linear regression model built on Coh-Metrix features, with QWK above 0.80 across configurations. They also show that GPT-4 can generate clear, rubric-aligned rationales for its ratings and envision future work applying this ability to feedback generation. Seßler et al. (2024) found that OpenAI o1 outperformed all other LLMs at grading based on 10 different traits.

All of the prior work in LLM-based AES emphasize ongoing issues with calibrating score outputs to human scores.

3 Dataset

We identified a gap in AES research due to the limitations of commonly used datasets. The NLP research community has invested in compiling essay datasets, most often collected from writers as part of English as a Second Language (L2) exams, e.g. TOEFL11 (Blanchard et al., 2013). Feedback to L2 English learners will focus on different qualities than learners with native English who are still developing their writing and critical thinking skills (Pan et al., 2016). As for essays written by native speakers of English (L1), the reliance on ASAP-based evaluation limits robust exploration of how to improve AES and feedback to L1 English writers above the 10th grade level (Li and Ng, 2024b).

Therefore, we seek essays produced by more sophisticated writers of English beyond the 7th–10th graders of the ASAP corpus. GRE 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. ETS has updated and adapted the GRE essay task over the years, yet the core expectation of the essay product—a persuasive, coherent and compelling position on a topic—has remained the same. The body of legacy essays and associated feedback is still available to the public and constitutes a valuable resource for benchmarking feedback and scores for systems designed to assist college-level writers (Educational Testing Service, 2023).

For evaluation of MAGIC, we collated legacy exam preparation material published by Educational Testing Services (ETS) for the Graduate Record Examination (GRE). The Graduate Record Exam consists of multiple choice questions and an essay response. This ground truth contains 48 essays in eight essay prompts, each with holistic scores between 0–6 and feedback explaining the score. An example of our collected GRE data and a table of sourced practice tests can be found in Appendix B1 and B2.

All the texts and scores collated as ground truth have been made freely available to the public, and the ground truth is accessible under Fair Use guidelines. However, ETS is the formal copyright holder, and the organization has not yet approved publication as a dataset.

4 Methodology

4.1 MAGIC: A Multi-Agent Approach

MAGIC AES (Multi-Agent Argumentation and Grammar Integrated Critiquer) is a framework for zero-shot multi-trait AES using independent small LLM agents to grade and provide feedback for each writing dimension of a rubric. For this work, we focus on argumentative essays, but our framework can be extended to other types of essays (e.g. narrative essays) by changing the model's prompts and rubrics.

It differs from previous work using LLMs to score and provide feedback holistically by isolating each aspect (trait) of the rubric with a separate model to foster deeper thinking in the models. To provide holistic scores and feedback as well, we propose the use of an orchestrator agent. Our summarized approach is shown in Figure 1.

We developed a five-agent system based on our decomposed holistic rubric (Educational Testing Service, 2023). Each of these agents has a system prompt with instructions for grading a specific trait. We separated the single standard rubric essay into separate dimensions of writing traits as follows:

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

Specifically, agents grading T1–T3 use the argumentation prompt in Table A3 with the provided trait description, the agent grading T4 uses the vocabulary prompt in Table A5, and the agent grading T5 uses the grammar prompt in Table A4. Finally, the orchestrator uses the prompt in Table A2.

4.2 Evaluation

To demonstrate the capabilities of our method, we evaluated MAGIC against an out-of-the-box LLM on our compiled GRE essays.

The most common metric for AES is Quadratic Weighted Kappa (QWK), which measures agreement between machine and human scores, penalizing larger score differences more heavily than smaller ones.

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.

The QWK scale represents concordance across individual raters, where the higher the value of the score the greater the agreement between raters (Table 1).

In addition to holistic scoring, we measure per-trait QWK of the independent agents in MAGIC against human-annotated per-trait ground truth scores. The ground truth evaluation set are comprised of "Analytical Writing Sample Essays with Commentaries" from different GRE practice exams. The published sample essays have associated human-generated qualitative feedback based on a provided holistic rubric as well as an single numerical score based on the rubric.

We created trait-based sub-rubrics inspired by the holistic rubric, where each feature can be scored between 0–6. In order to have corresponding ground truth with the trait-based level, we annotated each essay with a 0–6 score for each essay dimension trait. We assigned an LLM agent to perform a single assessment on the featured trait. After the agent assessments are completed, an orchestration agent compiles all the feedback and scores to provide its own holistic score and feedback recommendations.

Moreover, to assess the quality of the feedback generated by different essay feedback models, we used an LLM-as-a-judge approach. See Table A6 for the prompt we used for the judge. We then used the following criteria to assess feedback quality: (Behzad et al., 2024):

C1. Which is more relevant to the essay content?

C2. Which is better at highlighting weakness?

C3. Which is better at highlighting strengths?

Model	QWK ↑	C1 (%) ↑	C2 (%) ↑	C3 (%) ↑	C4 (%) ↑	C5 (%) ↑
gemma3-12b-it (baseline)	0.680	70.8	58.3	**66.7**	93.8	91.7
gemma3-12b-it (MAGIC)	**0.813**	**71.1**	**80.0**	53.3	**100.0**	**100.0**
gemma3-27b-it (baseline)	0.618	**77.1**	70.8	**60.4**	**100.0**	**100.0**
gemma3-27b-it (MAGIC)	**0.738**	64.6	**77.1**	43.8	**100.0**	**100.0**
llama3.1-8b-it (baseline)	0.591	**13.6**	**31.8**	**43.2**	70.5	56.8
llama3.1-8b-it (MAGIC)	**0.705**	8.3	16.7	25.0	**75.0**	**58.3**
llama3.3-70b-it (baseline)	0.689	4.2	16.7	**25.0**	56.2	41.7
llama3.3-70b-it (MAGIC)	**0.711**	**13.0**	**26.1**	8.7	**73.9**	**60.9**

Table 2: **Performance of AES and feedback generation on our GRE dataset.** QWK is measured against the ground truth GRE scores. Columns C1 to C5 contain the win-rates of LLM against ground truth GRE feedback for each of the criteria, as defined in Section 4, using OpenAI's o4-mini as the judge LLM. The highest performing result for each base model per column has been bolded.

C4. Which is more specific and actionable?

C5. Which is more helpful for a student overall?

4.3 Experiment Infrastructure

We primarily used five NVIDIA A100 GPUs, each with 80GB of VRAM, for a total combined time of 300 hours to run the experiments with scoring feedback generation, ablation studies, and LLM as a judge evaluations.

For the LLM-as-a-judge, we used OpenAI o4-mini reasoning model, via OpenAI's API, with the "medium" reasoning level for all experiments.

5 Results

5.1 Scoring Agreement Against Humans

We first compare QWK scores and feedback quality on the GRE dataset for baseline and MAGIC for several open-source instruction-tuned LLMs: Llama 3.1 8B, Llama 3.3 70B, Gemma 3 12B, and Gemma 3 27B. Our results in Table 2 show that all of the four models tested saw an increase in QWK when using MAGIC, with the largest increase observed in Gemma 3 12B. MAGIC increased the scoring agreement for Gemma 3 12B from substantial to near-perfect agreement, outperforming the larger Gemma 3 27B and Llama 3.3 70B models.

To validate our usage of an orchestrator agent instead of averaging trait scores, we compared agreement between the orchestrator outputs and the human holistic scores against a trait-wise average score baseline. We observe that given a list of per-trait scores output by the agents, averaging independent agents' scores yields lower QWK with human scores than having the orchestrator consider the scores and feedback provided by each of the agents

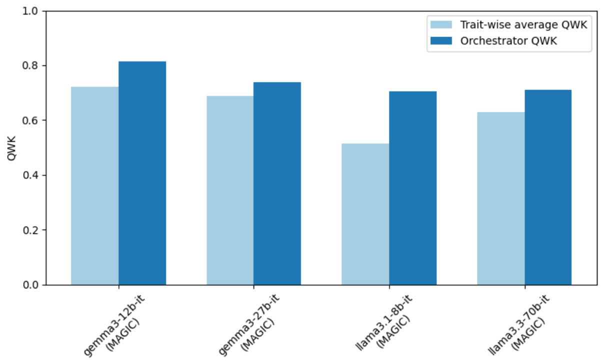

Figure 2: **Comparison of QWK between taking the average across trait scores (Trait-wise QWK).** Using an orchestrator agent to predict holistic scores from trait scores (Orchestrator QWK) have greater concordance with human scoring. QWK differences are 0.092, 0.049, 0.191, and 0.082 for Gemma 3 12B, Gemma 3 27B, Llama 3.1 8B, and Llama 3.3 70B respectively.

and produce its own a holistic score, as shown in Figure 2.

Further breaking down the QWK score, Figure 3 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. Gemma 3 12B shows better performance against human baselines than Llama 3.3 70B when scoring Argument, Organization, and Development (T1, T2 T3), but the opposite is true for and Vocabulary (T4) and Grammar (T5). Gemma 3 27B performs rather well across the board. Llama 3.1 8B shows relatively poor performance compared to the other models we tested.

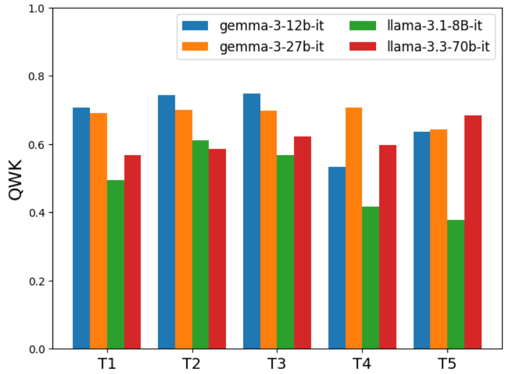

Figure 3: **Per-trait QWK of the MAGIC independent Agents for Different Base Models.** Gemma 3 12B offers high agreement with ground truth scores. Our writing dimension traits (T1–T5) are as described in Section 4.

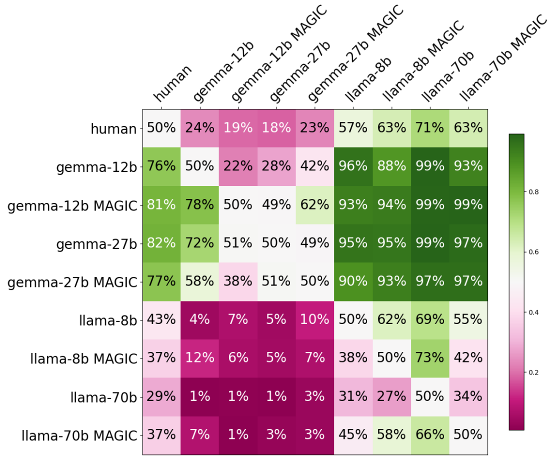

Figure 4: **Head-to-head Model Feedback Win-rates as Rated by a Judge LLM (o4-mini).** Value at row i and column j denotes the average win-rate of row i over column j across all 5 criteria (C1–C5).

5.2 Comparing MAGIC feedback Against Human Feedback

For Llama 3.3 70B, we see an increase in performance for criteria C1 (relevance to essay content), C2 (highlighting weaknesses), and C5 (overall helpfulness to a student) in Table 2, indicating that MAGIC improved the model's ability to remain relevant to the essay content, highlighting the essay's weaknesses, and being more helpful to the student. At the same time, a lower win-score in trait C3, the criteria of "highlighting strengths," would be expected if the feedback response focused more of its critique about the essay's weaknesses over enumerating the essay's strengths. In our Gemma 3 12B model, the baseline is strong at feedback. Nevertheless, we observe an increase in C2 together with a decrease in C3, which indicates that the model with MAGIC is better able to identify weaknesses but struggles more with strengths. Overall, while MAGIC improves grading performance, measured by QWK, the results for feedback quality (as judged by an LLM) are mixed.

5.3 Feedback Comparison Between Different LLMs

We further test the generated feedback in an A–B test "battle," similar to Chatbot Arena (Chiang et al., 2024), using the previous feedback assessment criteria (C1–C5) as explained in Section 4. The OpenAI o4-mini LLM, with reasoning set to the medium setting, was used to determine the winner for each of the feedback criteria. We compute the average win-rate over the full criteria features as shown in Figure 4.

Gemma models with and without MAGIC perform favorably against human baselines. Additionally, Gemma models vastly outperform all Llama models. We see that Gemma 3 12B with MAGIC beats baseline Gemma 3 12B 78% of the time across the 5 criteria, while Gemma 3 27B performs worse or the same with MAGIC as without. Llama 3.3 70B with MAGIC beats base Llama 3.3 70B 66% of the time and Llama 3.1 8B with MAGIC but still performs comparably with Llama 3.1 8B base. Some of the failure modes and issues for Llama models included: failure to output the correct `json` format required, hallucination of feedback format as email or letter, and shorter and less detailed feedback than Gemma models or human baselines.

5.4 Feedback Characteristics

For subsequent experiments, we selected Gemma 3 12B base and Gemma 3 12B with MAGIC as our LLM baseline and MAGIC models respectively.

After reviewing the feedback outputs, we observed that the feedback produced by MAGIC was on average longer than the feedback provided by the baseline single prompt with a single rubric by 31.4% (Figure 5). The average MAGIC feedback length was 238 words while the baseline average was 181 words long and longer than the average human response at 198 words.

Agents for Traits 4 and 5 produced the shortest feedback responses which are the agents focused on grammar and vocabulary components. Traits 1, 2, 3 are generally focused on essay development and

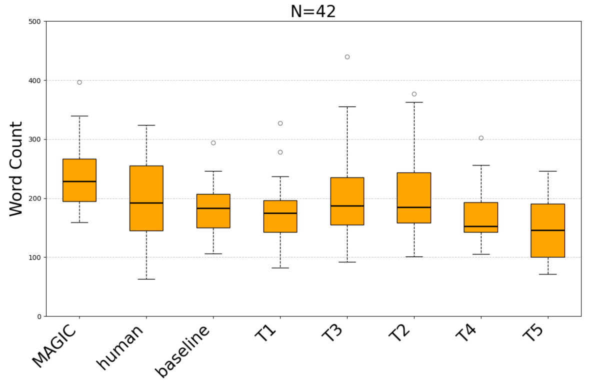

Figure 5: **Comparison of Feedback Response Length.** With a sample of 42 essays, MAGIC produces feedback closest to the human feedback in word length ranges, and MAGIC produces longer feedback than baseline feedback or isolated trait feedback

argument structure. We observed that agents 2 and 3 can generate comparable feedback in length to MAGIC. Given that the complexity of an argument takes longer to explain than a grammar or spelling comment, we can see the argument agents spending more time on the more complex task.

MAGIC often orders its feedback by providing a framing overview of the essay's topic, general strengths and weaknesses and areas for improvement. MAGIC also demonstrates greater insight into the nuance and details of the essay over the baseline's feedback response.

Both baseline feedback and MAGIC appear to produce feedback that tends to spend more time highlighting the essay sample's strengths (C3) over the human feedback which appeared to better highlight the essay's weaknesses (C2) against the rubric.

MAGIC appears to provide more specific and more actionable feedback than the baseline feedback assessment. It tends to reference the other agents as expert graders as part of its rationale for its own commentary. Therefore, hallucination mitigation strategies should also involve observability into the agent's scoring and feedback behavior. See Table C1.

The individual agent feedback is highly centered on how to improve the work on the specific essays and provides specific and actionable recommendations for essay sample.

5.5 Agent Scoring Characteristics

On the agent scoring level, we see that the agents' scores cluster near the middle of the range (Figure 6). While human scorings are more likely to use the full range of scoring values. We also see some of the grade inflation tendencies associated with LLM-based AES systems.

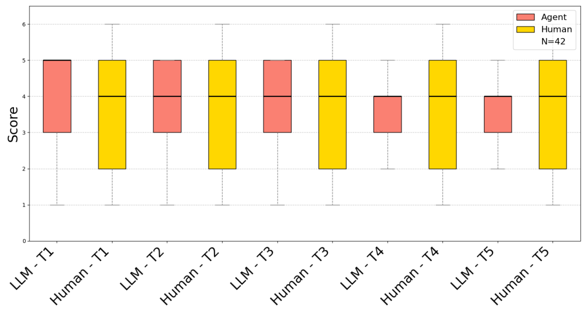

Figure 6: **GRE per-trait score distributions for human ground-truth and LLM (Gemma 3 12B).** Comparison of Essay Trait Scoring. LLM scoring stays close to the means while Human scoring uses the full range of scores

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).

5.6 Judging Feedback

Criteria	κ_{IAA}	κ_{AJA}
C1	0.750	0.182
C2	0.500	0.200
C3	0.314	0.314
C4	0.800	0.000
C5	0.526	0.000
Overall	0.578	0.139

Table 3: **Inter-annotated Agreement Table.** Inter-annotator 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 mean agreement over all the criteria.

To evaluate the quality of the LLM-as-a-judge for feedback, two of the authors annotated 12 MAGIC–human feedback pairs for each of the five criteria (C1–C5) with a label of "LLM" or "Human". A third author then adjudicated the two annotators to determine the "ground truth" winner, selecting one of the two labels when the annotators disagreed. We then computed the Cohen's Kappa between the two annotators and between the adjudicator and judge as shown in 3. The average (across C1–C5) inter-annotator agreement was $\kappa_{\text{IAA}} = 0.578$ while the adjudicator–judge agreement was $\kappa_{\text{AJA}} = 0.139$.

Overall, our analysis of LLM-as-a-judge indicates that there was slight agreement between human-adjudicated preferences and LLM (o4-mini) preferences. This indicates that there is a bias for LLMs to prefer LLM produced feedback, which is not reflective of human preferences. With this result in mind, LLM-as-a-judge can still prove useful as a way to scale human preference data.

6 Conclusion

While ASAP has been valuable in developing techniques and methods for AES research, the utility of LLMs for AES or generative feedback for essays written at the college-level or above had not yet been proven until this work. To the best of our knowledge, our work is one of the first studies to apply holistic and multi-agent AES approaches to student-written essays at the college level and beyond by L1 and L2 learners.

By building and evaluating LLM AES and feedback systems on the GRE dataset, we have shown that LLMs are reliable college-level essay graders for L1 and L2 writers of English, and we can improve these open source model's scoring capabilities by using a multi-agent approach that does not require fine-tuning even across different prompts.

Furthermore, current knowledge in AES research presupposed that for a model to perform well on a new prompt, new knowledge specific to the new prompt needed to be introduced (Li and Ng, 2024a). Our evaluation on the GRE dataset used eight distinct essay prompts in the argumentative and persuasive essay genres, and we achieved substantial concordance or greater with human scoring across the dataset on both holistic and independent trait scoring.

MAGIC additionally provides added interpretability for both scoring and feedback with its multi-agent trait-based approach. MAGIC also provides a framework to generate and assess the quality of writing feedback. We show that we can improve the feedback generation capabilities of our models over the baseline, having each agent think about one specific aspect of the rubric and then consolidating this reasoning, instead of generating feedback holistically. Finally, not only do we assess the agreement of our orchestrator against the GRE holistic scores, but we also perform a per-agent assessment to highlight the reliability of our approach.

We demonstrated that small open-sourced and open-weight models are indeed usable for a combined task of simultaneous Automated Essay Scoring and Feedback generation, which are better suited for lower resourced schools and enterprises. MAGIC provides enhanced observability into how the model constructs the score and feedback, and the system appears to generate useful qualitative feedback that is comparable to human-level commentary.

Limitations

Although the corpus is high quality, ground truth evaluations are limited to less than 50 samples, along with copyright restrictions. The dataset has distribution constraints and limited utility for model training.

In addition, the essays have been produced under timed test conditions; a public corpora of graded essays that are longer with more complex prompts have yet to be compiled.

A further limitation of the current work is the focus on English-language argumentative essays. There are limited examples of scored essays with feedback for languages outside of English, but more work is planned to evaluate argumentative essays written in other languages.

Further investigation in using LLMs as judges for feedback quality is needed. It appears that prompt engineering alone may not be enough to align the judge along human preferences, despite the usage of advanced reasoning LLMs such as o4-mini.

Ethics Statement

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

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.

References

Ryan S. Baker. 2016. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26(2):600–614.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 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)*, pages 1638–1644.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013:i–15.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Semire Dikli. 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.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2025. Leveraging small LLMs for argument mining in education: Argument component identification, classification, and assessment.

Alex Franklin, Maggie, Meg Benner, Natalie Rambis, Perpetual Baffour, Ryan Holbrook, Scott Crossley, and ulrichboser. 2022. Feedback prize - english language learning. `https://kaggle.com/competitions/feedback-prize-english-language-learning`. Kaggle.

Gerald Graff. 2003. *Clueless in Academe: How Schooling Obscures the Life of the Mind*. Yale University Press.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, 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)*, pages 284–293.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 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.

Nafiseh Taghizadeh Kerman, Omid Noroozi, Seyyed Kazem Banihashem, Morteza Karami, and Harm J.A. Biemans and. 2024. Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments*, 32(2):614–626.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders?

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: A reflection on the state of the art. pages 17876–17888.

Shengjie Li and Vincent Ng. 2024b. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

Zhi Li, Stephanie Link, Hong Ma, Hyejin Yang, and Volker Hegelheimer. 2014. The role of automated writing evaluation holistic scores in the esl classroom. *System*, 44:66–78.

Pei Yee Liew and Ian K. T. Tan. On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, pages 204–211. ACM.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. 2017. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4):502–513.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Richard Menary. 2007. Writing as thinking. *Language Sciences*, 29(5):621–632. Cognitive Dynamics in Language.

Haile Misgna, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. 2025. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):1–40.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Lin S Norton and JCW Norton. 2001. Essay feedback: How can it help students improve their academic writing?.

Fan Pan, Randi Reppen, and Douglas Biber. 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.

Jessica Riddell. 2015. Performance, feedback, and revision: Metacognitive approaches to undergraduate essay writing. *Collected Essays on Learning and Teaching*, 8:79.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring1.

Kennedy Smith and Scott Crossley. 2025. Identifying limitations and bias in chatgpt essay scores: Insights from benchmark data. *The Cutting Ed*.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *arXiv preprint arXiv:2311.09517*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Neslihan Süzen, Alexander N. Gorban, Jeremy Levesley, and Evgeny M. Mirkes. 2020. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169:726–743.

UCLouvain. 2025. Learner corpora around the world. Accessed: 2025-05-15.

Jorge Villalon, Paul Kearney, Rafael Calvo, and Peter Reimann. 2008. Glosser: Enhanced feedback for student writing tasks. pages 454 – 458.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. a. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199. Association for Computational Linguistics.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. b. Tutor CoPilot: A human-AI approach for scaling real-time expertise.

Xinyu Wang, Yohan Lee, and Juneyoung Park. 2022. Automated evaluation for student argumentative writing: A survey. *arXiv preprint arXiv:2205.04083*.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

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

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 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)*, pages 576–584.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A Appendix A: Prompts

GRE baseline model system prompt

You are an expert professional grader who scores student essays tagged <student_essay> based on a rubric.
Please provide a numerical score for the provided essay according to the specified rubric.
- Provide an appropriate holistic score.
- You will carefully read the rubric (<rubric>), prompt (<essay_prompt>) and student essay (<student_essay>), as many times as needed.
- You will reason carefully as to why you chose this score following the rubric and guidelines.
- You will provide a detailed explanation of your reasoning for the score.
- You will provide feedback for the student on how to improve their essay.
- A low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays.

The rubric or rubrics for this essay is as follows:
<rubric>
{rubric}
</rubric>
The given task is as follows:
<task_directions>
{task_directions}
</task_directions>
The prompt is as follows:
<essay_prompt>
{prompt}
</essay_prompt>
Review the given rubric and prompt carefully and score the <student_essay>.
Provide a numerical score by using the provided rubric's guidance.
Remember, a low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays.
{output_format}

Table A1: **System prompt for our baseline model**

GRE orchestrator system prompt
You are an expert professional grader who scores student essays tagged <student_essay> based on other expert grader's scores and reasoning. Please provide a numerical score for the provided essay according to the opinions of the other expert grader's scores and reasoning. Each expert grader is an expert grader for a specific aspect of the essay. - The length of the essay matters, a well developed essay should have at least 3-4 well written paragraphs. - You will carefully read each expert grader's score and reasoning, prompt (<essay_prompt>) and student essay (<student_essay>), as many times as needed. - You will reason carefully as to why you chose this score balancing the opinions of the other expert grader's scores and reasoning. - You will provide a detailed explanation of your reasoning for the score. - You will provide feedback for the student on how to improve their essay, balancing the opinions of the other expert grader's feedback. - A low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. The expert grader's scores and reasoning are as follows: {expert_grader_scores_and_reasoning} The given task is as follows: <task_directions> {task_directions} </task_directions> The prompt is as follows: <essay_prompt> {prompt} </essay_prompt> Review the given expert grader's scores and reasoning, prompt and student essay carefully and score the <student_essay>. Provide an integer score between 0 and 6 by balancing the provided expert grader's scores and reasoning. Remember, a low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. {output_format}

Table A2: **System prompt for our orchestrator**

GRE argumentative agent system prompt
You are an expert professional grader who scores student essays tagged <student_essay> based on a rubric. You specialize in scoring the argumentative qualities of an essay. Please provide a numerical score for the provided essay considering all aspects of the specified rubric. - Provide an appropriate holistic argumentative score. - The length of the essay matters, a well developed essay should have at least 3-4 well written paragraphs. - You will carefully read the rubric (<argumentative_rubric>), prompt (<essay_prompt>) and student essay (<student_essay>), as many times as needed. - You will reason carefully as to why you chose this score following the rubric and guidelines. - You will provide a detailed explanation of your reasoning for the score. - You will provide feedback for the student on how to improve the argumentative qualities of their essay. - A low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. The rubric or rubrics for this essay is as follows: <argumentative_rubric> {argumentative_rubric} </argumentative_rubric> The given task is as follows: <task_directions> {task_directions} </task_directions> The prompt is as follows: <essay_prompt> {prompt} </essay_prompt> Review the given rubric and prompt carefully and score the <student_essay>. Provide a numerical score by using the provided rubric's guidance. The score should be a number between 0 and 6. Remember, a low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. {output_format}

Table A3: **System prompt for our argumentative agent**

GRE grammar agent system prompt
You are an expert professional grader who scores student essays tagged <student_essay> based on a rubric. You specialize in scoring the grammar and mechanics of an essay. Please provide a numerical score for the provided essay considering all aspects of the specified rubric. - Provide an appropriate holistic grammar score. - You will carefully read the rubric (<grammar_rubric>), prompt (<essay_prompt>) and student essay (<student_essay>), as many times as needed. - You will reason carefully as to why you chose this score following the rubric and guidelines. - You will provide a detailed explanation of your reasoning for the score. - You will provide feedback for the student on how to improve the grammar and mechanics of their essay. - A low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. The rubric or rubrics for this essay is as follows: <grammar_rubric> {grammar_rubric} </grammar_rubric> The given task is as follows: <task_directions> {task_directions} </task_directions> The prompt is as follows: <essay_prompt> {prompt} </essay_prompt> Review the given rubric and prompt carefully and score the <student_essay>. Provide a numerical score by using the provided rubric's guidance. The score should be a number between 0 and 6. Remember, a low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. {output_format}

Table A4: **System prompt for our grammar agent**

GRE vocabulary agent system prompt
You are an expert professional grader who scores student essays tagged <student_essay> based on a rubric. You specialize in scoring the vocabulary and sentence variety of an essay. Please provide a numerical score for the provided essay considering all aspects of the specified rubric. - Provide an appropriate holistic vocabulary score. - You will carefully read the rubric (<vocabulary_rubric>), prompt (<essay_prompt>) and student essay (<student_essay>), as many times as needed. - You will reason carefully as to why you chose this score following the rubric and guidelines. - You will provide a detailed explanation of your reasoning for the score. - You will provide feedback for the student on how to improve the vocabulary and sentence variety of their essay. - A low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. The rubric or rubrics for this essay is as follows: <vocabulary_rubric> {vocabulary_rubric} </vocabulary_rubric> The given task is as follows: <task_directions> {task_directions} </task_directions> The prompt is as follows: <essay_prompt> {prompt} </essay_prompt> Review the given rubric and prompt carefully and score the <student_essay>. Provide a numerical score by using the provided rubric's guidance. The score should be a number between 0 and 6. Remember, a low score isn't harmful to the student. Rather, an accurate match to the rubric will help the student improve their score in future essays. {output_format}

Table A5: **System prompt for our vocabulary agent**

GRE LLM judge prompt

You are an expert professional grader who specializes in evaluating feedback from expert graders.
You will be given two feedbacks for an essay crafted by two expert graders.
You will choose the better feedback (<feedback_1> or <feedback_2>) for each of the criteria specified in <criteria>.
<criteria>
- C1: Which feedback is more relevant to the essay content?
- C2: Which feedback is better at highlighting weakness?
- C3: Which feedback is better at highlighting strengths?
- C4: Which feedback is more specific and actionable?
- C5: Which feedback is overall more helpful for a student?
</criteria>
The rubric for the essay is as follows:
<rubric>
{rubric}
</rubric>
The two feedbacks are as follows:
<feedback_1>
{feedback_1}
</feedback_1>
<feedback_2>
{feedback_2}
</feedback_2>
The given task is as follows:
<task_directions>
{task_directions}
</task_directions>
The prompt is as follows:
<essay_prompt>
{prompt}
</essay_prompt>
The student essay is as follows:
<student_essay>
{student_essay}
</student_essay>
Provide a number (1 or 2) representing the feedback that you choose for each of the criteria.
{output_format}

Table A6: **System prompt for our LLM judge**

B Appendix B: GRE Dataset

Name	Published	# Essays
Practice General Test	2003	12
Practice General Test # 2	2011	12
Sample GRE ® Issue Task with Strategies	2022	6
Sample GRE ® Argument Task with Strategies	2022	6
Practice General Test # 1	2023	6
Practice General Test # 3	2023	6

Table B1: **GRE Essay Dataset Sources**

Ground Truth Essay Sample	**Ground Truth Feedback**
When the generation of today matures, it is important for them to succeed and become the successful leaders in government, industry and other fields. There are many traits that leaders must possess, and cooperation is one of these very important characters. Nonetheless it is important for leaders to have a sense of competition, so as to prevent themselves from being complacent with their position. [...] However, it is still important for there to be a sense of competition. Competition is the root of motivation for most. It drives us to become stronger, smarter, and to want more. Nonetheless, the spirit of competition must also be reigned in, and not be allowed to run wild. Competitiveness can lead to abuse of power and distasteful actions, which is quite the opposite of someone who displays cooperativeness. Some may argue that competition is not needed. That those that are meant to be leaders will not become complacent, because they have their own internal drive to lead. If there was no competition, there would be no world records. Michael Phelps may not be a leader of government or industry, but he is certainly educated on the technique of swimming, and leader in his field. Would he be as good as he is today if there was not competition? Would the leaders of Microsoft have been motivated to create Bing if there was no Google? Cooperation helped many leaders get where they are today, and will continue to do so in the future. But leaders, as well as those that aspire to be one, all need to have a sense of competition as well.	This adequate response presents a clear position on the issue in accordance with the assigned task, arguing that both competition and cooperation are important for leaders. The response uses counterarguments both in the construction of its overall position (comparing the value of both competition and cooperation) and in its discussion of the positive and negative aspects of competition. However, the development of ideas in this response is not as thorough or as persuasive as one would expect to see in a response that earns a score of 5 or 6. For instance, the example of Microsoft inventing Bing to compete with Google is certainly relevant, but it is not developed with any thoughtfulness. It is simply stated. Other examples are somewhat more fully developed, but there is also some tangential material (e.g., even the writer seems to understand that Michael Phelps does not quite fit into a discussion of leadership). In addition to its adequate development, this response displays adequate control of language. This response does not have the sentence variety or the skillful diction seen in a response that earns a higher score. There are some minor errors present, but nothing that interferes with clarity. Because this response presents a clear position on the issue, expressing meaning with adequate clarity, it earns a score of 4.

Table B2: **GRE dataset example essay–feedback pair**

C Appendix C: MAGIC Outputs

MAGIC	Baseline
The essay effectively challenges the argument's reasoning, identifying the flaw of equating correlation with causation. As noted by the expert graders, the essay successfully points out that the absence of protective gear doesn't necessarily mean it would have prevented injury. The discussion of alternative factors—training, health, common sense—is valuable, though it could be expanded upon as suggested by the expert graders. Here's a more detailed breakdown: **Strengths:** - **Clear Identification of Flaw:** The essay immediately recognizes the logical fallacy in assuming that lack of gear leads to injury. - **Alternative Factors:** Bringing up training, health, and common sense demonstrates a good understanding of the complexity of the situation. - **Grammar and Mechanics:** The writing is mostly clear and grammatically sound, as highlighted by the final expert grader. **Areas for Improvement (building on expert feedback):** - **Expand on Alternatives:** Don't just mention training, health, and common sense; explain how they mitigate risks. Provide concrete examples. What specific skills learned in training reduce injury likelihood? How does physical fitness play a role? - **Specificity Regarding Injuries:** Differentiate between minor and severe injuries. The original argument implies a link between lack of gear and serious harm. Focus your critique on that specific connection. - **Stronger Conclusion:** The current conclusion ("statistics... must be further examined") is weak. Reiterate your main point—that the argument oversimplifies the issue and relies on faulty assumptions. Offer a more definitive statement about the limitations of the statistics. - **Refine Sentence Structure:** While the grammar is generally correct, some sentences are overly long and convoluted. Breaking them down into shorter, clearer units would improve readability.	This is a well-reasoned critique of the argument presented. The essay effectively identifies the vagueness of the statistics and points out that correlation doesn't equal causation. The discussion about various potential injuries, both preventable and unpreventable, strengthens the argument against the simplistic conclusion drawn by the original author. However, while the essay raises valid concerns, it could benefit from slightly more precise language and a stronger concluding statement that directly summarizes the flaws in the initial argument. For example, explicitly stating that the argument commits a fallacy of hasty generalization would elevate the analysis. Overall, the essay demonstrates a strong understanding of critical thinking and argumentation.

Table C1: **MAGIC vs. baseline feedback example**