From Advocacy to Judgment: Training-Free Analytic Essay Scoring with Multi-Agent Debate and Exemplar Retrieval

Anonymous ACL submission

Abstract

Automated Essay Scoring (AES) is shifting from feature-engineering to LLMs, yet current training-free approaches struggle with calibration, often exhibiting a "middle-score bias" that fails to distinguish between exceptional and weak writings. In this work, we introduce MADRAG (Multi-Agent Debate with Retrieval-Augmented Generation), a training-free framework designed to achieve the reliability of supervised models without the need for labeled training data. MADRAG decomposes the scoring process into a multi-agent interaction: an Advocate highlights essay strengths, a Skeptic critiques weaknesses, and a Judge synthesizes these arguments to assign a score. Crucially, we augment the Judge with RAG mechanism that retrieves rubric-aligned exemplar essays spanning the full score range, grounding the debate in concrete evidence. Evaluating our approach on the ASAP dataset for analytic trait scoring, we demonstrate that MADRAG significantly outperforms existing prompt-based LLM baselines and achieves performance competitive with state-of-the-art supervised models.

1 Introduction

Assessing student writing is labour-intensive and often inconsistent across raters. In large educational settings, teachers must score many essays under tight time constraints, leading to fatigue, delayed feedback, and imperfect reliability (Ramesh and Sanampudi, 2021). Even when essays are double-scored, inter-rater agreement remains limited: analyses of the Automated Student Assessment Prize (ASAP) dataset show that trained raters frequently disagree by more than one score point on individual traits (Crossley et al., 2025). These challenges motivate AES systems that aim to approximate human judgments at scale.

Early AES approaches relied on hand-crafted features such as word counts and readability metrics (Page, 1966; Attali and Burstein, 2006), followed by neural models including recurrent, convolutional, and transformer-based architectures (Taghipour and Ng, 2016; Dong et al., 2017; Wang et al., 2022). While these systems can produce reliable scores, most deployed tools output a single holistic score, offering little actionable feedback for instruction (Warschauer and Ware, 2006). In contrast, writing teachers typically prefer analytic trait scoring, which provides targeted feedback on dimensions such as ideas, organization, and conventions (Knoch, 2009). An ideal AES system should therefore produce accurate, transparent, and reliable *trait-level* scores.

Large language models (LLMs) have recently enabled training-free, prompt-based scoring across diverse rubrics (Fallah et al., 2024). However, prior work shows that direct LLM judging is often poorly calibrated, sensitive to prompt design, and prone to systematic biases (Mansour et al., 2024). In particular, LLM-as-judge tends to regress toward the middle of the scoring scale, failing to distinguish clearly between exceptional and weak inputs (Zheng et al., 2023). These issues are exacerbated by the multi-step, multi-trait nature of rubric-based scoring, which requires maintaining and coordinating multiple criteria and score ranges within a single judgment (Valmeekam et al., 2023). As a result, naïvely applying LLMs as essay graders yields inconsistent trait scores and misjudgments of extreme cases.

In this work, we ask whether a fully *training-free* LLM-based system can achieve trait-level scoring reliability comparable to human raters and strong supervised AES models. We propose **MADRAG** (Multi-Agent Debate with Retrieval-Augmented Generation), a framework that combines two complementary mechanisms. First, multi-agent debate (MAD): an Advocate highlights strengths, a Skeptic critiques weaknesses, and a Judge synthesizes their arguments to produce a score. Second, retrieval-augmented generation (RAG): before scor-

ing, the Judge retrieves rubric-aligned exemplar essays spanning the full score range to ground and calibrate its decision. By integrating debate with trait-specific retrieval, MADRAG provides external memory, encourages explicit comparison against exemplars, and mitigates middle-score bias.

2 Related Work

2.1 Supervised Analytic Trait Scoring

Early AES systems primarily focused on holistic scoring using hand-crafted features such as word counts and readability measures (Page, 1966; Attali and Burstein, 2006). While effective for large-scale testing, holistic scores provide limited diagnostic value for formative assessment, motivating a shift toward analytic trait scoring that evaluates dimensions such as content, style, and organization separately (Warschauer and Ware, 2006; Deane, 2013).

Neural approaches enabled this transition by modeling essays as structured representations amenable to multi-trait prediction. Mathias and Bhattacharyya (2020) pioneered attention-based architectures for analytic scoring, showing that trait-specific predictions better support instructional feedback. Subsequent work has focused on improving robustness in data-scarce and cross-prompt settings. ProTACT (Do et al., 2023), for example, introduces prompt-aware representations and a trait-similarity objective to exploit correlations among rubric dimensions, achieving strong performance on the ASAP dataset. Others explore multi-task transformers with trait-specific heads (Kumar et al., 2022) or reinforcement learning objectives to refine trait-level accuracy (Do et al., 2024). Despite their reliability, supervised trait scorers require labeled data for each new prompt and rubric, limiting their practicality in real-world classrooms where assignments and criteria change frequently.

2.2 Training-Free LLM Scoring

LLMs offer an appealing alternative by enabling training-free, prompt-based essay scoring that can be applied across prompts and rubrics without task-specific fine-tuning (Kojima et al., 2022). However, empirical studies consistently find that direct zero-shot scoring lags behind supervised AES systems in both accuracy and reliability. Mansour et al. (2024) show that even with careful prompt engineering and one-shot examples, models such as ChatGPT and LLaMA substantially underperform supervised baselines.

More structured prompting strategies improve performance but expose additional limitations. Tang et al. (2024) demonstrate that rubric-aligned exemplars and justification prompts can raise agreement on abstract traits like *Ideas*, yet performance remains highly sensitive to decoding parameters and degrades sharply on surface-level traits such as *Conventions*. Related analyses reveal systematic biases, including harshness on complex traits (Kundu and Barbosa, 2024), length bias, and central tendency bias, where models avoid extreme scores and regress toward the middle of the scale (Li et al., 2025). One line of work addresses calibration by reformulating scoring as a comparative task. LCES (Shibata and Miyamura, 2025) replaces absolute scoring with pairwise ranking, yielding more stable judgments. However, its quadratic complexity makes it difficult to scale to classroom-sized datasets, highlighting the need for training-free approaches that retain absolute scoring while improving calibration and extreme-score discrimination.

2.3 Multi-Agent Debate and Orchestration

To mitigate the limitations of single LLM judges, recent work has explored multi-agent frameworks in which multiple models collaborate or debate to reach a consensus. Debate has been shown to improve reasoning quality and factual accuracy by encouraging agents to critique and refine each other's arguments (Liang et al., 2024; Du et al., 2024). ChatEval (Chan et al., 2023) successfully applied this to text evaluation, showing that a "jury" of LLMs correlates better with human judgments than a single score. In the domain of AES, MAGIC (Jordán et al., 2025) applies this by assigning specialized agents to different rubric traits (e.g., a "Grammar Expert" and "Organization Expert") and synthesizing their outputs via an orchestrator, achieving substantial gains over single-agent baselines. Similarly, CAFES (Su et al., 2025) utilizes a reflective workflow where an initial scorer revises its judgments based on feedback from a "critic" agent. These systems underscore the potential of decomposing the scoring task, yet they often rely on static agent roles or lack access to external knowledge, which can limit their ability to ground scores in concrete evidence.

2.4 Positioning of Our Work

Taken together, prior work reveals a persistent trade-off: supervised AES models deliver reliable trait-level scores but lack flexibility, while training-

free LLM judges offer adaptability at the cost of calibration and consistency. Multi-agent systems partially bridge this gap, yet they remain vulnerable to groupthink and poorly grounded reasoning (Wu et al., 2023). MADRAG advances this line of research for analytic trait scoring. Unlike prior multi-agent AES systems that rely solely on internal model representations, MADRAG retrieves and conditions on rubric-aligned exemplar essays spanning the full score range during judgment. By combining adversarial reasoning with explicit evidence retrieval, our approach aims to reduce middle-score bias, improve extreme-score discrimination, and produce training-free trait scores that are both calibrated and grounded.

Research Questions. Guided by the limitations of both supervised AES systems and LLMs as judges, we structure our empirical evaluation around the following research questions:

- **RQ1:** Can MADRAG achieve competitive *training-free* analytic trait scoring performance relative to strong supervised AES models and prior training-free LLM baselines?
- **RQ2:** Does MADRAG mitigate the *middle-score bias* commonly observed in LLM-based judges and improve discrimination at the extremes of the scoring scale?
- **RQ3:** When MADRAG fails, what failure mechanisms dominate, and which components of the system plausibly contribute to these errors?

3 Methodology

3.1 Problem Setting and Notation

Let $\mathcal{E}$ denote the collection of student essays and $\mathcal{R}$ the set of rubric traits. Each essay $e \in \mathcal{E}$ consists of unstructured text and associated metadata (e.g., an identifier). Each rubric trait $r \in \mathcal{R}$ is represented by a structured object with fields including a name, minimum and maximum scores, and a description of the trait to evaluate. For a given essay e and trait r, our goal is to produce a numeric score $s(e, r)$ reflecting how well the essay satisfies the trait, along with a rationale.

3.2 Agents and Roles

We assign specialized roles to agents (Appendix A.6): the Advocate speaks first, the Skeptic responds, and the Judge synthesizes the contributions along with agent confidence scores. The confidence of each agent is approximated by the log-probability of the first token in its message—a coarse indicator of internal certainty inspired by confidence-aware debate systems (Lin and Hooi, 2025; Kadavath et al., 2022).

Supervisor. A coordinator that decomposes the rubric into its constituent traits, retrieves few-shot examples for each trait, instantiates debate agents and judges, and orchestrates the entire evaluation.

Advocate agent. Given an essay and trait, this agent argues in favor of the essay's performance on the trait. It receives the rubric description, but not the few-shot examples. The Advocate produces a single initial statement highlighting strengths; it does *not* assign a score or mention weaknesses.

Skeptic agent. This agent critically examines the essay with respect to the trait and points out shortcomings. It takes as input the Advocate's initial statement and the rubric and produces a single rebuttal. The Skeptic is prohibited from assigning a score or mentioning strengths.

Judge agent. An impartial arbiter that reads both the Advocate and Skeptic messages, along with their confidences, few-shot examples, and the rubric trait, and produces a final integer score.

3.3 Retrieval-Augmented Few-Shot Example Generation

In our setting, retrieval is used to provide previously scored essays as few shot calibration references (Appendix A.7), allowing judges to align its scoring decisions with examples that span the rubric's scoring range (Lewis et al., 2021).

3.3.1 Vector Database and Embeddings

To support retrieval augmented exemplar construction, we construct a vector database of scored essays. Each essay $e \in \mathcal{E}$ is embedded as a dense vector representation using all-MiniLM-L6-v2, a sentence-transformer embedding model $\phi(\cdot)$ (Reimers and Gurevych, 2019). It maps the full essay text into a fixed-dimensional vector $\mathbf{z} = \phi(e)$ optimized for semantic similarity search. These embeddings are stored in a Chroma vector database together with structured metadata. Formally, the vector database is defined as

$$\mathcal{V} = \{(\mathbf{z}, m) \mid \mathbf{z} = \phi(e),\ e \in \mathcal{E}\},$$

m is a metadata dictionary associated with essay e which contains the raw essay text, the overall

domain score, and a discrete score for each rubric trait. When multiple human raters provide scores for a given trait, their scores are aggregated by averaging and rounding to the nearest integer.

3.3.2 Retrieval Procedures

Given a query essay $e \in \mathcal{E}$, retrieval is performed by embedding the essay using the same encoder $\phi(\cdot)$ and searching the vector database $\mathcal{V}$ for relevant few-shot exemplars F. We define two retrieval procedures. The first procedure retrieves essays that are semantically similar to the query essay. Specifically, the query essay e is embedded as $\phi(e)$, and a nearest-neighbor search is performed in $\mathcal{V}$ to retrieve the top k essays with highest similarity in the embedding space which are used as few-shot examples. The second procedure retrieves exemplars conditioned on rubric scores for a specific trait. For a given trait $r \in \mathcal{R}$, we retrieve one exemplar essay for each score value within the valid score range. For each s, we perform nearest-neighbor search in $\mathcal{V}$ using the query embedding $\phi(e)$ while filtering candidates to essays whose metadata score for trait r equals s. When multiple candidates are available, the nearest remaining essay is selected. If no suitable essay exists for a given score value, the system records that no exemplar is available for that score.

3.4 MADRAG Workflow

Figure 1 provides an overview of MADRAG and shows how trait-wise debate is combined with retrieval-augmented exemplars within a single scoring pipeline. We now formalize the trait-level procedure, where retrieved exemplars augment the Judge alongside the Advocate–Skeptic debate transcript. Algorithm 1 summarizes the full trait-level procedure, including retrieval, debate, and confidence-aware judgment. For reproducibility, Appendix A documents the exact execution logic.

Debate dynamics. For essay e and rubric trait r_i with score range $[m_i, M_i]$, the shared input context:

$$x_i(e) \;=\; \langle q,\; e,\; r_i,\; [m_i, M_i],\; \alpha \rangle$$

where q denotes the essay prompt, and α collects fixed inference-time settings shared across agents (e.g., system instructions/role prompts and decoding hyperparameters such as temperature).

Advocate generation. Let G_A denote the Advocate policy (LLM with role prompt). The Advocate emits an argument a_i:

$$a_i \sim p_A(\cdot \mid x_i(e)).$$

Algorithm 1 Evaluate a rubric trait via MADRAG

function EVALUATETRAIT(e, r_i, q, α)
  $x \leftarrow \langle q,\ e,\ r_i,\ [m_i, M_i],\ \alpha \rangle$
  $F \leftarrow \mathrm{RAG}(e, r_i)$
  $(a, \ell^A) \leftarrow$ ADVOCATE(x) $\{a \sim p_A\}$
  $c^A \leftarrow \exp(\ell^A)$
  $(k, \ell^K) \leftarrow$ SKEPTIC(x, a) $\{k \sim p_K\}$
  $c^K \leftarrow \exp(\ell^K)$
  $\tau \leftarrow (a, k)$ $\{\tau \equiv \tau_i(e)\}$
  $\hat{s} \leftarrow$ JUDGE(x, F, τ, c^A, c^K)
  return $(\hat{s}, a, c^A, k, c^K)$
end function=0

We define an *internal confidence proxy* via the log-probability of the first emitted token $t^A_{i,1}$:

$$\ell^A_i \;=\; \log p_A\big(t^A_{i,1} \mid x_i(e)\big)\,,\; c^A_i \;=\; \exp(\ell^A_i).$$

Skeptic rebuttal. The Skeptic policy G_K conditions on the Advocate's message:

$$k_i \sim p_K(\cdot \mid x_i(e), a_i),$$

with first-token confidence

$$\ell^K_i \;=\; \log p_K\big(t^K_{i,1} \mid x_i(e), a_i\big)\,,\; c^K_i \;=\; \exp(\ell^K_i).$$

Debate transcript. The debate trace for trait r_i is the ordered pair

$$\tau_i(e) \;=\; (a_i,\; k_i).$$

Confidence-aware judging. The judge J produces a discrete score by maximizing a confidence-conditioned posterior over valid integers:

$$\mathcal{I}_i(e) \;=\; \big(x_i(e),\, \mathrm{RAG}(e, r_i),\, \tau_i(e),\, c^A_i,\, c^K_i\big),$$
$$s_i(e) = \arg\max_s \; p_J\Big(s \;\Big|\; \mathcal{I}_i(e)\Big)$$

4 Experiments

4.1 Experimental Setup

Data. We evaluate MADRAG on the ASAP [1] dataset, a widely used benchmark of student-written English essays scored by trained human raters using prompt-specific rubrics. ASAP consists of eight essay sets, but analytic trait annotations in the original release are available only for Essay Sets 7 and 8. Accordingly, all experiments in this paper are conducted on Sets 7 and 8, which provide multiple independent human ratings per essay at the trait level. Detailed dataset statistics, prompts, transcription procedures, and label construction are provided in Appendix B.

[1] https://www.kaggle.com/c/asap-aes/data

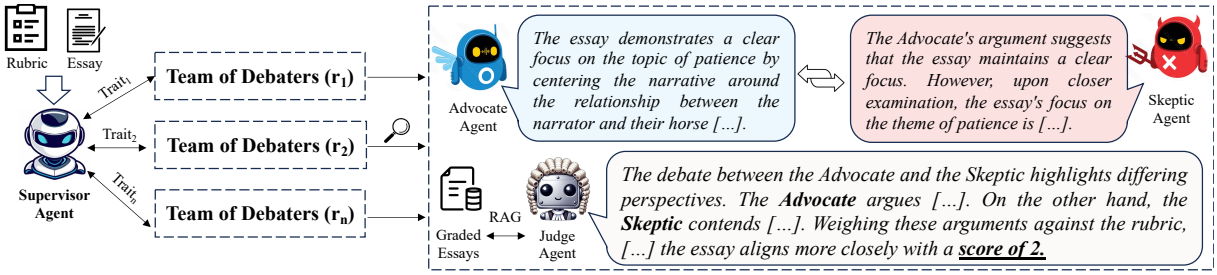

Figure 1: **Overview of the MADRAG scoring pipeline** The Supervisor routes each rubric trait to a dedicated debate team, retrieves few-shot exemplars to augment the Judge, and the Judge aggregates the Advocate and Skeptic exchanges together to produce the final trait score. The full sequence of agent messages is provided in Appendix A.8.

Evaluation Metrics. We evaluate trait-level scoring using Quadratic Weighted Kappa (QWK), a standard agreement metric in AES that accounts for ordinal score distances.

LLMs. MADRAG is evaluated with multiple LLM backbones (GPT-4o-mini, GPT-4o, GPT-5-mini, and GPT-5) using the same role prompts across models (OpenAI, 2024). To reduce run-to-run variance, we decode the Judge deterministically (temperature $= 0$) and use high-temperature decoding for the Advocate and Skeptic; when supported, we also log token-level log-probabilities to compute confidence proxies (Appendix A). During scoring, the Judge is augmented with retrieved, rubric-aligned exemplars spanning the trait's score range, following the retrieval procedure.

Baselines. We compare MADRAG against a diverse set of strong baselines spanning supervised and training-free paradigms. These include *training-based* neural models for analytic trait scoring—FeatEng-RF (Mathias and Bhattacharyya, 2020), and ProTACT (Do et al., 2023)—as well as *training-free*, prompt-engineered LLM scorers, including ZS-LLM (Mansour et al., 2024) and CSR-J (Tang et al., 2024). We additionally report Human–Human agreement as a reference ceiling. To improve readability, we defer detailed descriptions of each baseline's methodology, training regime, and evaluation protocol to Appendix C.

4.2 Comparative Performance on Analytic Traits (RQ1)

Table 1 reports trait-wise QWK on ASAP Essay Sets 7 and 8, comparing MADRAG against supervised and training-free baselines. Three main findings emerge. First, MADRAG substantially outperforms all prior training-free approaches. Second, despite requiring no labeled training data, MADRAG achieves performance competitive with—and in several cases exceeding—strong supervised systems. Third, gains are highly trait-dependent, with the largest improvements observed on discourse-oriented traits.

Across both essay sets, MADRAG consistently improves over training-free LLM baselines. The strongest configuration (with GPT-5) achieves a QWK of 0.75 on *Ideas* in Set 7, representing a 36% relative improvement over the CSR-J (0.55) and a near sevenfold increase over ZS-LLM (0.05–0.09). Similar margins are observed for *Organization* and *Style*, where MADRAG variants consistently rank among the top performers. These gains demonstrate that multi-agent debate and exemplar-grounded judging substantially improve calibration beyond prompt engineering alone. More strikingly, MADRAG is competitive with supervised models that rely on thousands of labeled training examples. On *Ideas*, MADRAG (GPT-5) exceeds ProTACT by 50% in Set 7 (0.75 vs. 0.50) and by 18% in Set 8 (0.67 vs. 0.57). It also surpasses Human–Human agreement on *Ideas* in both sets (0.75 vs. 0.69 in Set 7; 0.67 vs. 0.53 in Set 8), indicating that the framework yields more consistent content judgments than individual human raters. While the strongest supervised baseline (FeatEng-RF) remains dominant on several traits, MADRAG's performance without any task-specific training shows that structured reasoning over retrieved exemplars can approximate learned scoring functions.

Performance varies systematically by trait type. Discourse-oriented traits such as *Ideas* and *Organization* exhibit the largest gains, with MADRAG often matching or outperforming supervised systems. In contrast, surface-level traits show more modest improvements and greater variance across LLM backbones. For example, on *Conventions* in Set 7, all MADRAG variants trail FeatEng-RF by a wide margin (0.19–0.28 vs. 0.62), and in Set 8, performance on *Word Choice* and *Sentence Fluency*

remains inconsistent. The gap likely reflects both the debate structure, which emphasizes holistic argumentation over error counting, and the limited utility of exemplar retrieval for traits where score distinctions hinge on surface-level errors. At the same time, persistent gaps on surface traits suggest that fully replacing supervised models will require hybrid approaches that combine debate-based reasoning with specialized mechanisms for low-level linguistic analysis.

4.2.1 Ablation Study

To isolate the contribution of each component in MADRAG, we conduct an ablation study on the merged ASAP Sets 7 and 8, averaging results for overlapping traits. We compare the full model against five variants: **SA** (Single-Agent): a single LLM Judge scores all traits directly from the rubric; **SARAG** (Single-Agent+RAG): the same single Judge is additionally provided with retrieved exemplars; **MA** (Multi-Agent Decomposition): one independent Judge per trait scores that trait from the rubric (decomposition only); **MAD** (Multi-Agent Debate): an Advocate and Skeptic debate each trait and a Judge synthesizes their exchange; and **MARAG** (Multi-Agent+RAG): one Judge per trait receives retrieved exemplars but no debate transcript (retrieval without debate). As shown in Figure 2, performance improves incrementally from **SA** as we add decomposition (**MA**), debate (**MAD**), and retrieval (**MARAG**), with the full MADRAG configuration achieving the best overall performance, particularly on discourse traits such as *Organization*. However, on some surface traits (e.g., *Conventions*), **MARAG** can occasionally outperform MADRAG, suggesting that adversarial debate may introduce noise for fine-grained, error-based scoring. We investigate this trade-off further in RQ3 via targeted failure-mode analysis. A detailed, set-wise breakdown of the ablation results is provided in Appendix D.

4.3 Mitigating Middle-Score Bias (RQ2)

LLM-based judges exhibit middle-score bias, clustering predictions toward the center of the scoring scale and avoiding extreme values even when warranted (Zheng et al., 2023; Li et al., 2025). This is particularly problematic for formative assessment, where accurate identification of struggling and exceptional students matters most. We test whether MADRAG mitigates this bias on essay–trait instances where at least one rater assigned either the minimum or maximum score. For each instance, we report **Agree@1** (agreement within ± 1 point of the extreme score) and **MAE** (mean absolute error from the extreme reference).

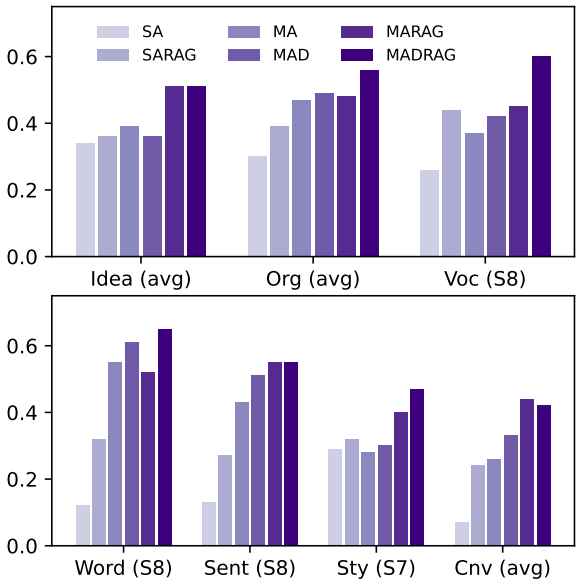

Figure 2: Merged ablation results (QWK). Overlapping traits between ASAP Sets 7 and 8 are averaged.

Table 2 shows that MADRAG consistently achieves the highest agreement and lowest error across both essay sets. Several ablation patterns clarify the sources of these gains. Decomposition alone (MA) yields limited improvement and even degrades performance in Set 8, indicating that naive task division without grounding is insufficient for extreme-score discrimination. Adding retrieval produces substantial gains: MARAG attains 75.3% (Set 7) and 85.5% (Set 8) Agree@1, capturing most of MADRAG's improvement. This confirms that access to score-calibrated exemplars is the primary driver of extreme-score calibration. The inclusion of SARAG further isolates this effect. Compared to SA, SARAG improves agreement in both sets, showing that retrieval alone already mitigates central tendency to some extent. However, SARAG consistently underperforms MARAG, showing that retrieval without trait-wise decomposition is less effective at resolving boundary cases.

Debate without retrieval (MAD) improves over SA and MA but substantially trails retrieval-based variants, reinforcing that debate alone does not correct central tendency. Full MADRAG adds a further 4–5 percentage points in Agree@1 over MARAG in Set 7 and yields the highest agreement overall, indicating that debate provides discriminative refinement on top of retrieval. In Set 8, MARAG achieves slightly lower MAE than

Table 1: Trait-wise QWK. **Bold** indicates the best-performing result, and underline indicates the second-best result.

Set	Type	Method	Idea	Org.	Voc.	Word	Sent	Sty.	Cnv.
7	Train-free	MADRAG (GPT-4o-mini)	0.43	0.64	—	—	—	0.47	0.26
7	Train-free	MADRAG (GPT-4o)	0.40	0.38	—	—	—	0.35	0.28
7	Train-free	MADRAG (GPT-5-mini)	0.69	0.62	—	—	—	0.33	0.22
7	Train-free	MADRAG (GPT-5)	0.75	0.63	—	—	—	0.47	0.19
7	Train	ProTACT (Do et al., 2023)	0.50	0.31	—	—	—	—	0.23
7	Train	FeatEng-RF (Mathias and Bhattacharyya, 2020)	**0.77**	**0.67**	—	—	—	**0.65**	**0.62**
7	Train-free	CSR-J (GPT-4) (Tang et al., 2024)	0.55	0.58	—	—	—	0.47	0.22
7	Train-free	ZS-LLM (GPT-3.5-turbo) (Mansour et al., 2024)	0.05	0.07	—	—	—	0.08	0.1
7	Train-free	ZS-LLM (LLaMA-2-13B-Chat) (Mansour et al., 2024)	0.09	0.02	—	—	—	0.15	0.32
7	—	Human–Human	0.69	0.58	—	—	—	0.54	0.57
8	Train-free	MADRAG (GPT-4o-mini)	0.59	0.47	0.60	**0.65**	0.55	—	0.58
8	Train-free	MADRAG (GPT-4o)	0.59	0.42	**0.63**	0.61	**0.59**	—	**0.62**
8	Train-free	MADRAG (GPT-5-mini)	0.60	**0.63**	0.62	0.28	0.34	—	0.36
8	Train-free	MADRAG (GPT-5)	**0.67**	0.61	0.55	0.32	0.35	—	0.42
8	Train	ProTACT (Do et al., 2023)	0.57	0.61	—	0.59	0.55	—	0.43
8	Train	FeatEng-RF (Mathias and Bhattacharyya, 2020)	0.58	**0.63**	0.54	0.55	0.58	—	0.55
8	Train-free	ZS-LLM (GPT-3.5-turbo) (Mansour et al., 2024)	0.18	0.25	0.15	0.15	0.20	—	0.31
8	Train-free	ZS-LLM (LLaMA-2-13B-Chat) (Mansour et al., 2024)	0.27	0.27	0.27	0.26	0.12	—	0.08
8	—	Human–Human	0.53	0.54	0.47	0.48	0.51	—	0.55

MADRAG despite lower agreement, suggesting that debate can occasionally introduce small deviations around the extreme boundary even while improving exact matches.

Across both sets, MADRAG's MAE of approximately 1.0 indicates that residual errors typically land within one score point of the extreme reference, whereas SA and MA exhibit MAE values above 1.5, reflecting systematic regression toward the center. Trait-level analysis shows the largest gains on discourse-oriented traits (e.g., *Ideas*: 82% vs. 39% for SA), while surface traits show more modest improvements (e.g., *Conventions*: 76% vs. 54%), consistent with the overall performance trends in RQ1. Together, these results indicate that mitigating middle-score bias in training-free AES requires explicit calibration mechanisms while debate provides targeted but secondary refinement.

4.4 Qualitative Error Analysis on High-Disagreement Cases (RQ3)

While QWK summarizes overall agreement with human raters, it does not explain *why* MADRAG succeeds or fails. To characterize failure modes of judge's *reasoning*, we analyze a targeted subset of **high-disagreement** essay–trait instances where human raters are relatively consistent (within 1 point) but MADRAG deviates by more than 1 point from the human average. Each unit of analysis is one row (e, t) (essay e and trait t), containing the essay text, the rubric trait, two human scores, and the outputs (score + rationale) of five systems: SA, SARAG, MAD, MARAG, and MADRAG. We focus on rows where MADRAG is wrong (N=173) and diagnose *how* its rationale becomes misleading.

Table 2: Standout subset performance on ASAP Essay Sets 7 and 8. N denotes the number of essay–traits.

Set	Method	N	Agree@1	MAE
7	MADRAG	2,474	**0.795**	**1.001**
7	MARAG	2,474	0.753	1.143
7	MAD	2,474	0.575	1.380
7	MA	2,474	0.424	1.561
7	SARAG	2,474	0.566	1.306
7	SA	2,474	0.477	1.518
8	MADRAG	67	**0.883**	**1.011**
8	MARAG	67	0.855	0.834
8	MAD	67	0.638	1.264
8	MA	67	0.759	1.230
8	SARAG	67	0.608	1.386
8	SA	67	0.520	1.389

Coding scheme and procedure. We annotate each wrong case with three categorical codes. (i) **Reasoning quality** B labels whether MADRAG is rubric-aligned and text-grounded (B2), partially grounded (B1), or misaligned/ungrounded (B0). (ii) **Primary failure mechanism** C assigns one dominant mechanism: trait-boundary confusion (C1), debate framing capture (C2), spurious debate claim accepted (C3), exemplar-induced calibration error (C4), rubric boilerplate collapse (C5), or anonymization distortion (C6). (iii) **Component attribution** D compares the same row across abla-

Table 3: Marginal distributions of primary failure mechanisms (C) and component attributions (D).

Primary failure mechanism (C)	**Count**	**%**
C2 Debate framing capture	56	32.7
C6 Anonymization distortion	46	26.9
C3 Spurious debate claim accepted	30	17.5
C5 Rubric boilerplate collapse	14	8.2
C1 Trait-boundary confusion	13	7.6
C4 Exemplar-induced calibration error	12	7.0
Component attribution (D)	**Count**	**%**
D1 Debate plausibly contributed	64	37.4
D3 Interaction plausible (debate+RAG)	60	35.1
D4 Not component-specific	32	18.7
D2 Retrieval plausibly contributed	15	8.8

Table 4: $A \times C$ on wrong MADRAG cases.

Counts						
	C1	C2	C3	C4	C5	C6
A1	12	50	23	6	9	46
A2	1	6	7	6	5	0
Row-normalized (%)						
A1	8.2	34.2	15.8	4.1	6.2	31.5
A2	4.0	24.0	28.0	24.0	20.0	0.0

Table 5: Reasoning quality (B) distribution.

B	**Count**	**%**	**Trait**	**N**	**B0**	**B1**	**B2**
B0	10	5.8	Conventions	49	6	41	2
B1	149	86.6	Organization	45	2	42	1
B2	13	7.6	Ideas/Content	22	0	19	3
			Sent Fluency	33	1	28	4
			Voice	13	1	10	2
			Word Choice	10	0	9	1

tions to test whether the *same* failure mechanism persists when a component is removed: debate plausibly contributed (D1), retrieval plausibly contributed (D2), interaction plausible (D3), or not component-specific (D4). Annotators were explicitly instructed to treat anonymization markers (e.g., `@PERSON`, `@DATE`) as placeholders, not true errors.

Dominant failure mechanisms and their component sources. Table 3 summarizes the marginal distributions of failure mechanisms (C) and component attributions (D) across all incorrect cases. The most frequent mechanism is **debate framing capture** (C2; 32.7%), followed by **anonymization distortion** (C6; 26.9%) and **spurious claim acceptance** (C3; 17.5%). Template-like failures occur less often but remain non-trivial (C4: 7.0%; C5: 8.2%). Attribution analysis indicates that errors are most often linked to **debate dynamics** (D1; 37.4%) or to **debate–retrieval interactions** (D3; 35.1%), whereas retrieval alone is comparatively rare as the primary driver (D2; 8.8%). Appendix E.1 details how specific mechanisms align with individual components, and Appendix E.3 provides deeper qualitative analyses of themes and micro-theories.

Error direction: underscoring dominates. We next examine whether the same mechanisms govern *under-scoring* vs. *over-scoring*. Table 4 shows that most wrong cases correspond to **under-scoring** relative to the human average (A1: 146/173), and that anonymization distortion (C6) is exclusively an under-scoring mechanism in our sample. Over-scoring cases (A2: 27/173) are comparatively more associated with spurious claim acceptance (C3) and template-driven failures (C4/C5), consistent with plausible-sounding but weakly grounded rationales inflating rubric placement in the absence of careful verification.

Errors are usually partially grounded. Table 5 summarizes reasoning quality for wrong cases. Most errors are **partially grounded** (B1; 86.6%): rationales often sound rubric-consistent but fail to cite decisive text evidence or make an explicit evidence→rubric→score link. Fully grounded rationales (B2; 7.6%) are rare by construction in the wrong subset, while misaligned/ungrounded rationales (B0; 5.8%) concentrate in *Conventions* (12.2% B0 within that trait), consistent with fragile surface-form judgments. Appendix E.2 further analyzes how reasoning quality interacts with specific failure mechanisms.

Conclusion

We presented MADRAG, a fully training-free framework for analytic essay trait scoring. Our experiments show that MADRAG significantly outperforms existing LLM judges and achieves parity with state-of-the-art supervised models. By grounding scores in both adversarial reasoning and rubric-aligned exemplars, MADRAG produces calibrated, interpretable trait-level assessments without task-specific training. The success of MADRAG underscores the importance of explicit calibration mechanisms and structured deliberation in LLM-based evaluation. Ultimately, MADRAG illustrates how hybrid LLM frameworks can combine the flexibility of prompt-based scoring with the reliability of supervised systems, paving the way for more accountable and scalable automated assessment.

Limitations

While MADRAG demonstrates consistent gains over single-agent prompting and training-free LLM judging baselines, several limitations remain. First, our experiments are limited to ASAP Essay Sets 7 and 8, covering only two narrative prompts from middle- and high-school settings; as a result, our conclusions may not fully transfer to other genres (e.g., argumentative or expository writing), grade levels, languages, or rubric structures that appear in real educational deployments. Second, the ASAP essays contain anonymization placeholders (e.g., `@PERSON`, `@DATE`) that can be misread as genuine grammatical or mechanical errors, particularly for surface-level traits such as Conventions, introducing bias that is unrelated to true writing quality and potentially distorting trait-specific scores.

In addition, MADRAG is more computationally demanding than single-agent prompting or supervised AES systems at inference time, since each trait evaluation requires multiple LLM calls (Advocate, Skeptic, and Judge) as well as embedding-based retrieval to construct exemplars; this cost may be prohibitive at large scale unless carefully optimized or selectively applied. Moreover, we report single-run results for each configuration rather than aggregates over repeated trials, as re-running the full MADRAG pipeline across multiple random seeds or configurations would require substantial additional computational resources due to repeated LLM invocations; consequently, the reported numbers should be interpreted as outcomes from one specific instantiation of the MADRAG pipeline rather than as mean or variance estimates over repeated runs.

Finally, although MADRAG is training-free in that it does not require parameter updates or supervised fine-tuning, it is not data-free: retrieval-augmented generation presupposes access to a bank of manually scored essays. That said, the amount of labeled data needed for RAG is typically much smaller than what is required to train or fine-tune a scoring model—at minimum, one well-chosen exemplar per trait–score level can provide basic coverage for retrieval. In settings with limited coverage, especially for certain traits or score levels, retrieval may fail to surface score-discriminative exemplars, weakening calibration with human preference.

Ethical Considerations

Deploying MADRAG in educational contexts raises ethical risks even when agreement with human raters appears strong. A primary concern is *automation bias*: because the system produces fluent rationales and competitive aggregate metrics, educators or institutions may over-trust its outputs and defer to them even when they are incorrect. This is most consequential in high-stakes settings, where scoring errors can affect students' educational trajectories and where students may have limited ability to contest decisions. For this reason, we view MADRAG as appropriate for formative feedback under human oversight, not as a substitute for human judgment in summative assessment.

MADRAG also carries environmental and access concerns due to its inference-time footprint: each trait requires multiple large-model calls plus embedding-based retrieval, which can scale to hundreds of calls even for a single classroom assignment. This cost has implications for carbon footprint and may exacerbate inequities, where well-resourced schools can afford such tooling while under-resourced schools cannot.

To mitigate these risks, MADRAG should be deployed only in low-stakes, human-in-the-loop contexts where machine scores are treated as advisory rather than definitive, since even the "fairest" automated scoring models exhibit measurable demographic bias and therefore require human adjudication in summative uses (Bridgeman, 2013). Ethical AI grading frameworks further show that embedding human oversight at multiple stages—rubric setup, initial model scoring, and human review with contextual adjustment—can reduce grading time while preserving educators' ability to override erroneous outputs (Litman et al., 2021). Studies on LLM-assisted grading also indicate that formal appeals and re-evaluation procedures can meaningfully correct scoring errors, motivating transparent rubrics, human oversight, and explicit recourse mechanisms (Aytutuldu et al., 2025). Transparent disclosure is equally important: institutions should clearly communicate when automated scoring is used and how grades are produced, as lack of transparency erodes trust (Conijn et al., 2023). By combining human oversight, transparency, and regular auditing, MADRAG can support formative assessment while reducing risks to equity and accountability.

To mitigate these risks, MADRAG should be de-

ployed only in low-stakes, human-in-the-loop contexts where machine scores are treated as advisory rather than definitive, since even the "fairest" automated scoring models exhibit measurable demographic bias and therefore require human adjudication in summative uses (Bridgeman, 2013). Ethical AI grading frameworks demonstrate that embedding human oversight at multiple stages—rubric setup, initial model scoring, human review and contextual adjustment—both reduces grading time and allows educators to override erroneous outputs (Litman et al., 2021). Studies on LLM-assisted grading further show that appeals processes meaningfully correct AI errors; in one case, 74 % of appealed grades were adjusted, leading the authors to conclude that transparent rubrics, human oversight and formal appeal mechanisms are essential for fair outcomes (Aytutuldu et al., 2025). Transparent disclosure of AI involvement is equally important: ethical guidelines emphasize that institutions must clearly explain how data are collected and used and how automated grades are generated, because lack of transparency erodes trust (Conijn et al., 2023). Finally, fairness audits and diverse training data are critical: comparative studies find that models trained on skewed subsets (e.g., only high- or low-ability students) mis-score students outside those groups, underscoring the need to curate exemplar pools that represent all relevant user groups and to evaluate model performance across demographic and cognitive dimensions (Schaller et al., 2024). By combining human oversight, transparency, robust audits and a cautious, research-prototype framing, MADRAG can provide formative feedback without undermining equity or trust.

References

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

I. Aytutuldu, O. Yol, and Y. S. Akgul. 2025. Integrating llms for grading and appeal resolution in computer science education. *Preprint*, arXiv:2504.13557.

Brent Bridgeman. 2013. Human ratings and automated essay evaluation. *Handbook of automated essay evaluation: Current applications and new directions*, pages 221–232.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Rianne Conijn, Patricia Kahr, and Chris Snijders. 2023. The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, 10(1):37–53.

Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954.

Paul Deane. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24. Automated Assessment of Writing.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Avisa Fallah, Ali Keramati, Mohammad Ali Nazari, and Fatemeh Sadat Mirfazeli. 2024. Automating theory of mind assessment with a llama-3-powered chatbot: Enhancing faux pas detection in autism. In *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 365–372.

Joaquín Jordán, Xavier Yin, Melissa Fabros, Gireeja Ranade, and Narges Norouzi. 2025. Magic: Multi-agent argumentation and grammar integrated critiquer. *Preprint*, arXiv:2506.13037.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Ute Knoch. 2009. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2):275–304.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders? *Preprint*, arXiv:2409.13120.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025. Evaluating scoring bias in llm-as-a-judge. *Preprint*, arXiv:2506.22316.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zijie Lin and Bryan Hooi. 2025. Enhancing multi-agent debate system performance via confidence expression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6453–6471, Suzhou, China. Association for Computational Linguistics.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I*, page 255–267, Berlin, Heidelberg. Springer-Verlag.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Dadi Ramesh and Suresh Kumar Sanampudi. 2021. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55:2495 – 2527.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221, Mexico City, Mexico. Association for Computational Linguistics.

Takumi Shibata and Yuichi Miyamura. 2025. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29976–29989, Suzhou, China. Association for Computational Linguistics.

Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. 2025. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *Preprint*, arXiv:2505.13965.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14):e34262.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Preprint*, arXiv:2206.10498.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):157–180.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix: Implementation and Reproducibility Details

This appendix provides implementation details for reproducing the MADRAG pipeline described in Section 3, including decoding settings, log-probability extraction, asynchronous orchestration, and retrieval configuration. For transparency, we summarize the exact execution logic used to generate the scores reported in Section 4.

A.1 Runtime Environment and Execution

All experiments were executed in a single experimental run using a Python pipeline that processes each essay independently and evaluates it across all rubric traits in parallel (one debate instance per trait), due to the computational cost of repeated LLM evaluations. Each trait evaluation consists of (i) Advocate generation, (ii) Skeptic rebuttal generation conditioned on the Advocate, and (iii) Judge scoring conditioned on the debate transcript and retrieved exemplars.

The implementation uses asynchronous execution via `asyncio` to concurrently evaluate all rubric traits for a given essay:

- For each essay, we spawn one task per trait using `asyncio.gather`.
- Each task runs a three-step debate (Advocate → Skeptic → Judge).

Output logging. For each (essay, trait) pair, the system logs: (i) the Judge rationale text, (ii) the parsed final integer score, (iii) raw model outputs, (iv) token-level log-probabilities (where supported), (v) last-token alternative candidates (top logprobs), and (vi) the full debate transcript (Advocate + Skeptic).

A.2 Decoding and Inference Settings

We use deterministic decoding for the **Judge** to reduce run-to-run variance and ensure that score outputs are stable under identical inputs. Concretely:

- **Judge:** temperature set to 0.0 (greedy / deterministic decoding).

For the **Advocate** and **Skeptic**, we use highest-temperature sampling to allow diversity in argumentation:

- **Advocate and Skeptic:** temperature set to 0.7.

Temperature selection for debate agents. We set the decoding temperature for the **Advocate** and **Skeptic** to 0.7 based on a lightweight manual tuning procedure. Concretely, we varied temperature on a small subset of essay–trait instances and qualitatively inspected how it affected the agents' argumentative behavior (e.g., specificity, coverage, and willingness to challenge or defend claims). Because this tuning focused on reasoning style rather than end-task accuracy, we did not treat performance as the selection criterion. Moreover, sweeping temperatures over the full evaluation pipeline would be computationally expensive due to the multi-step, multi-trait debate structure. We therefore fix $T = 0.7$ for all Advocate/Skeptic generations in all experiments.

A.3 Models and API Calls

We evaluate MADRAG with multiple underlying LLMs by swapping the model used in the chat completion call. The codebase includes wrappers for OpenAI chat completion models (e.g., GPT-4o-mini, GPT-4o, GPT-5-mini, GPT-5) called via `AsyncOpenAI`.

A.4 Confidence Proxy from Token Log-Probabilities

To implement confidence-aware judging, we extract a proxy confidence score from the **first generated token** of each debate agent response (Advocate and Skeptic). When available, the OpenAI

Advocate Agent

You are an Advocate Agent in a multi-agent debate system for essay scoring. Your role is to support the essay by highlighting its strengths with respect to strictly within the single trait "$TRAIT_NAME". You must analyze the essay and provide detailed, text-based evidence of what is done well according to the rubric's expectations for "$TRAIT_NAME" trait.
Do not assign a score. Do not summarize or critique weaknesses. Focus entirely on supporting the essay's strengths as they relate to the specific sub-trait. Use quotes or paraphrased excerpts from the essay when needed. Be specific and detailed in your analysis.
Anonymization $ANON_CONTEXT

Figure 3: Adocate system prompt.

API returns per-token log-probabilities. Let ℓ denote the log-probability of the first emitted token; we compute:

$$c = \exp(\ell) \in (0, 1].$$

This value is passed to the Judge as a soft indicator of the agent's internal certainty. In addition, for debugging/analysis we log:

- the full token-level logprob sequence when returned by the API; and
- the top-k alternatives for the last generated token (from `top_logprobs`) when present.

For models without logprob support (GPT-5), this signal is extracted from the models self-report.

A.5 Score Parsing and Output Constraints

The Judge is instructed to output an integer score within the valid trait range. The pipeline parses the final score using a regular expression that extracts:

`Final Score: {integer}`.

All text preceding the final score marker is stored as the Judge rationale. If the score cannot be parsed, the system records the rationale but flags the score as missing.

Skeptic Agent

You are a Skeptic Agent in a multi-agent debate system for essay scoring. Your role is to critically analyze the essay and identify weaknesses according to strictly within the single trait "$TRAIT_NAME". Focus on providing detailed, evidence-based critiques of how the essay falls short for "$TRAIT_NAME" trait.
Do not assign a score. Do not mention positive aspects. Concentrate only on identifying issues, weaknesses, and areas where the essay does not meet the rubric's expectations. Use specific excerpts or descriptions to support your critique.
Anonymization $ANON_CONTEXT

Figure 4: Skeptic system prompt.

A.6 Prompt Templates

All agent prompts are stored as external template files and rendered at runtime using a shared context dictionary. The context includes the trait name, the full rubric trait serialized as JSON, the essay text, the essay prompt/question, and the valid score range. In addition, the Judge templates include (i) retrieved few-shot exemplars spanning the full score range for the current trait and (ii) the debate transcript (Advocate opening + Skeptic rebuttal). We arrived at the final prompt settings through an extensive, iterative prompt-engineering process involving multiple rounds of pilot runs and refinements to enforce role constraints, improve output format reliability, and reduce failure modes (e.g., agents assigning scores or mixing traits). For ease of inspection and reproducibility, we include the exact prompt templates used in our experiments below. Figures 3–5 show the system instructions used for the Advocate, Skeptic, and Judge roles, respectively.

A.7 Retrieval-Augmented Exemplar Construction

We use retrieval augmentation to provide the Judge with calibration examples spanning the full score range for each trait. For each essay and trait, we construct a **trait-specific exemplar prompt** by retrieving one example essay for each valid score in $[s_{\min}, s_{\max}]$.

Judge Agent

You are "The Synthesizer-Judge," an impartial arbiter for the single trait "$TRAIT_NAME" multi-agent debate system for essay scoring.
Your Job - Read the debate transcript between Advocate and Skeptic agents who previously debated regarding the essay strengths and weaknesses. - Weigh the arguments against the rubric for "$TRAIT_NAME". - Produce a final integer score from $MIN_POINTS to $MAX_POINTS.
Anonymization $ANON_CONTEXT

Figure 5: Judge system prompt.

Operationally, for each trait name τ and score set $\{s_{\min}, \ldots, s_{\max}\}$, we call a function and concatenate the returned exemplars into a few-shot block that is injected into the Judge context.

RAG isolation across roles. To prevent debate agents from anchoring on retrieved examples, we provide retrieved exemplars to:

- **Judge only**, as part of its context prompt.

The Advocate and Skeptic receive only the rubric trait and essay content.

A.8 Example Multi-Agent Debates

We present representative examples of the multi-agent debate process used in MADRAG, including the Advocate, Skeptic, and Judge agents. Each example corresponds to a single essay–Idea Trait (Figures 6, 7).

B Dataset Details and Preprocessing

ASAP overview. The ASAP dataset is a widely used benchmark for automated essay scoring, originally released as part of a Kaggle competition sponsored by the William and Flora Hewlett Foundation. The dataset consists of anonymized English essays written by students in grades 7–10 in response to eight distinct prompts, each defining a separate *essay set*. Essay sets vary substantially in genre (persuasive, narrative, and source-dependent response), length, grade level, and scoring rubric, making ASAP a challenging and diverse evaluation benchmark for AES systems.

All essays were scored by trained human raters following prompt-specific guidelines. Each essay receives a resolved (overall) score, and for a subset of prompts, additional analytic trait scores are available. Due to these properties, ASAP has been extensively adopted in prior work evaluating both holistic and trait-level essay scoring models.

While all essay sets include holistic scores, the ASAP release provides trait-level annotations only for Essay Sets 7 and 8. Accordingly, we document the full dataset for completeness and reproducibility, but restrict our experiments to Sets 7 and 8, the only subsets that provide multiple independent human ratings at the trait level.

B.1 ASAP Essay Set Statistics

Table 6 summarizes key properties of all eight ASAP essay sets, including essay type, grade level, training set size, and the availability of trait-level annotations. Consistent with prior analyses, essay lengths range from short source-dependent responses (approximately 150 words) to long narrative essays exceeding 600 words on average, with score ranges varying substantially across prompts.

Rationale for focusing on Essay Sets 7 and 8. Although ASAP contains eight essay sets, only Essay Sets 7 and 8 provide independent trait-level scores from at least two human raters per essay. This property is essential for our study, which explicitly examines trait-level reliability, Human–Human agreement, and model calibration under rater disagreement. Consequently, all quantitative evaluations in the main paper are conducted exclusively on Sets 7 and 8.

B.2 Essay Set 7: Prompt

Prompt. *Write about patience. Being patient means that you are understanding and tolerant. A patient person experiences difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.*

B.3 Essay Set 8: Prompt

Prompt. *We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important*

Table 6: Summary of the ASAP dataset across all eight essay sets. Trait-level annotations are available only for Essay Sets 7 and 8 in the ASAP release.

Set	Essay Type	Grade	Train Size	Traits
1	Persuasive / Narrative / Expository	8	1,783	—
2	Persuasive / Narrative / Expository	10	1,800	—
3	Source-dependent responses	10	1,726	—
4	Source-dependent responses	10	1,772	—
5	Source-dependent responses	8	1,805	—
6	Source-dependent responses	10	1,800	—
7	Persuasive / Narrative / Expository	7	1,569	4 traits
8	Persuasive / Narrative / Expository	10	723	6 traits

part of any relationship. Tell a true story in which laughter was one element or part.

B.4 Text Transcription and Fidelity

ASAP essays were transcribed from handwritten student responses following strict transcription guidelines. Misspellings and grammatical errors were preserved exactly as written, and no normalization or correction was applied that could alter surface-level evidence relevant to traits such as *Conventions*. When a handwritten word could not be reliably inferred, it was omitted according to the original transcription protocol. In our experiments, we didn't apply any preprocessing: No spelling correction, grammar normalization, or sentence restructuring is performed.

B.5 Trait Labels and Preprocessing

Rater scores. For Essay Sets 7 and 8, each essay–trait instance includes scores from at least two independent human raters. These scores are retained explicitly to compute Human–Human agreement and to define evaluation targets.

Reference label construction. For model-vs-human evaluation, we construct a single reference score per essay–trait pair by averaging the available rater scores and rounding to the nearest valid integer within the trait's scoring range. This procedure avoids privileging any individual rater while remaining consistent with the discrete rubric scales.

Retrieval pool and data leakage prevention. For retrieval-augmented judging, exemplar essays are drawn exclusively from the training split of the same essay set. The evaluated essay is never eligible to be retrieved as an exemplar. Additional details of exemplar construction and role-specific access are provided in Appendix A.7.

C Baseline Models

ProTACT. We include ProTACT (Do et al., 2023) as a strong *training-based* neural baseline for cross-prompt analytic trait scoring. ProTACT learns prompt-aware essay representations via essay–prompt attention and augments them with engineered essay-quality features (including a topic-coherence feature), while a trait-similarity objective encourages consistent predictions across correlated traits. Because the original paper does not report a complete set of trait-wise results for ASAP Sets 7–8, we run the authors' released implementation on the public data and reproduce the evaluation pipeline to obtain the missing sub-trait QWK scores reported in our tables.

Feature-Engineered Trait Scorer (FeatEng-RF). We include the supervised trait-scoring baseline of Mathias and Bhattacharyya (Mathias and Bhattacharyya, 2020), which predicts analytic trait scores on ASAP using a Random Forest model trained on a large set of hand-crafted linguistic features (e.g., length, punctuation, syntax, style, and cohesion indicators such as discourse connectives and entity-grid features) under a five-fold cross-validation protocol. Because this approach is trained directly on ASAP trait labels, we treat it as a strong *supervised* reference point. (Mathias and Bhattacharyya, 2020)

Prompt-Engineered Zero-Shot LLM Judge (ZS-LLM). We report training-free, single-agent LLM baselines from Mansour et al. (Mansour et al., 2024), who evaluate ChatGPT (gpt-3.5-turbo) and LLaMA-2-13B-Chat on ASAP via rubric-aware, prompt-engineered scoring without any supervised fine-tuning. Their prompts provide the essay prompt, score range, and rubric guidelines, and are progressively strengthened with structured instructions, role formatting, and one-shot exemplars; decoding is deterministic (temperature $= 0$). The

model outputs holistic and trait-level scores (optionally with feedback), and the authors find performance is highly prompt- and task-dependent, yet remains well below supervised and cross-prompt SOTA in QWK—especially for trait scoring on Sets 7 and 8.

Criteria & Sample-Referenced LLM Scoring (CSR-J). We include the training-free prompt-based LLM scorer of Tang et al. (Tang et al., 2024), which uses a single GPT-4 judge to assign analytic trait scores from the rubric and to produce brief rationales grounded in *sample-referenced* exemplars (i.e., human-scored example essays provided in the prompt). The method evaluates ASAP Essay Set 7 on Ideas, Organization, Style, and Conventions, and reports trait-wise QWK under deterministic decoding (temperature $= 0$), serving as a strong single-agent, prompt-engineered reference.

Human–Human agreement. Finally, we report Human–Human agreement as QWK between the two human raters for each trait, serving as an approximate reference ceiling given inherent rater variability.

C.1 Why One Round and Why Advocate→Skeptic

Finally, we analyze design choices in the debate transcript provided to the judge. Our objective is to feed the judge a *minimal but sufficient* argumentative context: an evidence-heavy pro argument followed by a targeted rebuttal. We compare (i) advocate-only transcripts, (ii) skeptic-only transcripts, and (iii) concatenating both threads, as well as the effect of adding an additional debate round.

One round is substantially more stable than two rounds. Across both sets, extending debate beyond a single exchange causes a sharp drop in QWK (e.g., `both_round2` collapses relative to `both_round1`). This is consistent with the hypothesis that longer debates amplify verbosity, drift, and non-local contradictions, which can degrade the judge's calibration even when RAG exemplars are provided.

Advocate-first is a better conditioning signal than skeptic-first. Skeptic-only transcripts yield near-zero agreement in both sets, indicating that leading with exclusively negative framing can push the judge toward systematic under-scoring or rubric-misaligned reasoning. In contrast, advocate-first provides a structured inventory of rubric-aligned evidence, after which the skeptic rebuttal can selectively challenge specific claims. This ordering preserves *coverage* (positives are surfaced) while still introducing *adversarial pressure* (weaknesses are surfaced) in a controlled way.

Table 7: Essay Set 7: Transcript configuration analysis (QWK). Adv.=Advocate-only, Skp.=Skeptic-only, R=Round. * denotes our proposed method (MADRAG).

Method	Idea	Org.	Sty.	Cnv.
Adv. (R1)*	0.43	**0.64**	**0.45**	**0.26**
Adv. (R2)	0.46	0.24	0.10	0.02
Both (R1)	**0.48**	0.24	0.14	0.08
Both (R2)	0.14	0.02	0.03	0.01
Skp. (R1)	0.07	0.05	0.01	0.00
Skp. (R2)	0.06	0.01	0.01	0.01

Why not concatenating both full threads. Although `both_round1` can be competitive on some traits, it is less reliable across traits and prompts than the advocate-first exchange used in MADRAG. Empirically, concatenation appears to dilute the discourse structure (two parallel narratives with competing local context), increasing the judge's burden to resolve inconsistencies. The advocate→skeptic exchange yields a single, linear argumentative path that is easier for the judge to synthesize.

Quantitative evidence. Tables 7 and 8 summarize QWK under different transcript configurations. The overall pattern is clear: skeptic-first is consistently poor; two rounds is unstable; and a short advocate-led exchange offers the most reliable trade-off between signal and noise.

D Detailed Ablation Study

This appendix provides a detailed breakdown of the ablation study discussed in Section 4.2.1. The study evaluates the incremental impact of each major component in the MADRAG pipeline by comparing the following configurations:

- **SA (Single-Agent):** A single LLM acts as a judge, scoring the essay directly using the rubric without debate or retrieved exemplars.
- **SARAG (Single-Agent+RAG):** A single LLM judge scores the essay using the rubric *and* retrieved, rubric-aligned exemplar essays spanning the score range, but *without* trait decomposition or debate.
- **MA (Multi-Agent):** Multiple, independent LLM agents score the essay for a trait. Their

Table 8: Essay Set 8: Transcript configuration analysis (QWK). * denotes our proposed method (MADRAG).

Method	Idea	Org.	Voc.	Word	Sent.	Cnv.
Adv. (R1)*	**0.59**	**0.47**	**0.60**	**0.65**	**0.55**	**0.58**
Adv. (R2)	0.37	0.20	0.39	0.37	0.54	0.62
Both (R1)	0.57	0.45	0.52	0.63	0.53	0.54
Both (R2)	0.17	0.17	0.19	0.30	0.30	0.27
Skp. (R1)	0.11	0.03	0.13	0.14	0.13	0.10
Skp. (R2)	0.15	0.10	0.15	0.15	0.23	0.18

scores are averaged, simulating a multi-rater setup without interaction or retrieval.

- **MAD (Multi-Agent Debate):** Introduces the Advocate and Skeptic agents who generate a debate transcript. The Judge scores the essay based on this transcript, but *without* access to retrieved exemplars for calibration.
- **MARAG (Multi-Agent with RAG):** Multi-agent scoring is combined with RAG. The Judge receives exemplars spanning the score range but does *not* see a debate transcript.
- **MADRAG:** The full proposed framework, combining the Advocate–Skeptic debate transcript with retrieval-augmented exemplars for the Judge.

Table 9 presents the complete trait-wise results for Essay Sets 7 and 8 separately. The merged view, which averages overlapping traits, is shown in the main paper as Figure 2.

Key observations. Several consistent patterns emerge from the detailed ablation results.

- **Retrieval provides the largest single gain in calibration.** Comparing SA to SARAG reveals that exemplar-based retrieval alone yields substantial improvements across both essay sets, particularly on surface-oriented traits such as *Conventions*. This confirms that access to score-calibrated exemplars is a primary driver of improved agreement, even in the absence of decomposition or debate.
- **Debate and decomposition offer complementary but trait-dependent benefits.** Moving from SA to MA yields modest gains, indicating that trait-wise decomposition and multiple perspectives help stabilize judgments but are insufficient on their own. Adding debate (MAD) further improves performance on several discourse-oriented traits, most notably *Organization* and *Sentence Fluency*, suggesting that adversarial reasoning is particularly beneficial when evaluating higher-level structure and coherence.
- **Multi-agent retrieval (MARAG) outperforms single-agent retrieval (SARAG).** Across nearly all traits, MARAG consistently improves over SARAG, indicating that trait-wise decomposition remains valuable even when retrieval is present. This gap highlights that retrieval alone does not fully resolve rubric alignment issues without trait-specific conditioning.
- **Debate can introduce noise on surface-level traits.** For some surface traits, MARAG slightly outperforms MADRAG. For example, on *Conventions* in Set 7, MARAG achieves higher agreement than MADRAG (0.35 vs. 0.26), and a similar pattern appears for *Word Choice* in Set 8. These regressions suggest that adversarial debate can amplify spurious or surface-form cues, motivating a closer analysis of debate-induced failure modes.
- **Overall, MADRAG delivers the strongest and most consistent performance.** Despite occasional regressions on individual surface traits, the full MADRAG framework achieves the best or near-best performance on the majority of traits across both essay sets, particularly for discourse-oriented dimensions. This pattern confirms the intended synergy between adversarial reasoning and exemplar-based calibration.

In Section 4.4), we analyze the qualitative failure mechanisms that arise from debate–retrieval interactions.

Table 9: Ablation study (QWK) on ASAP Essay Sets 7 and 8. SA: Single-Agent, MA: Multi-Agent, MAD: Multi-Agent Debate, MARAG: Multi-Agent with RAG, MADRAG: Full framework.

Set	Method	Idea	Org.	Voc.	Word	Sent	Sty.	Cnv.
7	SA	0.27	0.33	—	—	—	0.29	0.06
7	SARAG	0.27	0.36	—	—	—	0.32	0.20
7	MA	0.31	0.39	—	—	—	0.28	0.17
7	MAD	0.25	0.40	—	—	—	0.30	0.20
7	MARAG	**0.56**	0.47	—	—	—	0.40	**0.35**
7	MADRAG	0.43	**0.64**	—	—	—	**0.47**	0.26
8	SA	0.41	0.26	0.26	0.12	0.13	—	0.08
8	SARAG	0.45	0.42	0.44	0.32	0.27	—	0.28
8	MA	0.47	0.54	0.37	0.55	0.43	—	0.34
8	MAD	0.47	0.57	0.42	0.61	0.51	—	0.46
8	MARAG	0.45	**0.49**	0.45	0.52	0.55	—	0.53
8	MADRAG	**0.59**	0.47	**0.60**	**0.65**	**0.55**	—	**0.58**

E Detailed Qualitative Analysis

E.1 Primary Failure Mechanisms and Component Attribution Links

To probe *which* mechanisms are linked to which components, Table 10 reports the $C \times D$ matrix. Debate framing capture (C2) is predominantly debate-linked (55.4% D1 row-normalized), consistent with the judge inheriting the debate stance without verifying against the essay. In contrast, anonymization distortion (C6) is disproportionately interaction-coded (63.0% D3), suggesting that surface-form cues often become harmful when debate and retrieval jointly increase attention to token-level artifacts. Finally, the template-like mechanisms (C4/C5) are most often *not* component-specific (66.7% and 64.3% D4), indicating that generic rubric prose and mid-band defaults are largely baseline judge limitations rather than uniquely induced by debate or retrieval.

E.2 Primary Failure Mechanisms and Reasoning Quality Links

Table 11 links reasoning quality to failure mechanisms. Anonymization distortion (C6) accounts for the majority of B0 cases (8/10), indicating that truly ungrounded rationales often arise when placeholders are mistaken as genuine mechanical errors. In contrast, exemplar-induced calibration errors (C4) are the one mechanism that frequently yields *high-quality* rationales (B2: 46.2% within C4) despite being wrong, suggesting that these errors are less about incoherent reasoning and more about systematic miscalibration toward a rubric band.

Table 10: $C \times D$ contingency table for wrong MADRAG cases. Top: counts. Bottom: row-normalized percentages (each row sums to 100).

Counts				
	D1	D2	D3	D4
C1	5	2	3	3
C2	31	6	13	6
C3	13	2	13	2
C4	2	0	2	8
C5	1	4	0	9
C6	12	1	29	4
Row-normalized (%)				
C1	38.5	15.4	23.1	23.1
C2	55.4	10.7	23.2	10.7
C3	43.3	6.7	43.3	6.7
C4	16.7	0.0	16.7	66.7
C5	7.1	28.6	0.0	64.3
C6	26.1	2.2	63.0	8.7

Table 11: $B \times C$ on wrong MADRAG cases. Top: counts. Bottom: row-normalized percentages.

Counts			
	B0	B1	B2
C1	0	10	3
C2	0	53	3
C3	2	27	1
C4	0	7	6
C5	0	14	0
C6	8	38	0
Row-normalized (%)			
C1	0.0	76.9	23.1
C2	0.0	94.6	5.4
C3	6.7	90.0	3.3
C4	0.0	53.8	46.2
C5	0.0	100.0	0.0
C6	17.4	82.6	0.0

E.3 Mechanism deep dives: themes and micro-theories

We synthesize the most frequent failure mechanisms into three recurring themes, each expressed as a micro-theory about how debate and retrieval shape the judge's attention and calibration.

Theme 1: Token myopia (C6) — Anonymization treated as real error. A dominant pattern is that MADRAG cites anonymization markers (e.g., `@CAPS`/`@PERSON`/`@DATE`) as "capitalization" or "formatting" failures, especially in *Conventions* (and occasionally *Fluency* or *Voice*). **Micro-theory:** when the surface form contains many anonymization tokens, MADRAG over-weights them as evidence of convention breakdown and readability loss, leading to systematic under-scoring even

when the underlying prose is readable.

Theme 2: Debate capture (C2) — Judge inherits its stance without verification. In many failures, the judge echoes Advocate/Skeptic framing (often the Skeptic) without checking whether the claimed defect is supported by the essay (e.g., "no thesis," "no paragraph breaks," "disorganized" despite clear temporal markers and closure). **Micro-theory:** debate increases the salience of critique, but the judge sometimes substitutes "debate resolution" for "text verification," producing overconfident misdiagnoses about structure and coherence.

Theme 3: Rubric-template collapse (C4/C5) — Generic band language replaces close reading. A smaller but important class of errors reflects template-driven justifications (e.g., "clear but limited development", "errors impede readability") that are weakly tied to the essay and insensitive to strong counter-evidence. **Micro-theory:** under uncertainty, MADRAG falls back on plausible-sounding rubric prose, reducing sensitivity to extremes and enabling large deviations when the essay is clearly strong or clearly weak on the target trait.

E.3.1 Implications for MADRAG design

The qualitative results suggest that MADRAG's components change *what the judge attends to*, not only the final score. Debate often improves structured critique, but it also creates frequent failure via framing capture (C2) and acceptance of unsupported debate claims (C3), indicating the need for explicit *text verification* constraints in the judge prompt. Retrieval is less often the sole driver of failure (D2), but interaction effects are common (D3), especially when surface-form noise is present. Finally, anonymization tokens represent a systematic confound for convention-heavy traits: without explicit normalization or masking, placeholders are repeatedly treated as genuine mechanical errors, producing predictable under-scoring.

Table 12: Representative Token Myopia case (C6): MADRAG treats anonymization tokens as mechanical errors, driving an overly low conventions judgment.

Trait	Conventions
Memo (E)	For conventions, the essay shows generally strong control of standard conventions: many sentences have correct end-of-sentence punctuation and accurate spelling. MADRAG emphasizes that capitalization/placeholders and punctuation errors significantly impede readability, but this is misleading because many flagged issues are anonymization tokens (e.g., `@CAPS`, `@PERSON`) rather than author errors, and multiple clearly punctuated sentences remain readable.

Table 13: Representative Debate Capture case (C2): MADRAG adopts a negative framing about organization despite clear sequencing cues.

Trait	Organization
Memo (E)	For organization, the essay shows clear chronological sequencing and a recognizable beginning, middle, and reflective ending with explicit temporal markers. MADRAG emphasizes abrupt transitions and disjointedness, but this is misleading because the text uses markers such as "The next day" and "To my surprise" and provides a coherent arc with closure, indicating functioning structure.

Table 14: Representative Rubric-Template case (C4/C5): MADRAG uses mid-band generic rationale that underresponds to essay evidence.

Trait	Ideas and Content
Memo (E)	For ideas_and_content, the essay demonstrates a clear, focused narrative with a developed main idea supported by concrete scenes and an explicit resolution. MADRAG emphasizes that the piece is fragmented and insufficiently developed, but this is misleading because the text provides escalating, sensory detail and a decisive emotional payoff that aligns with high-band rubric traits.

Turn 1: Advocate Message

```
As the Advocate for the "Ideas" criterion, my role is to highlight the strengths of the essay in
↪  terms of its focus and development of ideas related to the topic of patience.

The essay demonstrates a clear focus on the topic of patience by centering the narrative around
↪  the relationship between the narrator and their horse, @CAPS1. The story is consistently
↪  anchored in the theme of patience, as evidenced by the repeated references to the horse's
↪  impatience and the narrator's attempts to manage it. This focus is maintained throughout the
↪  essay, ensuring that the reader understands the central theme.

The development of ideas is supported by specific and relevant details that bring the story to
↪  life. For instance, the essay describes @CAPS1 as "a chestnut quarter horse with a white star
↪  on his forehead," which not only provides a vivid image of the horse but also personalizes
↪  the narrative, making the story more engaging. The description of @CAPS1's behavior, such as
↪  being "hipper and ready to go" and not standing still when being tacked up, adds depth to the
↪  portrayal of the horse's impatience. These details are specific and relevant, as they
↪  directly illustrate the challenges the narrator faces in dealing with an impatient horse.

Furthermore, the essay includes specific examples of situations where patience is tested, such
↪  as when @CAPS1 is "so far up ahead" during a ride that the narrator has to stop him to wait
↪  for the group. This scenario effectively highlights the theme of patience, as it shows the
↪  narrator's efforts to manage the horse's behavior and maintain harmony within the group. The
↪  narrative also conveys the narrator's ongoing struggle and acceptance of the horse's nature,
↪  as seen in the line, "Meany times I have tried training him to be patient but I have always
↪  failed." This acknowledgment of the narrator's attempts and eventual acceptance adds a layer
↪  of complexity to the story, demonstrating a thorough development of the theme.

In conclusion, the essay excels in maintaining a clear focus on the theme of patience and
↪  develops this idea with specific, relevant details that enrich the narrative. The consistent
↪  emphasis on the horse's impatience and the narrator's response to it effectively illustrates
↪  the challenges and nuances of being patient.
```

Turn 2: Skeptic Rebuttal

```
The Advocate's argument suggests that the essay maintains a clear focus on the theme of patience
↪  and develops this idea with specific, relevant details. However, upon closer examination, the
↪  essay's focus on the theme of patience is not as clear or thoroughly developed as claimed.

Firstly, the essay primarily describes the horse's impatience rather than exploring the concept
↪  of patience itself. The narrative is centered around the horse's behavior, with limited
↪  exploration of the narrator's experience of patience. The essay mentions the horse being
↪  "hipper and ready to go" and not standing still, but these details focus more on the horse's
↪  characteristics rather than the narrator's patience. The narrative lacks depth in
↪  illustrating how the narrator embodies patience or how they cope with the challenges
↪  presented by the horse's behavior.

Moreover, the essay does not provide a comprehensive exploration of the narrator's attempts to
↪  manage the horse's impatience. While it mentions that the narrator has "tried training him
↪  to be patient," it does not delve into specific strategies or experiences that demonstrate
↪  the narrator's patience. The lack of detailed examples of the narrator's actions or
↪  reflections on their patience limits the development of the theme.

Additionally, the essay concludes with the narrator's acceptance of the horse's nature, stating,
↪  "I will always love him no matter how un patience he is." This statement shifts the focus
↪  from patience to acceptance, which, while related, is not the same concept. The narrative
↪  does not sufficiently explore the process or challenges of being patient, which is the
↪  central task of the essay prompt.

In summary, the essay's ideas are not thoroughly developed in relation to the theme of patience.
↪  The focus is more on the horse's impatience rather than a nuanced exploration of the
↪  narrator's patience, resulting in a narrative that lacks depth and specificity in addressing
↪  the essay prompt.
```

Figure 6: Example multi-agent debate for the *Ideas* trait (Part I): Advocate initiation and Skeptic rebuttal.

Turn 3: Judge Decision

```
"The debate between the Advocate and the Skeptic highlights differing perspectives on the
↪  essay's focus and development of ideas related to the theme of patience.

The Advocate argues that the essay maintains a clear focus on patience by centering the
↪  narrative around the relationship between the narrator and their horse, @CAPS1. The Advocate
↪  points out that the essay includes specific and relevant details, such as the horse's
↪  description and behavior, which illustrate the challenges the narrator faces. The Advocate
↪  also notes that the essay provides examples of situations where patience is tested, adding
↪  depth to the narrative.

On the other hand, the Skeptic contends that the essay primarily describes the horse's
↪  impatience rather than exploring the concept of patience itself. The Skeptic argues that the
↪  narrative lacks depth in illustrating how the narrator embodies patience or copes with the
↪  challenges presented by the horse's behavior. The Skeptic also points out that the essay
↪  does not provide detailed examples of the narrator's actions or reflections on their
↪  patience, limiting the development of the theme. Additionally, the Skeptic notes that the
↪  conclusion shifts the focus from patience to acceptance, which is not the same concept.

Weighing these arguments against the rubric, the essay does tell a story that is somewhat
↪  focused on the topic of patience. However, the development of ideas is more centered on the
↪  horse's impatience rather than a thorough exploration of the narrator's patience. The essay
↪  includes a mix of specific and general details, but it lacks a comprehensive exploration of
↪  the narrator's experience of patience. The narrative does not delve deeply into the
↪  strategies or experiences that demonstrate the narrator's patience, which limits the
↪  development of the theme.

Based on this analysis, the essay aligns more closely with a score of 2, as it tells a story with
↪  ideas that are somewhat focused on the topic and are developed with a mix of specific and
↪  general details.

Final score: 2"
```

Figure 7: Example multi-agent debate for the *Ideas* trait (Part II): Judge synthesis and final score.