THE FIRST TOKENS MATTER: EARLY CONFIDENCE SIGNALS FOR EVALUATING LLM REASONING

Ali Keramati, Justin Cheok , Jacob Horne & Mark Warschauer
University of California, Irvine
{a.kera,jcheok,jhorne1,markw}@uci.edu

ABSTRACT

Assessing the logical reasoning quality of large language models (LLMs) remains a central challenge, particularly in open-ended settings where reference answers are unavailable. In this work, we investigate whether intrinsic confidence signals, specifically token-level log-probabilities produced during decoding, can serve as reliable indicators of reasoning quality in multi-agent LLM systems. We study this question in a debate-based framework, where agents generate competing arguments and are evaluated by an LLM-as-judge along structured reasoning dimensions. We find that early-generation confidence signals, especially dispersion within the first few tokens, consistently outperform full-sequence statistics in predicting judged reasoning quality. Trajectory analysis reveals that the opening phase of generation is the most heterogeneous, making it the most informative region for distinguishing high- and low-quality reasoning. We further identify a systematic asymmetry between reasoning roles: confidence aligns more strongly with supportive reasoning than with adversarial critique, reflecting differences in their underlying failure modes. These findings suggest that early decoding dynamics provide a lightweight and scalable signal for estimating reasoning reliability, offering a bridge between intrinsic model uncertainty and external evaluation of logical reasoning in LLMs.

1 INTRODUCTION

Recent advances in large language models (LLMs) have enabled the development of *multi-agent systems*, in which multiple specialized agents collaborate to solve complex tasks Wu et al. (2023). By decomposing problems into role-specific subtasks, such systems have been shown to improve performance, robustness, and consistency across a range of applications, including reasoning, planning, and automated decision-making Parmar et al. (2025); Han et al. (2026). Among interaction paradigms, *debate* has emerged as a particularly effective mechanism: by eliciting both supporting and opposing arguments, it encourages exploration of diverse reasoning paths and exposes errors that may remain hidden in a single-agent trajectory Du et al. (2023).

Rubric-based scoring provides a concrete and high-impact setting in which these benefits are especially relevant. In this setting, a system assigns scores according to a predefined rubric that specifies evaluation criteria and score ranges Fallah et al. (2024). A canonical example is *automated essay scoring (AES)*, where models aim to approximate human judgments of student writing quality Dikli (2006). Public benchmarks such as the ASAP [1] dataset include prompts that provide trait-level rubric scores (rather than a single holistic score), enabling trait-specific feedback and analysis Crossley et al. (2025). At the same time, recent work has explored using LLMs directly for essay scoring, highlighting both the promise of scalable rubric-based evaluation and the need to better understand the reliability of LLM-driven scoring behavior Pack et al. (2024).

Multi-agent debate is a natural fit for rubric scoring because it produces inspectable intermediate reasoning artifacts Keramati & Warschauer (2025). In these systems, agents adopt complementary roles, generating diverse perspectives on the same input. This structured disagreement can help the system consider alternative interpretations of the rubric and mitigate single-path scoring bias by

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

forcing explicit engagement with counterevidence Du et al. (2023). However, debate also increases system complexity: multiple agents, multiple messages, and multiple opportunities for subtle procedural failures Wynn et al. (2025). As multi-agent pipelines become more elaborate, it becomes essential to add an evaluation layer that measures not only whether a final score matches a reference, but also whether agents' reasoning is high quality and reliable Chen et al. (2025).

A growing body of work addresses this need through *LLM-as-judge* evaluation, where a separate language model is used to score generated outputs along predefined criteria. This paradigm has become a scalable alternative to human evaluation, particularly for open-ended tasks where reference answers are unavailable Zheng et al. (2023). However, LLM-as-judge provides only an *external* signal of quality, and an important open question remains: *to what extent do these judgments reflect the true reliability of the underlying reasoning process?* In particular, can we identify *intrinsic signals* within the generating model that correlate with externally judged reasoning quality?

To connect judge-based reasoning evaluation with model-intrinsic signals, we turn to *confidence estimation* and *uncertainty quantification* Kang et al. (2025). Neural probabilities are not automatically calibrated, and language model confidence can be misaligned with correctness. Nevertheless, recent research shows that language models can provide meaningful self-evaluations under appropriate formats and that uncertainty estimation for LLM generation is an active area of study Mavi et al. (2025). In this work, we operationalize model confidence using token-level log-probabilities produced during decoding. Intuitively, if an agent follows a more coherent and evidentially grounded reasoning path, the model should assign higher probability mass to the tokens it generates along that path, yielding more confident logprob trajectories.

2 Related Work

LLM-as-Judge Evaluation. Recent work has established *LLM-as-judge* as a practical paradigm for evaluating open-ended generation in settings where reference answers are weak or unavailable. Prior studies show that strong language models can correlate well with human judgments on instruction-following and related tasks, making them a scalable alternative to manual evaluation Chiang & Lee (2023); Dubois et al. (2025); Zheng et al. (2023); Fu et al. (2024); Liu et al. (2023). Beyond coarse pairwise or scalar judgments, subsequent work emphasizes the need for more structured and interpretable evaluation. For example, FLASK introduces fine-grained, rubric-based assessment and demonstrates improved interpretability and reliability compared to skill-agnostic scoring Ye et al. (2024).

Despite these advances, a growing body of meta-evaluation work highlights fundamental limitations of LLM judges. Prior studies document systematic biases such as verbosity and positional bias, limited self-consistency, and sensitivity to prompt design and evaluation protocols Wang et al. (2024); Zeng et al. (2024); Zheng et al. (2023); Liu et al. (2023). In response, recent approaches propose more elaborate judging strategies, including chain-of-thought and decomposition-based evaluation, multi-aspect scoring, reference-based comparisons, and multi-agent or debate-style evaluators such as PRD and ChatEval Gong & Mao (2023); Saha et al. (2024); Li et al. (2024); Chan et al. (2023); Jeong et al. (2024). However, evidence on the effectiveness of these methods remains mixed. REIFE shows that gains from evaluation protocols depend strongly on the base model and dataset, underscoring the need for diverse and well-calibrated evaluation setups Liu et al. (2025). Similarly, Huang et al. demonstrate that fine-tuned judge models (e.g., JudgeLM, PandaLM, Auto-J, Prometheus) often fail to generalize beyond their training domain, behaving more like task-specific classifiers than robust evaluators Huang et al. (2025).

Confidence and Uncertainty in LLMs. A parallel line of work investigates whether *intrinsic confidence signals* can be used to assess the reliability of LLM outputs. Research in calibration and uncertainty quantification shows that neural probabilities are informative but not inherently well calibrated, meaning that high confidence does not always correspond to correctness Desai & Durrett (2020); Kadavath et al. (2022); Quevedo et al. (2024). Nevertheless, token-level probabilities remain one of the most direct signals available during generation, and have been widely used to detect hallucinations, factual inconsistencies, and uncertain outputs through log-probability- and entropy-based features Liu et al. (2022); Manakul et al. (2023); Mallen et al. (2023).

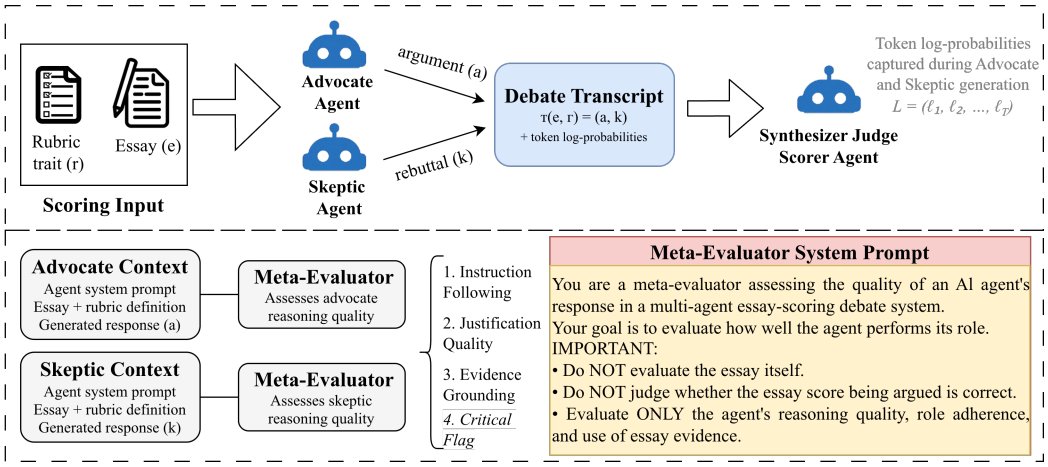

Figure 1: Overview of the proposed multi-agent debate and LLM-as-judge evaluation framework.

In addition, work on self-evaluation suggests that LLMs can sometimes produce useful confidence estimates in natural language, though these verbalized signals may diverge from underlying model uncertainty, particularly in multi-step reasoning settings Kadavath et al. (2022); Mavi et al. (2025). Recent surveys therefore advocate for scalable uncertainty estimation methods that combine intrinsic decoding-time signals with downstream evaluation metrics Kang et al. (2025).

3 Methodology

Figure 1 provides an overview of our framework, which builds upon the multi-agent debate architecture introduced in prior work Keramati & Warschauer (2025) and extends it with an LLM-as-judge meta-evaluation module for reasoning analysis. In the first stage, an **Advocate** and a **Skeptic** produce opposing arguments for a given essay–rubric pair while exposing token-level log-probabilities as intrinsic confidence signals. In the second stage, a separate meta-evaluator scores each argument along rubric-based dimensions such as instruction following, justification quality, and evidence grounding. This design enables systematic analysis of the relationship between internal confidence signals and externally judged reasoning quality.

3.1 Problem Setting

Let $\mathcal{E}$ denote the set of essays and $\mathcal{R}$ the set of rubric traits. Each essay $e \in \mathcal{E}$ consists of unstructured text together with optional metadata, and each rubric trait $r \in \mathcal{R}$ specifies a textual description and a scoring range $[m_r, M_r]$. For each essay–trait pair (e, r), the debate system produces a transcript

$$\tau(e, r) = (a,\, k),$$

where a is the Advocate's argument and k is the Skeptic's rebuttal. Both arguments are generated by a language model conditioned on the essay, rubric, and conversation history; the model simultaneously produces token-level log-probabilities reflecting its internal confidence over candidate continuations. Given a collection of debate responses $\{(a_i, k_i)\}$, each paired with confidence signals c_i and meta-evaluation scores q_i, our objective is to analyze whether token-level probability signals correlate with the externally judged quality of agent reasoning, and thus whether intrinsic confidence can serve as an indicator of reasoning reliability in multi-agent LLM systems.

3.2 Agents and Roles

The debate framework comprises three specialized agents that interact sequentially for each essay–trait pair. The **Advocate** initiates the debate by constructing an argument that highlights the essay's strengths relative to the rubric trait, drawing exclusively on supporting evidence from the essay text without assigning a score. The **Skeptic** responds by identifying limitations or shortcomings

in the essay with respect to the same criterion, producing an evidence-based rebuttal that challenges the Advocate's claims, again without scoring. The **Synthesizer-Judge Scorer** reads the completed transcript and produces the final trait-level score within the allowed rubric range. Because this agent performs a constrained decision-making task whose output can be evaluated directly against ground-truth scores using accuracy-based metrics, it falls outside the scope of the present study. Our analysis focuses exclusively on the open-ended reasoning produced by the Advocate and Skeptic. Full system prompts for all three agents are provided in **Appendix C**.

3.3 Confidence Signals from Token Log-Probabilities

We estimate model confidence using token-level log-probabilities obtained during generation. For a generated response of T tokens, the model produces a log-probability at each decoding step:

$$\ell_t = \log p(t_t \mid t_{<t},\, x),$$

where x is the prompt context and $t_{<t}$ the preceding tokens. The resulting sequence $L = (\ell_1, \ldots, \ell_T)$ forms a log-probability trajectory over the full response.

3.3.1 Window-Based Segmentation

Rather than summarizing L with a single statistic, we extract contiguous sub-sequences to examine how confidence evolves across different phases of generation. We use two complementary strategies:

Fixed-length windows. For window size k:

$$W_{\text{first}}(k) = (\ell_1, \ldots, \ell_k), \qquad W_{\text{last}}(k) = (\ell_{T-k+1}, \ldots, \ell_T).$$

Percentage-based windows. To normalize across responses of varying length, we define windows as a fraction $\alpha \in (0, 1]$ of the total response:

$$W_{\text{first}}(\alpha) = (\ell_1, \ldots, \ell_{\lfloor \alpha T \rfloor}), \qquad W_{\text{last}}(\alpha) = (\ell_{T-\lfloor \alpha T \rfloor+1}, \ldots, \ell_T).$$

3.3.2 Statistical Aggregation

For each window W, we compute the following summary statistics:

Mean and median.

$$\mu_W = \frac{1}{|W|} \sum_{\ell \in W} \ell.$$

The mean reflects overall token likelihood; the median provides a robust central-tendency estimate less sensitive to outlier tokens.

Minimum and maximum. $\min(W)$ and $\max(W)$ bound the range of token confidence within the segment.

Variance, standard deviation, and range.

$$\sigma^2_W = \frac{1}{|W|} \sum_{\ell \in W} (\ell - \mu_W)^2, \qquad \text{range}_W = \max(W) - \min(W).$$

These statistics quantify the dispersion and volatility of the generation process within a window.

Trajectory slope. We fit a linear regression to the log-probability sequence over the window:

$$\ell_t \approx a\, t + b.$$

The slope coefficient a captures directional trends: $a > 0$ indicates growing confidence across the segment, while $a < 0$ indicates declining confidence.

3.4 LLM-AS-JUDGE META-EVALUATION

Because the Advocate and Skeptic generate open-ended argumentative reasoning rather than discrete labels, their outputs cannot be evaluated with reference-based metrics such as accuracy or n-gram overlap. We therefore introduce a secondary evaluation stage in which a separate language model judges the quality of each agent's reasoning along rubric-based dimensions.

3.4.1 PROMPT RECONSTRUCTION

For each agent response, we reconstruct the complete prompt context the agent originally received, consisting of: (i) the agent's system instructions describing its role and behavioral constraints, (ii) the rubric trait definition, (iii) the essay text, and (iv) the agent's generated response. Supplying this full context enables the evaluator to assess both role adherence and the appropriateness of the evidence used.

3.4.2 EVALUATION DIMENSIONS

The meta-evaluator scores each response along three dimensions:

- **Instruction Following.** Whether the agent maintained its assigned role throughout and avoided prohibited behaviors.
- **Justification Quality.** Whether claims are supported by explicit reasoning that coherently links evidence to conclusions.
- **Evidence Grounding.** Whether the argument references concrete, specific passages from the essay rather than relying on vague or generic statements.

3.4.3 SCORING PROTOCOL

Each dimension is scored on a three-point ordinal scale ($1 = \text{Low},\ 2 = \text{Medium},\ 3 = \text{High}$), assigned independently. The evaluator also raises a **critical issue flag** when the response contains a severe failure that invalidates the reasoning, including hallucinated evidence, major internal contradictions, role-constraint violations, or incoherent output.

We summarize reasoning quality using an aggregate score, $Q_1 = s_{\text{instruction}} + s_{\text{justification}} + s_{\text{evidence}}$.. If a critical failure is detected, the aggregate score is overridden:

$$Q = \begin{cases} 0 & \text{if critical issue is present,} \\ Q_1 & \text{otherwise.} \end{cases}$$

This formulation ensures that responses containing severe reasoning failures are penalized regardless of their dimension-level scores, yielding a composite score in the range $[0, 9]$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. We evaluate our framework on the ASAP[2] dataset, a widely used benchmark of student-written English essays scored by trained human raters against prompt-specific rubrics. Although ASAP comprises eight essay sets, analytic trait-level annotations are available only for Essay Sets 7 and 8; all experiments are therefore conducted on these two sets, which provide multiple independent human ratings per essay at the trait level. Full dataset statistics, rubric descriptions, and label construction details are provided in Appendix A.

Evaluation Metrics. For ordinal evaluation targets—instruction following, justification quality, evidence grounding, and aggregate score—we report Spearman's ρ and Kendall's τ to capture rank-order agreement between confidence proxies and LLM-as-judge scores. For the binary `critical flag`, we report AUROC and point-biserial correlation. To keep results interpretable, we report only the best-performing proxy per target–role combination.

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

Set	Role	Target	Top confidence features (ranked)	Best score
7	Advocate	Aggregate	Full-response median, Final-half median, Full-response mean	0.379
7	Advocate	Instruction	Full-response median, Final-half median, Full-response mean	0.394
7	Advocate	Justification	Full-response median, Final-half median, Full-response mean	0.353
7	Advocate	Evidence	Full-response median, Final-half median, Full-response mean	0.242
7	Advocate	Critical	Max of first 3 tokens, Max of first 5 tokens, Max of first 10 tokens	0.849†
7	Skeptic	Aggregate	Range of first 3 tokens, Full-response slope, Slope over first 30 tokens	0.208
7	Skeptic	Instruction	Range of first 3 tokens, Full-response slope, Slope over first 30 tokens	0.163
7	Skeptic	Evidence	Range of first 3 tokens, Slope over first 30 tokens, Full-response slope	0.114
7	Skeptic	Critical	Mean of first 3 tokens, Min of first 3 tokens, Slope over first 3 tokens	0.638†
8	Advocate	Aggregate	Range of first 3 tokens, Range of first 5 tokens, Std. dev. of first 3 tokens	0.320
8	Advocate	Instruction	Range of first 3 tokens, Std. dev. of first 3 tokens, Variance of first 3 tokens	0.373
8	Advocate	Justification	Range of first 3 tokens, Range of first 5 tokens, Final-half mean	0.284
8	Advocate	Evidence	Range of first 3 tokens, Range of first 5 tokens, Std. dev. of first 5 tokens	0.336
8	Advocate	Critical	Median of first 5 tokens, Range of first 3 tokens, Std. dev. of first 3 tokens	0.759†
8	Skeptic	Aggregate	Range of first 3 tokens, Std. dev. of first 3 tokens, Variance of first 3 tokens	0.196
8	Skeptic	Instruction	Range of first 3 tokens, Std. dev. of first 3 tokens, Variance of first 3 tokens	0.176
8	Skeptic	Justification	Range of first 3 tokens, Std. dev. of first 3 tokens, Variance of first 3 tokens	0.231
8	Skeptic	Evidence	Range of first 3 tokens, Std. dev. of first 3 tokens, Variance of first 3 tokens	0.092
8	Skeptic	Critical	Median of first 5 tokens, Mean of first 3 tokens, Std. dev. of first 3 tokens	0.630†

Table 1: Top-3 confidence features for each role–target pair across Essay Sets 7 and 8. For ordinal targets, the final column reports Spearman correlation; for critical detection, it reports AUROC. Early-k refers to statistics computed over the first k generated tokens, while final-half refers to the last 50% of the response. Full rankings and additional metrics are provided in Appendix B. †AUROC reported for critical detection.

Models. Advocate and Skeptic responses are generated using **GPT-4o-mini**, and meta-evaluation is performed by **GPT-5-mini** instance acting as the LLM-as-judge. To reduce run-to-run variance, the meta-evaluator decodes deterministically, while the Advocate and Skeptic use standard sampling. Token-level log-probabilities are collected during Advocate and Skeptic decoding to compute the confidence proxies described in Section 3.

4.2 Cross-Dataset Analysis

Table 1 summarizes the top-performing confidence features across Essay Sets 7 and 8. While both datasets exhibit a consistent relationship between token-level confidence and judged reasoning quality, the structure of this relationship varies notably across roles and datasets.

For the Advocate, the dominant signal shifts from global to local confidence. In Essay Set 7, the strongest predictors are full-response summaries such as *full-response median* and *final-half median*, which consistently lead across all ordinal targets. In contrast, Essay Set 8 shows a clear transition toward early-generation dispersion, with *range of first 3 tokens* emerging as the top feature across all ordinal targets. This shift suggests that the informativeness of confidence signals depends on dataset characteristics, with some settings favoring globally stable confidence while others emphasize variability at the start of generation. In contrast, the Skeptic exhibits a stable pattern across both datasets. Early-window dispersion features—particularly *range of first 3 tokens*—consistently dominate all ordinal targets, with only minor variation in secondary features such as slope-based measures. Correlation magnitudes are uniformly lower than for the Advocate, indicating a weaker alignment between confidence and judged quality for adversarial reasoning.

A similar pattern holds for critical failure detection. Advocate failures are best captured by sharp early-token signals, such as *max of first 3 tokens* in Set 7 (AUROC $= 0.849$) and *median of first 5 tokens* in Set 8 (AUROC $= 0.759$), suggesting that severe errors manifest as localized confidence spikes or drops early in generation. Skeptic detection performance is both weaker and more stable across datasets (AUROC ≈ 0.63), with early mean- and median-based features performing best.

Taken together, three findings are consistent across both essay sets. First, early-generation signals are broadly informative: even when not dominant (as in Advocate Set 7), they remain among the top-performing features across nearly all role–target pairs. Second, the opening phase of generation is the most diagnostically useful region, aligning with trajectory analyses that show higher variability in early tokens compared to later segments. Third, the Advocate–Skeptic asymmetry is robust:

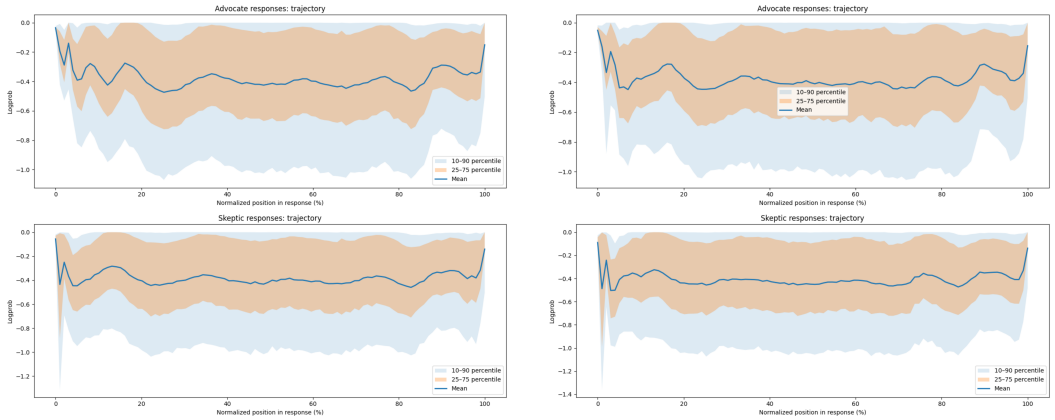

Figure 2: Token-level log-probability trajectories for Advocate and Skeptic responses on Essay Sets 7 (left) and 8 (right). Solid lines denote mean trajectories; shaded regions indicate percentile bands (25–75 and 10–90).

confidence aligns more strongly with supportive reasoning than with adversarial critique, both in ordinal correlations and in critical-failure detection. Full rankings and additional metrics are reported in Appendix B.

4.3 Trajectory Analysis

To understand why early-window features consistently dominate across both datasets, we analyze token-level log-probability trajectories for Advocate and Skeptic responses. Because responses vary in length, each trajectory is normalized to a 0%–100% position scale via interpolation. For each role and dataset, we compute the mean trajectory along with 25–75 and 10–90 percentile bands across responses.

Figure 2 reveals a consistent structural pattern across both essay sets and roles: responses begin with relatively high confidence, followed by a sharp early decline, a prolonged mid-response plateau, and a modest recovery toward the end. Since log-probabilities closer to zero indicate higher confidence, this pattern suggests that initial tokens are easy to predict, uncertainty increases as the model transitions into substantive reasoning, and confidence stabilizes once the response structure is established.

A central observation is that variability is concentrated at the beginning of the response. The percentile bands are widest in the first few tokens, indicating substantial heterogeneity in early-generation behavior: some responses start with stable, high-confidence trajectories, while others exhibit immediate volatility. In contrast, the middle and later portions of the response are comparatively flat and tightly clustered. This explains why early-window dispersion features (e.g., *range of first 3 tokens*) consistently emerge as strong predictors—they capture precisely the region where responses differ most in confidence. Once generation reaches the plateau phase, trajectories become too similar for full-response or late-window features to remain discriminative.

The trajectories also reveal a persistent role asymmetry. Skeptic responses exhibit slightly lower average confidence and wider low-confidence tails throughout generation, particularly in the lower percentile bands. This aligns with the weaker correlations observed for Skeptic features and suggests that adversarial reasoning introduces greater variability in generation dynamics. In contrast, Advocate responses follow more stable trajectories, making confidence a more reliable signal of judged quality.

4.4 Failure Analysis

LLM-as-Judge Meta-Evaluator. Figure 3 shows that the meta-evaluator exhibits strong score concentration across all dimensions, assigning the maximum score (3) in the majority of cases: 76.5% for instruction following, 79.3% for justification quality, and 93.1% for evidence grounding. Despite this overall skew, the three dimensions differ markedly in their ability to discriminate be-

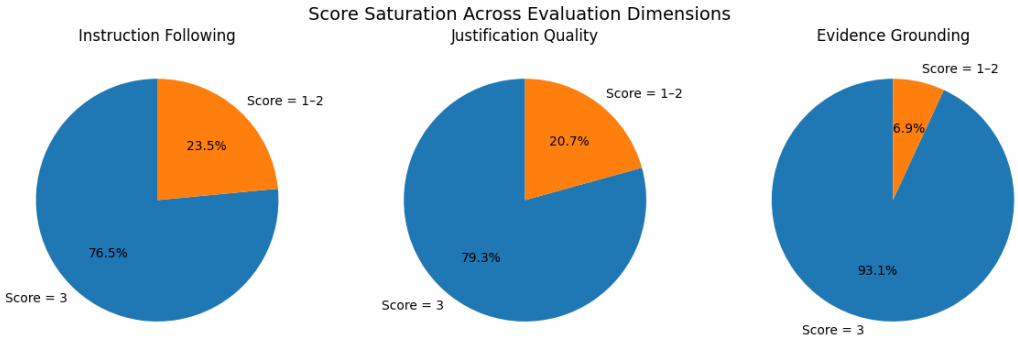

Figure 3: Distribution of LLM-as-judge scores across evaluation dimensions. Evidence grounding is highly saturated at the maximum score, while instruction following and justification quality exhibit greater variability.

tween responses. Evidence grounding is highly saturated and behaves almost as a binary signal, contributing little variation. In contrast, instruction following and justification quality account for most of the observable differences in scores.

Among these, justification quality shows the strongest role asymmetry, with a 13% gap in pass rates between the Advocate and Skeptic. Instruction following captures broader procedural failures across both roles, with the Skeptic penalized more heavily (84.2% vs. 93.6%), primarily due to violations of explicit instructions.

These patterns reflect fundamentally different failure modes across roles. Advocate failures are primarily associated with justification quality, especially through overstatement. The Advocate frequently amplifies weak or incorrect evidence, presenting flawed reasoning as strong or mischaracterizing surface-level features. Importantly, these errors occur along a continuum, ranging from mild exaggeration to clear misrepresentation.

In contrast, Skeptic failures are dominated by instruction-following violations. The most common issue arises from engaging with anonymization placeholders (e.g., `@CAPS1`, `@NUM2`) despite explicit instructions to ignore them. Unlike Advocate errors, these failures are largely binary: the Skeptic either adheres to the procedure or violates it.

This distinction explains the persistent role asymmetry observed in the correlation analysis. Because Advocate errors vary continuously, they induce a broader distribution of scores, enabling stronger alignment with confidence signals (e.g., $\rho = 0.373$ for instruction following). Skeptic errors, by contrast, collapse into near binary outcomes, limiting score variability and compressing rank-based correlations regardless of the underlying signal.

A similar pattern appears in critical failure detection. Advocate failures—often driven by hallucinated or fabricated evidence—produce sharper confidence anomalies, leading to stronger detection performance (e.g., AUROC ≈ 0.76). These errors typically involve claims about nonexistent essay features, introducing low-probability tokens during generation. In contrast, Skeptic failures are predominantly procedural and do not produce comparable confidence deviations, resulting in weaker detection signals (AUROC ≈ 0.63).

5 Limitations and Future Directions

Several limitations bound the scope of these conclusions. First, our experiments are restricted to two essay sets from a single benchmark; it remains to be seen whether the early-token informativeness pattern generalizes to other domains, rubric types, or longer-form generation tasks. Second, all results are obtained with a single model family (GPT-4o-mini for both generation and meta-evaluation), and the confidence patterns we observe may be specific to this model's decoding behavior. Finally, our confidence proxies are derived from token log-probabilities, which are not avail-

able from all model APIs; this limits the practical applicability of the approach to settings where decoding-time probability access is granted.

Several directions follow naturally from this work. Extending the analysis to more essay sets, additional rubric domains, and a wider range of model families would establish how broadly the early-token signal generalizes. Investigating whether confidence-aware decoding strategies, such as selectively resampling responses that exhibit high early-window variance, can improve output quality in practice would translate these findings into an actionable intervention. More broadly, developing evaluation protocols that are sensitive to the structural differences between supportive and adversarial debate roles would strengthen the reliability of multi-agent reasoning systems as a whole.

CONCLUSION

We presented a framework that couples multi-agent debate with LLM-as-judge meta-evaluation to study whether token-level confidence signals can predict the quality of open-ended argumentative reasoning in LLM-based essay scoring. Across both ASAP essay sets, we find that early-window log-probability statistics, particularly dispersion measures over the first few generated tokens, are consistently the strongest predictors of externally judged reasoning quality. This finding is supported both by the correlation analysis and by trajectory-level evidence showing that the opening segment of generation is the most heterogeneous, and therefore the most informative, region of the response. We also identify a stable Advocate–Skeptic asymmetry: confidence proxies correlate more reliably with Advocate reasoning quality than with Skeptic reasoning quality, a difference traceable to the distinct failure modes of each role.

6 ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. 2315294.

REFERENCES

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL https://arxiv.org/abs/2308.07201.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation, 2025. URL https://arxiv.org/abs/2507.21028.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954, 2025. ISSN 1075-2935. doi: https://doi.org/10.1016/j.asw.2025.100954. URL https://www.sciencedirect.com/science/article/pii/S1075293525000418.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

Semire Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), Aug. 2006. URL https://ejournals.bc.edu/index.php/jtla/article/view/1640.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

Avisa Fallah, Ali Keramati, Mohammad Ali Nazari, and Fatemeh Sadat Mirfazeli. 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)*, pp. 365–372, 2024. doi: 10.1109/ICCKE65377.2024.10874775.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

Peiyuan Gong and Jiaxin Mao. Coascore: Chain-of-aspects prompting for nlg evaluation, 2023. URL https://arxiv.org/abs/2312.10355.

Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. Llm multi-agent systems: Challenges and open problems, 2026. URL https://arxiv.org/abs/2402.03578.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.306. URL https://aclanthology.org/2025.findings-acl.306/.

Hawon Jeong, Chaehun Park, Jimin Hong, and Jaegul Choo. Prepair: Pointwise reasoning enhance pairwise evaluating for robust instruction-following assessments, 06 2024.

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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL https://openreview.net/forum?id=nddwJseiiy.

Ali Keramati and Mark Warschauer. Madest: Multi-agent debate essay scoring triangulation, September 2025. URL https://doi.org/10.5281/zenodo.17196206.

Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations, 2024. URL https://arxiv.org/abs/2307.02762.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6723–6737, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.464. URL https://aclanthology.org/2022.acl-long.464/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yixin Liu, Kejian Shi, Alexander Fabbri, Yilun Zhao, PeiFeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. ReIFE: Re-evaluating instruction-following evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12247–12287, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.610. URL https://aclanthology.org/2025.naacl-long.610/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Vaibhav Mavi, Shubh Jaroria, and Weiqi Sun. Self-evaluating llms for multi-step tasks: Stepwise confidence estimation for failure detection, 2025. URL https://arxiv.org/abs/2511.07364.

Austin Pack, Alex Barrett, and Juan Escalante. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234, 2024. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2024.100234. URL https://www.sciencedirect.com/science/article/pii/S2666920X24000353.

Mihir Parmar, Xin Liu, Palash Goyal, Yanfei Chen, Long Le, Swaroop Mishra, Hossein Mobahi, Jindong Gu, Zifeng Wang, Hootan Nakhost, Chitta Baral, Chen-Yu Lee, Tomas Pfister, and Hamid Palangi. Plangen: A multi-agent framework for generating planning and reasoning trajectories for complex problem solving, 2025. URL https://arxiv.org/abs/2502.16111.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024. URL https://arxiv.org/abs/2405.19648.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8352–8370, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.462. URL https://aclanthology.org/2024.naacl-long.462/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

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. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate, 2025. URL https://arxiv.org/abs/2509.05396.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets, 2024. URL https://arxiv.org/abs/2307.10928.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following, 2024. URL https://arxiv.org/abs/2310.07641.

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. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Data and Preprocessing

A.1 Dataset Selection

We conduct our analysis on the ASAP (Automated Student Assessment Prize) dataset, a standard benchmark for essay scoring. The dataset consists of eight prompt-specific essay sets with varying genres, scoring rubrics, and grade levels.

Our study focuses exclusively on **Essay Sets 7 and 8**, as these are the only subsets that provide *trait-level annotations* from multiple human raters. This property is essential for our setup, since we evaluate reasoning quality at the level of individual rubric traits rather than holistic scores. The remaining essay sets are not used, as they only provide single aggregated scores and therefore do not support fine-grained evaluation.

A.2 Relevant Dataset Characteristics

Table 2 summarizes the key properties of the two essay sets used in our experiments.

Table 2: Summary of the ASAP subsets used in this work.

Set	Grade	# Essays	# Traits
7	7	1,569	4
8	10	723	6

These two sets differ in both rubric complexity and score ranges, providing a useful testbed for analyzing how confidence signals interact with reasoning quality under different evaluation conditions.

A.3 Prompt Context

Each essay is written in response to a fixed prompt. For completeness, we include simplified versions of the prompts used in the selected sets:

Role	Target	Feature	Spearman	Kendall	AUROC	PB
Advocate	Aggregate	Full-response median	0.379	0.295	–	–
Advocate		Final-half median	0.363	0.281	–	–
Advocate		Full-response mean	0.353	0.273	–	–
Advocate	Evidence	Full-response median	0.242	0.196	–	–
Advocate		Final-half median	0.237	0.193	–	–
Advocate		Full-response mean	0.237	0.192	–	–
Advocate	Instruction	Full-response median	0.394	0.317	–	–
Advocate		Final-half median	0.383	0.308	–	–
Advocate		Full-response mean	0.379	0.305	–	–
Advocate	Justification	Full-response median	0.353	0.284	–	–
Advocate		Final-half median	0.332	0.267	–	–
Advocate		Full-response mean	0.323	0.260	–	–
Advocate	Critical	Max of first 3 tokens	–	–	0.849	0.261
Advocate		Max of first 5 tokens	–	–	0.849	0.261
Advocate		Max of first 10 tokens	–	–	0.830	0.253
Skeptic	Aggregate	Range of first 3 tokens	0.208	0.166	–	–
Skeptic		Full-response slope	0.125	0.100	–	–
Skeptic		Slope over first 30 tokens	0.121	0.097	–	–
Skeptic	Evidence	Range of first 3 tokens	0.114	0.093	–	–
Skeptic		Slope over first 30 tokens	0.067	0.055	–	–
Skeptic		Full-response slope	0.060	0.049	–	–
Skeptic	Instruction	Range of first 3 tokens	0.163	0.130	–	–
Skeptic		Full-response slope	0.102	0.081	–	–
Skeptic		Slope over first 30 tokens	0.098	0.078	–	–
Skeptic	Critical	Mean of first 3 tokens	–	–	0.638	0.194
Skeptic		Min of first 3 tokens	–	–	0.635	0.177
Skeptic		Slope over first 3 tokens	–	–	0.634	0.180

Table 3: Top-3 confidence features for Essay Set 7.

Set 7. Students are asked to write a story about patience, either from personal experience or imagination.

Set 8. Students are asked to write a true story in which laughter plays an important role.

These prompts define the context in which both the debate agents and the evaluator operate.

A.4 Text Handling

The ASAP essays are transcriptions of handwritten student responses. We use the text *as provided*, without any normalization or correction. In particular, spelling errors, grammatical inconsistencies, and informal structures are preserved.

This choice is important because the evaluation criteria (e.g., evidence grounding and justification) depend on the original textual content, and preprocessing could alter signals that are relevant to both the agents and the evaluator.

B Top-3 Confidence Feature Rankings by Dataset

Tables 3 and 4 report the top three confidence features for each role–target pair on Essay Sets 7 and 8. For ordinal targets, features are ranked by Spearman correlation with the LLM-as-judge score, with Kendall's τ reported as a secondary measure. For critical failure detection, features are ranked by AUROC, with point-biserial correlation (PB) included for completeness.

To improve interpretability, we present features using descriptive names rather than implementation-specific identifiers. Early-k refers to statistics computed over the first k generated tokens, while final-half refers to the last 50% of the response. Full-response features are computed over the entire generated sequence.

Role	Target	Feature	Spearman	Kendall	AUROC	PB
Advocate	Aggregate	Range of first 3 tokens	0.320	0.248	–	–
Advocate		Range of first 5 tokens	0.274	0.214	–	–
Advocate		Std. dev. of first 3 tokens	0.267	0.205	–	–
Advocate	Evidence	Range of first 3 tokens	0.336	0.272	–	–
Advocate		Range of first 5 tokens	0.326	0.263	–	–
Advocate		Std. dev. of first 5 tokens	0.305	0.247	–	–
Advocate	Instruction	Range of first 3 tokens	0.373	0.302	–	–
Advocate		Std. dev. of first 3 tokens	0.338	0.274	–	–
Advocate		Variance of first 3 tokens	0.338	0.274	–	–
Advocate	Justification	Range of first 3 tokens	0.284	0.228	–	–
Advocate		Range of first 5 tokens	0.284	0.228	–	–
Advocate		Final-half mean	0.273	0.221	–	–
Advocate	Critical	Median of first 5 tokens	–	–	0.759	0.285
Advocate		Range of first 3 tokens	–	–	0.754	-0.287
Advocate		Std. dev. of first 3 tokens	–	–	0.736	-0.286
Skeptic	Aggregate	Range of first 3 tokens	0.196	0.157	–	–
Skeptic		Std. dev. of first 3 tokens	0.195	0.157	–	–
Skeptic		Variance of first 3 tokens	0.195	0.157	–	–
Skeptic	Evidence	Range of first 3 tokens	0.092	0.075	–	–
Skeptic		Std. dev. of first 3 tokens	0.090	0.074	–	–
Skeptic		Variance of first 3 tokens	0.090	0.074	–	–
Skeptic	Instruction	Range of first 3 tokens	0.176	0.140	–	–
Skeptic		Std. dev. of first 3 tokens	0.174	0.139	–	–
Skeptic		Variance of first 3 tokens	0.174	0.139	–	–
Skeptic	Justification	Range of first 3 tokens	0.231	0.189	–	–
Skeptic		Std. dev. of first 3 tokens	0.230	0.188	–	–
Skeptic		Variance of first 3 tokens	0.230	0.188	–	–
Skeptic	Critical	Median of first 5 tokens	–	–	0.630	0.151
Skeptic		Mean of first 3 tokens	–	–	0.630	0.156
Skeptic		Std. dev. of first 3 tokens	–	–	0.617	-0.138

Table 4: Top-3 confidence features for Essay Set 8.

C Agent Prompt Templates

This section provides the system instructions used to define the roles of the agents in the debate framework. All prompts are implemented as template files and rendered at runtime using a shared context dictionary. The context includes the rubric trait name, the full rubric definition serialized as JSON, the essay text, the essay prompt or question, and the valid scoring range for the trait.

In addition, the Synthesizer-Judge receives the debate transcript produced by the Advocate and Skeptic agents. These prompts establish strict role boundaries to ensure that each agent performs a specialized function in the debate process.

The prompt configurations were developed through iterative experimentation, including pilot runs and refinements designed to enforce role adherence, maintain output consistency, and reduce undesired behaviors such as assigning scores prematurely or mixing multiple rubric traits in a single argument.

For transparency and reproducibility, we provide the exact system instructions used to define each agent role.

C.1 Advocate Agent

The Advocate agent is responsible for presenting arguments that highlight the strengths of the essay with respect to a single rubric trait. The agent receives the essay text and the rubric definition and produces an evidence-based argument explaining how the essay satisfies the expectations of the trait.

The Advocate is explicitly instructed to focus only on positive aspects of the essay and to avoid assigning a score or discussing weaknesses. The agent may reference specific passages from the essay to support its claims.

Advocate Agent System Prompt

You are an Advocate Agent participating in a multi-agent debate system for essay evaluation.
Your task is to analyze the essay and highlight strengths that demonstrate how the essay satisfies the rubric expectations for the trait "$TRAIT_NAME".
Focus exclusively on positive evidence from the essay. Provide detailed reasoning supported by specific excerpts or paraphrased examples from the essay.
Do not assign a score and do not critique weaknesses. Your role is solely to present arguments supporting the essay's strengths with respect to the specified rubric trait.
Anonymization $ANON_CONTEXT

Figure 4: System instructions for the Advocate agent.

C.2 Skeptic Agent

The Skeptic agent provides a counterargument to the Advocate by identifying weaknesses or limitations in the essay relative to the same rubric trait.

The Skeptic receives both the essay text and the Advocate's argument and produces a critique that challenges the strengths presented or highlights aspects where the essay fails to meet the rubric expectations.

The agent is instructed not to assign a score and to focus exclusively on critical analysis.

Skeptic Agent System Prompt

You are a Skeptic Agent participating in a multi-agent debate system for essay evaluation.
Your task is to critically analyze the essay and identify weaknesses related to the rubric trait "$TRAIT_NAME".
Provide detailed critiques supported by specific references to the essay text. Focus only on identifying shortcomings or areas where the essay does not fully satisfy the rubric expectations.
Do not assign a score and do not discuss strengths. Your role is to challenge the essay's performance with respect to the specified rubric trait.
Anonymization $ANON_CONTEXT

Figure 5: System instructions for the Skeptic agent.

C.3 Synthesizer-Judge Agent

The Synthesizer-Judge serves as the final decision-maker in the debate process. This agent reads the arguments produced by the Advocate and Skeptic and determines a final score for the rubric trait.

The agent synthesizes the competing arguments and evaluates them against the rubric definition before assigning a score within the permitted range.

C.4 LLM-as-Judge Meta-Evaluator

The meta-evaluator agent is responsible for assessing the quality of each Advocate and Skeptic response. The agent receives the original system prompt given to agents, context including the essay text and rubric, and the agent's response.

The meta-evaluator scores each response along three dimensions: instruction following, justification quality, and evidence grounding, each on a three-point ordinal scale (1 = Weak, 2 = Adequate, 3 = Strong). The evaluator also flags a critical issue when the response contains hallucinated evidence, severe deviation or violation of instructions, or internal contradictions.

Synthesizer-Judge Agent System Prompt

You are the Synthesizer-Judge in a multi-agent debate system for essay evaluation.
Your task is to read the debate transcript between the Advocate and Skeptic agents regarding the rubric trait "$TRAIT_NAME".
Carefully consider the arguments presented by both agents and evaluate them against the rubric expectations.
Based on the combined evidence, assign a final integer score between $MIN_POINTS and $MAX_POINTS for the essay on this rubric trait.
Anonymization $ANON_CONTEXT

Figure 6: System instructions for the Synthesizer-Judge agent.

The meta-evaluator is explicitly instructed to evaluate only the agent's reasoning quality and role adherence. It does not assess the essay itself or judge whether the score being argued is correct.

The complete system instructions and output schema are provided in Figures 4- 8.

Meta-Evaluation System Prompt

You are a meta-evaluator assessing the quality of an AI agent's response in a multi-agent essay-scoring debate system. Your goal is to evaluate how well the agent performs its role.
Important: Do NOT evaluate the essay itself. Do NOT judge whether the essay score being argued is correct. Evaluate ONLY the agent's reasoning quality, role adherence, and use of essay evidence.
You will receive: (1) the agent's system prompt, (2) the task prompt given to the agent, and (3) the agent's response. Evaluate the response across three dimensions using the full range of the scale: **1 = Weak**, **2 = Adequate**, **3 = Strong**. Avoid defaulting to the middle score. Evaluate each dimension independently.

Dimension 1 – Instruction Following. 3: Fully maintains role; completes all task components; no deviations. 2: Generally follows instructions with a minor omission or slight deviation. 1: Major or multiple deviations; neglects important instructions.

Dimension 2 – Justification Quality. 3: Multiple claims with clear reasoning; claim $\rightarrow$ explanation $\rightarrow$ implication structure. 2: At least one supported claim; reasoning understandable but shallow or repetitive. 1: Minimal or vague reasoning; assertions without explanation.

Dimension 3 – Evidence Grounding. 3: Two or more precise references including quotes or detailed paraphrases. 2: One clear identifiable reference; other claims rely on general statements. 1: Evidence vague, indirect, or missing.

Critical Issues Flag. Set `critical_flag = 1` if any of the following occur: hallucinated essay evidence, severe internal contradiction, explicit instruction violation, or nonsensical output. Otherwise `critical_flag = 0`.

Output: Return only a JSON object with fields `instruction_following`, `justification_quality`, `evidence_grounding`, `critical_flag`, `critical_issues_description`, and `reasoning` (2–3 sentences).

Figure 7: System instructions for the Meta-Evaluator agent.

Meta-Evaluation Task Prompt

```
# Agent Being Evaluated: {AGENT_TYPE}

# Agent's System Prompt (Instructions Given to the Agent)
{AGENT_SYSTEM_PROMPT}

# Agent's User Prompt (Task Context Given to the Agent)
{AGENT_USER_PROMPT}

# Agent's Actual Response
{AGENT_RESPONSE}
```

Now evaluate this agent's response. Use the 3-point scale (1=Low, 2=Medium, 3=High) for each scored dimension and set the `critical_flag` to 0 or 1. Output ONLY the JSON object.

Figure 8: Task prompt provided to the Meta-Evaluator agent.