MAPLE: Multi-Aspect Panels of LLM Evaluators for Open-Ended Questions

Anonymous ACL submission

Abstract

LLM-as-a-Judge, which uses LLMs to evaluate responses to open-ended questions, has seen significant growth in recent years. It has been adopted as a scalable alternative to manual human evaluation, such as crowdsourcing, which is often time-consuming and costly. However, the discrepancy between LLM-generated evaluations and human evaluations remains a critical problem in this field. To bridge this gap, we propose Multi-Aspect Panels of LLM Evaluators (MAPLE), a framework that orchestrates evaluations across multiple criteria using multiple LLMs. MAPLE integrates criterion-wise pairwise evaluations from multiple LLMs by estimating the importance of criteria and the reliability of individual evaluators. We conduct experiments with both open-source and closed-source models. Our results demonstrate that MAPLE achieves superior alignment with human evaluations compared to baselines, highlighting the importance of employing multi-agent and multi-criteria evaluation strategies.

1 Introduction

Preference-based post-training methods such as Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2023; Song et al., 2024; Zeng et al., 2024) have been central to aligning Large Language Models (LLMs) with human preferences. However, these methods depend on large datasets annotated with human preferences, and human labeling itself is inherently costly and slow. Given this limitation, recent work has explored the use of LLMs as scalable evaluators to score open-ended outputs or generate preference labels (Ding et al., 2023; Zheng et al., 2023; Chen et al., 2024; Liu et al., 2023). This line of work is commonly referred to as LLM-as-a-Judge. Prior studies have also reported that LLM evaluations can be competitive with crowdworker annotations (Gilardi et al., 2023). Despite this promise, evaluations produced by LLMs often do not align with human labels, with prior work pointing to factors such as systematic biases and sensitivity to the chosen evaluator model and task (Ye et al., 2025; Chen et al., 2024; Wang et al., 2024; Li et al., 2024).

To improve human alignment, prior studies have explored two directions: aggregating evaluations from multiple LLM agents (Qian et al., 2025; Chan et al., 2024) and decomposing evaluation into multiple explicit criteria (Lu et al., 2024; Jordán et al., 2026). While both directions improve human alignment, these approaches still have limitations. In many settings, multi-agent methods lack clear evaluation criteria and assume all models are equally reliable, which can hurt accuracy and make it unclear which aspects contribute to the final evaluation, especially when low-quality models are involved. On the other hand, multi-criteria methods often use only one LLM, so their results depend heavily on that model and its biases.

Therefore, we combine both methods in a single framework. We propose Multi-Aspect Panels of LLM Evaluators (MAPLE), a framework that takes a holistic rubric and target items as input. MAPLE breaks the rubric down into concrete evaluation criteria and collects criterion-wise pairwise comparisons from multiple LLM evaluators using the Analytic Hierarchy Process (AHP), a structured decision-making method. These comparisons are then aggregated using a reliability-aware statistical model that extends Crowd-BT (Chen et al., 2013) based on the Bradley–Terry model (Bradley and Terry, 1952).

We evaluate MAPLE on an essay scoring task and show that it achieves higher agreement with human scores than baselines that rely solely on multi-agent aggregation or multi-criteria prompting. MAPLE further distinguishes reliable from unreliable evaluators and remains robust even when biased or low-performing LLMs are included, pro-

viding a practical way to obtain more human-aligned and stable evaluations without prior knowledge of which single LLM is the best evaluator.

Our contributions are summarized as follows:

- We propose MAPLE, an LLM-as-a-Judge framework that combines multi-agent aggregation and multi-criteria evaluation.
- Through experiments on an essay scoring task, we show that MAPLE achieves higher agreement with human scores than baselines that rely on either multi-agent aggregation or multi-criteria prompting alone.
- Our results indicate that MAPLE attains strong average performance over individual evaluators and remains robust to low-performing or biased LLMs.

2 Related Work

Chan et al. (2024) proposed ChatEval, a framework for opinion integration that assigns distinct roles to multiple LLMs and engages them in discussion. They explored various discussion strategies, showing that diverse roles improve alignment with human evaluations. Conversely, Qian et al. (2025) argued that debate-based methods are inefficient. They introduced CollabEval for collaborative opinion integration among multiple agents, and showed that it maintains robust performance even when individual models struggle.

Jordán et al. (2026) proposed MAGIC, a framework that uses criteria-specific prompts with an orchestrator integrating results across criteria. In essay scoring, the authors emphasized the importance of multi-criteria assessment. Additionally, Lu et al. (2024) introduced a method for multi-criteria pairwise comparisons using the Analytic Hierarchy Process (AHP). By comparing against criterion-free pairwise comparisons, they demonstrated the effectiveness of multi-criteria decision-making.

In summary, existing LLM-as-a-Judge research improves alignment either by aggregating multiple judges or by using explicit evaluation criteria, but these two approaches are rarely combined in a single framework. Multi-agent methods often leave evaluation criteria implicit and may overweight unreliable judges, while multi-criteria methods are frequently run with a single judge and thus remain sensitive to evaluator choice. MAPLE addresses this gap by jointly aggregating criterion-wise comparisons from a multi-model panel with reliability-aware statistical modeling.

3 Preliminaries

3.1 Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP), proposed by Saaty (1987, 2004), is a decision-making framework that evaluates multiple candidates based on multiple criteria. AHP ranks candidates by integrating pairwise comparison results under multiple criteria, while also comparing criteria to estimate their relative importance.

In this study, we adopt this pairwise comparison process. However, for the aggregation step, we employ an extension of Crowd-BT, which is described in the following section.

3.2 Crowd-BT

The Bradley-Terry model (1952) estimates the scores of candidates based on the results of pairwise comparisons. The model assumes that each candidate o_i possesses a latent score w_i. The probability that candidate o_i is judged superior to candidate o_j is defined as

$$P(o_i \succ o_j) = \frac{e^{w_i}}{e^{w_i} + e^{w_j}}. \tag{1}$$

The parameter w_i is estimated via Maximum Likelihood Estimation (MLE) using the pairwise comparison data.

Crowd-BT (Chen et al., 2013) is an extension of the Bradley-Terry model designed to aggregate opinions from multiple annotators. When aggregating multiple opinions, a naive approach is to determine the outcome of each pairwise comparison by majority vote. However, since annotators vary in their ability and diligence, assigning equal weight to all participants is suboptimal. To address this, Crowd-BT introduces a reliability parameter η_k for each participant k. The probability that participant k judges candidate o_i as superior to o_j is defined as

$$P(o_i \succ_k o_j) = \eta_k \frac{e^{w_i}}{e^{w_i} + e^{w_j}} + (1 - \eta_k) \frac{e^{w_j}}{e^{w_i} + e^{w_j}}. \tag{2}$$

The parameters w_i and η_k are estimated from the comparison results of all participants using optimization techniques such as gradient descent.

In this study, we extend Crowd-BT to a multi-criteria setting to facilitate the aggregation of comparisons within the AHP framework.

4 Proposed Framework: MAPLE

Our proposed framework, MAPLE, accepts a predefined holistic rubric and the target texts as input,

outputting scores for each target. The framework comprises three distinct steps: Criteria Generation, Comparison, and Aggregation.

Step 1: Criteria Generation A single LLM specialized in criteria generation formulates evaluation criteria based on the provided holistic rubric and background information regarding the target texts. The Generator produces 4–6 analytic criteria with level descriptors; details are provided in Appendix C.

Step 2: Comparison (AHP) We compare target texts and criteria using an AHP-like method with the criteria from Step 1. Multiple LLMs perform pairwise comparisons in two settings: (1) comparing all text pairs under each criterion, and (2) comparing all criterion pairs to assess relative importance. Ties are not permitted; models must strictly choose the superior option.

Step 3: Aggregation We integrate the comparison results collected in Step 2 using an extension of the Crowd-BT method. We introduce the following parameters: the weight w_c of each criterion c; the score $\theta_{i,c}$ of target o_i under criterion c; and the reliability η_k of LLM k. We model the probability that LLM k judges criterion c as superior to criterion c' as

$$P(c \succ_k c') = \eta_k \frac{e^{w_c}}{e^{w_c} + e^{w_{c'}}} + (1 - \eta_k) \frac{e^{w_{c'}}}{e^{w_c} + e^{w_{c'}}}, \tag{3}$$

and the probability that LLM k judges target o_i as superior to target o_j under criterion c as

$$P(o_i \succ_k o_j | c) = \eta_k \frac{e^{\theta_{i,c}}}{e^{\theta_{i,c}} + e^{\theta_{j,c}}} + (1 - \eta_k) \frac{e^{\theta_{j,c}}}{e^{\theta_{i,c}} + e^{\theta_{j,c}}}. \tag{4}$$

Let $y_{c,c',k}$ represent the result of the comparison between criteria ($y_{c,c',k} = 1$ if LLM k judges c to be more important than c', and 0 otherwise), and let $y_{i,j,k,c}$ represent the result of the comparison between targets ($y_{i,j,k,c} = 1$ if LLM k judges o_i to be better than o_j under criterion c, and 0 otherwise). The loss function $L(\eta, \theta, \mathbf{w})$ is defined as

$$\begin{aligned} L(\eta, \theta, \mathbf{w}) = -\sum_k [& \sum_{i,j,c} \{ y_{i,j,k,c} \log P(o_i \succ_k o_j | c) \\ & + (1 - y_{i,j,k,c}) \log P(o_j \succ_k o_i | c) \} \\ & + \sum_{c,c'} \{ y_{c,c',k} \log P(c \succ_k c') \\ & + (1 - y_{c,c',k}) \log P(c' \succ_k c) \}]. \end{aligned} \tag{5}$$

We estimate the parameters $\mathbf{w}$, θ, and η by minimizing this loss function. The final score is determined by calculating the weighted sum of the target scores $\theta_{i,c}$ under each criterion, based on the criterion weights w_c. We apply the softmax to the weights so that they are positive and sum to one. The final score s_i for target o_i is calculated as

$$s_i = \sum_c \frac{e^{w_c}}{\sum_{c'} e^{w_{c'}}} \cdot \theta_{i,c}. \tag{6}$$

5 Experiments

We investigate the following research questions (RQs) through experiments:

RQ1: Can MAPLE correctly distinguish between reliable and unreliable models?

RQ2: Can MAPLE accurately score responses to open-ended questions?

RQ3: Is MAPLE robust against LLMs that exhibit strong bias or possess inferior capabilities?

To address RQ1, we use synthetic data to examine the aggregation behavior in Step 3. For RQ2, we compare the proposed framework with multiple baselines. For RQ3, we evaluate performance when highly biased or low-performing models are introduced as Evaluators.

5.1 RQ1 : Can MAPLE correctly distinguish between reliable and unreliable models?

We evaluate whether MAPLE can accurately distinguish between reliable and unreliable models and effectively integrate their predictions during the aggregation step using synthetic data.

5.1.1 Settings

Datasets We generate a synthetic dataset consisting of 50 target items and 5 grading criteria. Each criterion is assigned a distinct score from 1 to 5, and under each criterion, each target item is assigned a distinct score from 1 to 50, ensuring a strict ranking without ties. Based on these scores, we define the ground-truth pairwise comparisons for criteria and, under each criterion, for items.

Models We employ models with artificially controlled accuracy rates. Here, "accuracy" means that a model outputs the reverse of the ground-truth pairwise comparison at a rate of $1 -$ accuracy. For example, a model with 70% accuracy outputs the opposite result for a random 30% of the data. We conduct the experiment under the three settings:

	Consistency (CI)		
Acc. Range	**Reliability**	**Criteria**	**Targets**
60% – 100%	1.00	1.00	0.997
10% – 100%	1.00	1.00	0.998
0% – 90%	0.00	0.00	0.544

Table 1: Results of Experiments for RQ1. We report the Concordance Index (CI) for the estimated reliability weights, criterion scores, and final integrated target scores across three different accuracy settings.

- 5 models with accuracies from 60% to 100% in 10% increments.
- 10 models with accuracies from 10% to 100% in 10% increments.
- 10 models with accuracies from 0% to 90% in 10% increments.

Metrics We evaluate the results based on the following three aspects:

- Consistency between the estimated reliability values and the actual accuracy rates.
- Consistency between the magnitude of true criterion scores and estimated criterion scores.
- Consistency between the magnitude of integrated target scores.

Ranking consistency is measured using the Concordance Index (CI) defined as

$$\mathbf{CI}(f, g) = \frac{\sum_{i,j} \mathbb{I}(f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j))\mathbb{I}(g(\boldsymbol{x}_i) > g(\boldsymbol{x}_j))}{\sum_{i,j} \mathbb{I}(g(\boldsymbol{x}_i) > g(\boldsymbol{x}_j))}, \tag{7}$$

where x denotes the dataset, f represents the method being evaluated, g represents the ground truth, and $\mathbb{I}$ is the indicator function, which equals 1 if the input condition is true and 0 otherwise.

Hyperparameters We minimize eq. (5) using gradient descent with 5,000 epochs and a learning rate of 0.01, which are used for all experiments.

5.1.2 Results

Table 1 presents the results of Experiments for RQ1. For model accuracies of 60%–100% and 10%–100%, MAPLE correctly predicts the rank order of model reliability and criterion weights. Furthermore, Table 2 shows that the estimated reliability values closely match ground-truth accuracies, indicating that MAPLE can distinguish reliable from unreliable models.

Model Acc.	**60%–100%**	**10%–100%**	**0%–90%**
1.0 (100%)	1.000	1.000	–
0.9 (90%)	0.909	0.908	0.092
0.8 (80%)	0.810	0.809	0.191
0.7 (70%)	0.711	0.710	0.291
0.6 (60%)	0.605	0.604	0.396
0.5 (50%)	–	0.502	0.498
0.4 (40%)	–	0.398	0.603
0.3 (30%)	–	0.292	0.708
0.2 (20%)	–	0.191	0.809
0.1 (10%)	–	0.092	0.908
0.0 (0%)	–	–	1.000

Table 2: Comparison of final η values by model accuracy across three experimental settings. Cells with "–" indicate the model is not included in that setting.

In the 0%–90% setting, performance drops substantially, and the method fails to accurately estimate model reliability or criterion weights. This occurs because the aggregation is effectively driven by majority voting among evaluators. Without ground truth, judges that agree with the majority are given higher weight, while those that disagree are given lower weight. If malicious models form the majority, the resulting consensus becomes misleading and leads to poor performance. In practice, such a scenario is unlikely in LLM-as-a-Judge settings, and we therefore regard this as a minor limitation.

5.2 RQ2 : Can MAPLE accurately score responses to open-ended questions?

We evaluate the performance of MAPLE on an essay scoring task to demonstrate its effectiveness.

5.2.1 Settings

Dataset We use the ASAP 2.0 dataset (Crossley et al., 2025). This dataset contains essays across seven themes (prompts), written by US students in grades 6 through 10. Each entry includes the essay, student information, and a score (ranging from 1 to 6) assigned by experts based on a specific rubric (see Appendix A). For evaluation, we randomly sample 50 essays per theme. We repeat this sampling and evaluation process 10 times and report the average performance.

Models MAPLE uses LLMs in two roles: a Generator for criterion creation and an Evaluator for assessment. We use GPT-5.2 Pro as the Generator. Evaluators include both open- and closed-source models. The open-source models

are Llama-3-8B-instruct (Grattafiori et al., 2024), Ministral-8B (Mistral AI Team, 2024), Qwen-2.5-7B-instruct (Yang et al., 2024), Llama-3.3-70B-instruct (Grattafiori et al., 2024), and Qwen-3-32B (Yang et al., 2025), while the closed-source models are GPT-5.1, GPT-4o-mini, Gemini-2.5-flash, Gemini-2.5-flash-lite, and Claude-3.5-haiku. Prompts are provided in Appendix C.

Baselines We compare our proposed method against the two multi-model methods and the two single-model methods:

- **Multi-model baselines**
  - **Crowd-BT**: Multiple Evaluator LLMs perform holistic pairwise comparisons of essays (without specific criteria), and the results are aggregated using Crowd-BT.
  - **Majority Vote**: Pairwise outcomes are determined by majority voting across all criteria and evaluators, then aggregated using the Bradley–Terry model.
- **Single-model baselines**
  - **Pairwise (single model)**: A single Evaluator LLM performs holistic pairwise comparisons of essays (without specific criteria), and the results are aggregated using the Bradley-Terry model.
  - **Proposed framework (single model)**: The proposed MAPLE framework is executed using a single evaluator LLM.

Metrics We compare performance with the baselines using the Concordance Index (eq. (7)).

Hyperparameters For aggregation, we use the same hyperparameters as in Experiments for RQ1. Regarding LLM generation parameters, the detailed settings are provided in Appendix B.

5.2.2 Results

Comparison with multi-model baselines Table 3 presents the comparison results between MAPLE, Crowd-BT, and Majority Vote. MAPLE outperforms both Crowd-BT and Majority Vote across all themes. The superiority of MAPLE over Crowd-BT suggests the distinct advantage of multi-criteria evaluation. Furthermore, outperforming Majority Vote indicates that, given the variability in LLM capabilities, weighing opinions based on individual reliability yields better prediction accuracy than treating all models equally.

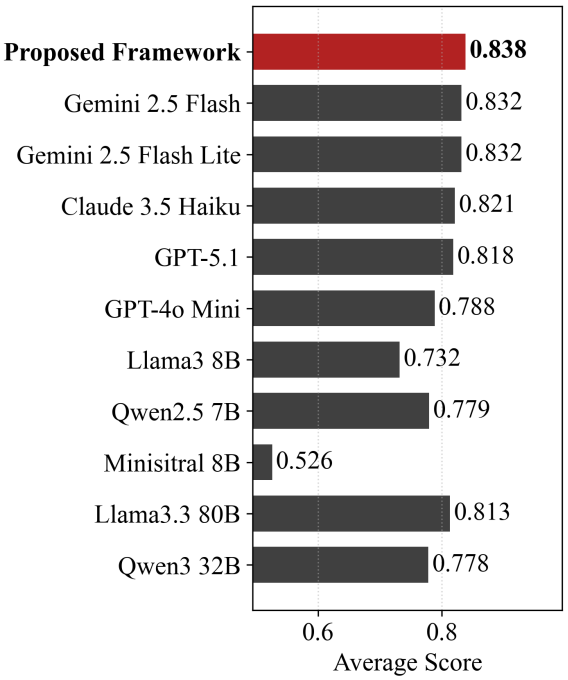

Figure 1: Performance comparison between the proposed framework (using all models) and individual models. Values represent the average Concordance Index across all themes and iterations. The proposed framework outperforms all single-model configurations.

Comparison with single-model baselines Figure 1 shows the average performance (CI) of each LLM when running MAPLE as a single model, while Figure 2 displays the average performance (CI) of individual LLMs (Standard BT). Both figures represent the average performance (CI) across all themes and sampling iterations. Additionally, Figure 3 breaks down the single-model MAPLE performance (CI) averaged by theme. Furthermore, Figure 4 presents a histogram illustrating the rank of the full-model MAPLE performance relative to single-model performance. Specifically, we rank the full-model score against single-model scores for each theme and sampling iteration to show the distribution of rankings.

Figures 1 and 2 reveal that the average performance (CI) of the proposed method (using all models) exceeds that of single-model pairwise comparisons or single-model MAPLE runs. On the other hand, Figure 3 shows that for certain themes, individual models often outperform the full-model ensemble. This observation is corroborated by Figure 4, which shows that the full-model performance rarely ranks first and most frequently ranks second, third, or fourth. As discussed in Experiments for RQ1, this occurs because the aggregation funda-

Prompt	MAPLE (Ours)	Crowd-BT	Majority Vote
A Cowboy Who Rode the Waves	**0.794 ± 0.061**	0.768 ± 0.054	0.765 ± 0.062
Car-free cities	**0.832 ± 0.059**	0.792 ± 0.063	0.790 ± 0.062
The electoral college	**0.880 ± 0.032**	0.870 ± 0.032	0.847 ± 0.037
Driverless cars	**0.833 ± 0.049**	0.832 ± 0.073	0.807 ± 0.074
Exploring Venus	**0.840 ± 0.050**	0.829 ± 0.040	0.823 ± 0.034
Facial action coding system	**0.876 ± 0.026**	0.853 ± 0.040	0.850 ± 0.041
The Face on Mars	**0.810 ± 0.047**	0.807 ± 0.048	0.790 ± 0.051

Table 3: Performance comparison of the proposed framework (MAPLE) and the baselines. MAPLE outperforms the baselines across all themes.

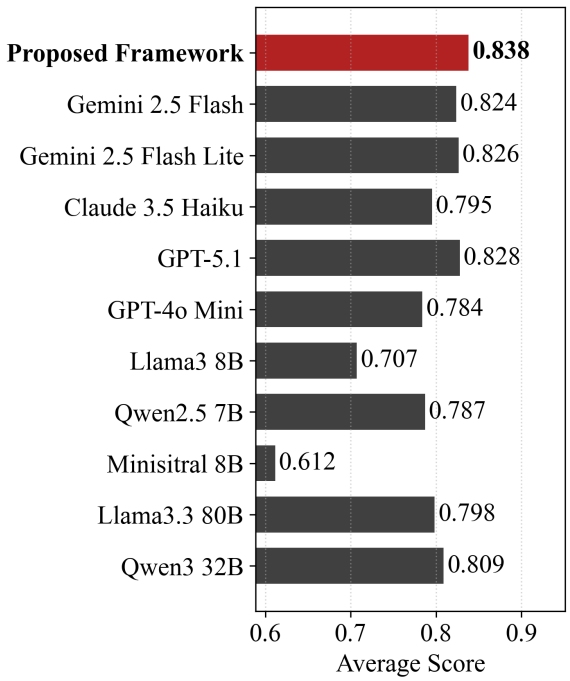

Figure 2: Performance comparison between the proposed framework and pairwise (single model). Values represent the average Concordance Index across all themes and iterations. The proposed framework using all models outperforms all single-model configurations.

mentally favors a majority-vote mechanism, meaning it rarely surpasses the absolute best model in the pool.

The fact that the proposed method yields the best performance on average, despite occasionally falling behind single models in specific instances, stems from the fact that LLM capabilities are not absolute but vary by task. As seen in Figure 3, the top-performing model differs depending on the theme. This highlights that even within the narrow task of essay scoring, the optimal model cannot be determined without empirical testing.

In scenarios where the best model cannot be identified in advance, employing MAPLE ensures consistently high-ranking performance. This stability constitutes a key advantage of our framework.

5.3 RQ3 : Is MAPLE robust against LLMs that exhibit strong bias or possess inferior capabilities?

We assess MAPLE's robustness to biased or low-capability LLMs. LLMs exhibit inherent biases such as position bias (Chen et al., 2024), with susceptibility varying across models (Ye et al., 2025), creating a risk of biased evaluators. Moreover, as shown in Experiments for RQ2, task proficiency is difficult to determine in advance, potentially leading to the selection of low-performing evaluators. We therefore test whether MAPLE maintains robust performance under such adverse conditions.

5.3.1 Settings

We examine how the performance of MAPLE changes when the following models are introduced as Evaluators into the experimental setup for RQ2:

- **Random Predictor**: A model that predicts the outcome of each comparison at random.
- **Position-1-Biased Model**: A model that consistently predicts the option presented first (Position 1) in the prompt as superior.
- **Position-2-Biased Model**: A model that consistently predicts the option presented second (Position 2) in the prompt as superior.

We investigate performance degradation by adding 1–5 instances of a biased model to the evaluator ensemble used for RQ2. Framework performance is measured using the average Concordance Index (eq. (7)) across all themes and iterations under identical conditions.

5.3.2 Results

Figure 5 presents the results of Experiments for RQ3. The framework remains robust with negligible degradation when up to four adversarial models are added. While adding five Random Predic-

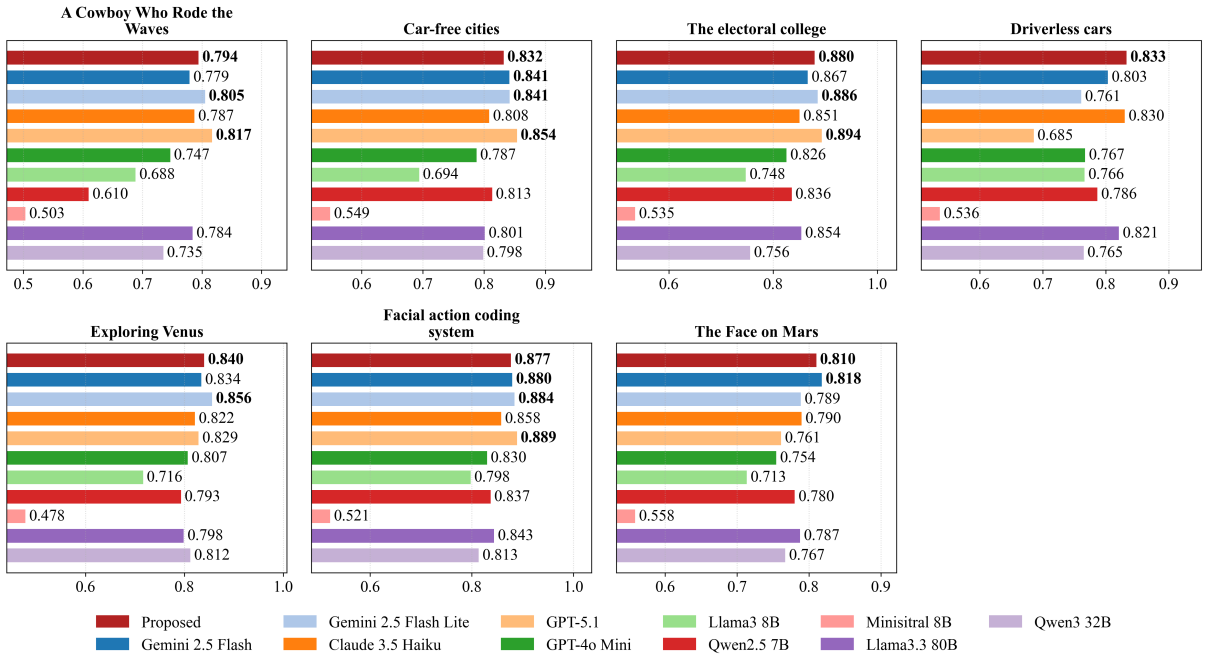

Figure 3: Performance comparison between the proposed framework (using all models) and individual models by theme. In some themes, some single models outperform the framework, and model performance varies by theme.

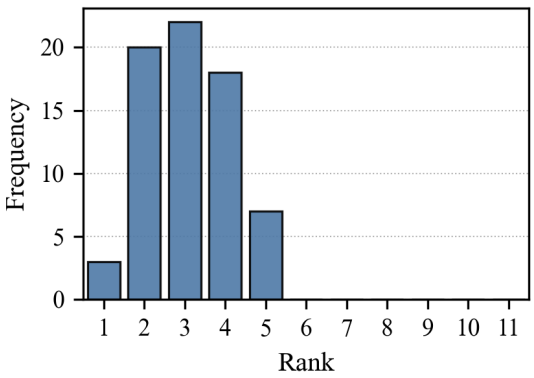

Figure 4: Histogram showing the rank of MAPLE using all models relative to single models. Aggregated across themes and iterations. MAPLE consistently ranks high, most frequently occupying the 2nd, 3rd, or 4th positions.

Figure 5: Results of Experiments for RQ3. MAPLE remains robust with negligible performance loss when up to four adversarial models are added, but declines when five Position-1– or Position-2–biased models are introduced.

tors does not affect performance, introducing five position-biased models causes a decline, with five Position-2-Biased Models leading to a pronounced drop.

As discussed in Section 5.1, MAPLE's aggregation step can degrade when biased evaluators form a majority, as it inherently relies on majority agreement among evaluators. The original pool already includes a Position-2-biased model, Ministral-8B, as shown in Figure 6. Consequently, injecting additional Position-2-biased models shifts the majority more strongly than injecting Position-1-biased models, resulting in a larger performance drop.

6 Ablation Study

We confirm that MAPLE improves alignment with human scores on an essay scoring task. We conduct an ablation study to clarify which components in the aggregation step (Step 3) are responsible for this gain. Concretely, we analyze the contributions of (i) learning model reliability η_k, (ii) learning criterion weights w_c, and (iii) modeling criterion-specific latent scores $\theta_{i,c}$.

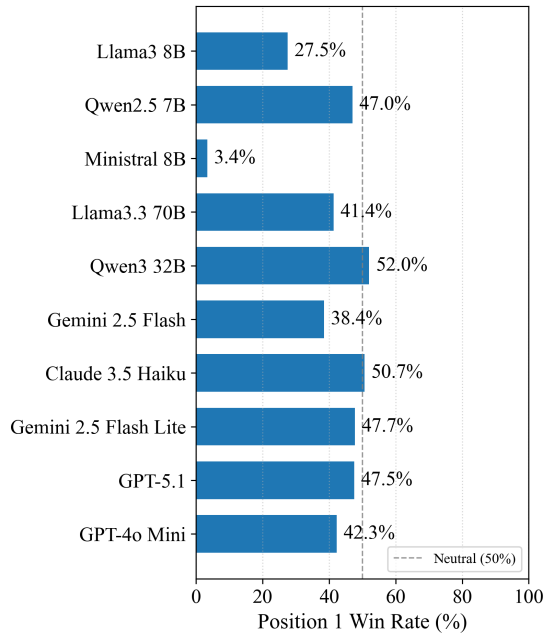

Figure 6: Proportion of instances where each model selects Position 1 in Experiments for RQ2. The data show that Ministral-8B has a strong bias toward Position 2.

6.1 Settings

We follow the same evaluation protocol as Section 5.2: for each of the seven prompts, we randomly sample 50 essays and repeat this process 10 times, resulting in 70 prompt-resampling runs in total. For all ablations, we reuse the same pairwise comparison outcomes collected in Step 2 (essay comparisons under criteria and criterion-importance comparisons). Thus, the variants differ only in how Step 3 fits and uses the parameters in eqs. (3) to (6), and no additional LLM queries are required. We evaluate each variant using CI (eq. (7)).

We compare the following variants:

- **MAPLE (Full)** learns evaluator reliability, criterion weights, and criterion-wise latent scores, and computes final scores (eq. (6)).
- **Fixed-Reliability** removes reliability learning by fixing $\eta_k = 1$ for all evaluators and learning only w_c and $\theta_{i,c}$.
- **Fixed-Criterion Weights** removes criterion-weight learning by fixing $w_c = 0$ for all criteria, which yields uniform weights after the softmax in eq. (6), while learning η_k and $\theta_{i,c}$.
- **Collapsed-Criteria** removes explicit multi-criteria modeling by collapsing the criterion index, such that all essay comparisons are treated as belonging to a single pseudo-criterion. As a result, the model learns only evaluator reliability η_k and a single latent score per essay, equivalent to Crowd-BT.

Variant	CI ↑	Top-1 ↑
MAPLE (Full)	**0.814±0.059**	**50/70**
Fixed-Reliability	0.789±0.058	2/70
Fixed-Criterion Weights	0.800 ±0.061	11/70
Collapsed-Criteria	0.793 ±0.064	3/70

Table 4: Ablation results under the same setting as Section 5.2. We report mean±std CI over 70 prompt-resampling runs (7 prompts × 10 resamplings). Top-1 counts runs where a variant achieves the highest CI (ties excluded).

6.2 Results

Table 4 reports results over 70 runs. Full MAPLE attains the highest mean CI (0.814) and ranks first in 50 out of 70 runs (71.4%), indicating both strong average performance and stability. When averaged over 10 resamplings, Full MAPLE also shows the highest mean CI across all seven prompts (Appendix D).

Compared to MAPLE (Full), Fixed-Reliability reduces the mean CI from 0.814 to 0.789, indicating that learning evaluator reliability η_k plays a central role in performance gains. Furthermore, Fixed-Criterion Weights achieves a lower mean CI (0.800) than MAPLE (Full), suggesting that learning criterion importance w_c provides additional benefits beyond reliability learning, although its impact is smaller than that of η_k. Collapsed-Criteria attains a mean CI of 0.793, highlighting the importance of explicitly modeling criterion-specific latent scores $\theta_{i,c}$. Notably, Fixed-Criterion Weights still outperforms Collapsed-Criteria (0.800 vs. 0.793), indicating that maintaining criterion-specific representations is beneficial even when criterion weights are fixed to be uniform.

7 Conclusion

We propose MAPLE, a framework that evaluates outputs using multiple LLMs across multiple criteria. MAPLE demonstrates robust, stable performance regardless of the evaluators used. Our results further show that MAPLE achieves higher performance than methods that either (i) aggregate multiple LLMs without explicit criteria or (ii) use a single LLM for multi-criteria evaluation.

Limitations

This study focuses on essay scoring as a representative open-ended evaluation task to assess the proposed framework. While MAPLE is designed to be task-agnostic, we leave validation on additional domains and datasets for future work.

In addition, MAPLE relies on multi-criteria pairwise comparisons, which increases the number of required evaluations as the number of criteria grows. This reflects a trade-off between evaluation granularity and cost, and motivates future work on more efficient comparison or sampling strategies.

Ethical Considerations

We conduct experiments using the publicly available ASAP 2.0 dataset (Crossley et al., 2025). Our use of ASAP 2.0 follows its license (CC BY-NC-SA 4.0) and is limited to research purposes consistent with the dataset's stated potential uses, such as automated essay scoring research and analyses of student writing. We therefore do not redistribute the raw essays and do not release any derived text from the essays. We report only aggregated experimental results. In the same spirit, any artifacts produced by our pipeline, including generated evaluation criteria and pairwise comparison labels, are used solely for research and analysis. These artifacts are not deployed for operational grading or other high-stakes decision-making. We also follow the usage conditions of the evaluator models. Open-source LLMs are used under their respective model licenses, and closed-source models are accessed only through official APIs in accordance with the providers' terms of service. We do not collect any new data from human participants. We rely on the dataset creators' documentation regarding data collection, consent, and permitted research use, and we use the dataset solely for research purposes in accordance with its terms. Although the dataset includes student demographics, we use only the essay text and grade information in this study. We do not use, process, or transmit any demographic or other explicit student metadata. We note that free-form essays may contain self-disclosed identifying information. To mitigate this risk, we (i) limit the information sent to any model API to only the essay text and the task prompt, (ii) do not attempt re-identification, (iii) do not reproduce any raw essays in the paper or release any new text derived from the dataset, and (iv) report only aggregate results. LLM-based evaluation methods may be misused in high-stakes educational settings without sufficient validation, potentially resulting in unfair or biased assessments. In addition, evaluator models may exhibit systematic biases or sensitivity to prompt design. Our work is intended for research purposes only, and we caution against deploying such methods as standalone decision-makers without human oversight and task-specific validation.

References

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 193–202, New York, NY, USA. Association for Computing Machinery.

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.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq R. Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joaquín Jordán, Xavier Yin, Melissa Fabros, Gireeja Ranade, and Narges Norouzi. 2026. MAGIC: Multi-agent argumentation and grammar integrated critiquer. In *Proceedings of the 16th Symposium on Educational Advances in Artificial Intelligence (EAAI-26)*, Singapore, Singapore. Accepted into EAAI 2026; program lists this paper for Jan 24–26, 2026.

Ruosen Li, Teerth Patel, and Xinya Du. 2024. PRD: Peer rank and discussion improve large language model based evaluations. *Transactions on Machine Learning Research*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Xiaotian Lu, Jiyi Li, Koh Takeuchi, and Hisashi Kashima. 2024. AHP-powered LLM reasoning for multi-criteria evaluation of open-ended responses. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1847–1856, Miami, Florida, USA. Association for Computational Linguistics.

Mistral AI Team. 2024. Un ministral, des ministraux: Introducing the world's best edge models. `https://mistral.ai/news/ministraux`. Mistral AI official blog post (Oct 16, 2024).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katherine Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Yiyue Qian, Shinan Zhang, Yun Zhou, Haibo Ding, Diego Socolinsky, and Yi Zhang. 2025. Enhancing LLM-as-a-judge via Multi-Agent collaboration. In *Proceedings of the AAAI 2025 Workshop on Advancing LLM-Based Multi-Agent Collaboration (WMAC 2025)*, Philadelphia, Pennsylvania, USA.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track*. NeurIPS.

R. W. Saaty. 1987. The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling*, 9(3–5):161–176.

Thomas L. Saaty. 2004. Decision making—the analytic hierarchy and network processes (AHP/ANP). *Journal of Systems Science and Systems Engineering*, 13(1):1–35.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024)*, pages 18990–18998.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *International Conference on Learning Representations (ICLR) 2025*. Poster. OpenReview: https://openreview.net/forum?id=3GTtZFiajM.

Xiaohui Zeng, Bingzhe Wu, Yutong Ban, and Xipeng Qiu. 2024. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Appendix

This appendix provides supplementary material for transparency and reproducibility of MAPLE. Appendix A reproduces the holistic rating form used to score essays in ASAP 2.0. In our pipeline, this rubric is provided only to the criterion-generation step (Step 1) to derive analytic evaluation criteria, and it is shared across all themes used in our experiments. Appendix B documents the LLM inference

settings and output-validation policy. Appendix C lists the exact prompt templates used in Steps 1–2. Appendix D reports additional experimental results and breakdowns omitted from the main text due to space constraints.

A Rubric (ASAP 2.0)

Explanation The text below is the holistic rating form used for expert essay scoring in ASAP 2.0 on a 1–6 scale. The same holistic rubric is used for scoring across all themes evaluated in this paper. In our pipeline, the holistic rubric is given only to the criterion-generation model in Step 1, which uses it to generate independent evaluation criteria. The evaluator models in Step 2 do not see the holistic rubric and instead receive only the generated criterion text to perform pairwise comparisons.

Holistic Rating Form After reading each essay and completing the analytical rating form, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 6 (maximum). As with the analytical rating form, the distance between each grade (e.g., 1–2, 3–4, 4–5) should be considered equal.

SCORE OF 6: An essay in this category demonstrates clear and consistent mastery, although it may have a few minor errors. A typical essay effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking; the essay uses clearly appropriate examples, reasons, and other evidence taken from the source text(s) to support its position; the essay is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas; the essay exhibits skillful use of language, using a varied, accurate, and apt vocabulary and demonstrates meaningful variety in sentence structure; the essay is free of most errors in grammar, usage, and mechanics.

SCORE OF 5: An essay in this category demonstrates reasonably consistent mastery, although it will have occasional errors or lapses in quality. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking; the essay generally using appropriate examples, reasons, and other evidence taken from the source text(s) to support its position; the essay is well organized and focused, demonstrating coherence and progression of ideas; the essay exhibits facility in the use of language, using appropriate vocabulary demonstrates variety in sentence structure; the essay is generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: An essay in this category demonstrates adequate mastery, although it will have lapses in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking; the essay using adequate examples, reasons, and other evidence taken from the source text(s) to support its position; the essay is generally organized and focused, demonstrating some coherence and progression of ideas exhibits adequate; the essay may demonstrate inconsistent facility in the use of language, using generally appropriate vocabulary demonstrates some variety in sentence structure; the essay may have some errors in grammar, usage, and mechanics.

SCORE OF 3: An essay in this category demonstrates developing mastery, and is marked by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence taken from the source texts to support its position; the essay is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas displays; the essay may demonstrate facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice and/or lacks variety or demonstrates problems in sentence structure; the essay may contain an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: An essay in this category demonstrates little mastery, and is flawed by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited, and demonstrates weak critical thinking; the essay provides inappropriate or insufficient examples, reasons, or other evidence taken from the source text to support its position; the essay is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas; the essay displays very little facility in the use of language, using very limited vocabulary or incorrect word choice and/or demonstrates frequent problems in sentence structure; the essay contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: An essay in this category demonstrates very little or no mastery, and is severely flawed by ONE OR MORE of the following weak-

nesses: develops no viable point of view on the issue, or provides little or no evidence to support its position; the essay is disorganized or unfocused, resulting in a disjointed or incoherent essay; the essay displays fundamental errors in vocabulary and/or demonstrates severe flaws in sentence structure; the essay contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.

B LLM Settings

For all LLM calls in Steps 1–2, we set the temperature to 0.2 whenever the API supports this parameter. For GPT-5.1, we additionally set reasoning_effort to none. All prompts enforce a strict JSON-only output schema (e.g., a single key "winner" or "priority_criterion"). If a model response does not conform to the required format (e.g., invalid JSON or extra text outside the JSON object), we re-query the same model with the same inputs until a valid response is obtained; only the first valid response is used for aggregation.

C Prompts

Analytic-criterion generation (Step 1). Figure 7 shows the prompt used to generate a small set of independent analytic criteria from the holistic rubric and the assignment prompt. The generator outputs 4–6 criteria, each with (i) a definition, (ii) a short checklist, and (iii) six level descriptors aligned to the 1–6 scale. We use the generator output to instantiate the criterion strings inserted into the evaluator prompt in Figure 8.

```
ROLE & GOAL
You are an experienced secondary-level
writing assessment designer and
rater-trainer. From the inputs below,
produce a clear set of **independent
analytic scoring criteria**. Each
criterion must include **six score levels
(1-6)** with concise, task-specific
descriptors and a short **"what to look
for"** checklist. Do not invent weights or
an aggregation method; the output is
intended for criterion-only scoring.

INPUTS
- HOLISTIC_RUBRIC:
  {{PASTE THE FULL HOLISTIC RUBRIC TEXT
  HERE}}
- TASK_PROMPT:
  {{PASTE THE ASSIGNMENT / ESSAY PROMPT
  HERE}}
- STUDENT_CONTEXT:
  {{GRADE LEVEL, COURSE, LANGUAGE PROFILE,
  ACCOMMODATIONS, ANY LOCAL CONSTRAINTS}}
- OPTIONAL_TASK_CONSTRAINTS (optional):
  {{E.G., REQUIRED USE OF SOURCES, CITATION
  RULES, LENGTH/TIME EXPECTATIONS,
  PROHIBITED ELEMENTS, ETC.}}

DESIGN REQUIREMENTS
1) Propose **4-6 independent criteria**
that are appropriate to the TASK_PROMPT
and faithful to the HOLISTIC_RUBRIC. Do
**not** assume default criterion names;
derive them from the inputs.
2) For **each criterion**, create:
   • A 1-3 sentence **definition**
   tailored to the task.
   • A **"What to look for"** checklist
   (3-7 observable indicators).
   • **Six level descriptors (1-6)**, each
   1-2 lines, expressed in behaviorally
   observable terms and reflecting equal
   steps of quality from the holistic
   rubric's intent.
   • **Common pitfalls** (2-5 brief items)
   the rater should watch for on this
   criterion.
3) Write criterion language that is
specific to the TASK_PROMPT and
STUDENT_CONTEXT, but avoid copying the
holistic rubric verbatim. Do not include
concrete examples of content that would
bias raters toward a specific answer.
4) Provide **scoring notes for borderline
cases** on each criterion (how to decide
between adjacent levels) without proposing
any cross-criterion integration.
5) Include **edge-case and fairness
guidance** applicable to the task and
context (e.g., off-prompt responses,
excessive quotation, fabricated or
unverifiable evidence, minimal length,
language mismatch, plagiarism indicators,
accessibility and ELL considerations).
Keep this generic and policy-neutral; do
not guess intent.

OUTPUT FORMAT (produce all sections in
this order)
A) **Rubric Overview (for humans)**
   • 1 concise paragraph explaining how
   the analytic criteria operationalize
   the holistic rubric for this specific
   task and context.

B) **Criteria (rater-usable detail)**
   For each criterion, present:
   - **Name**
   - **Definition**
   - **What to look for** (bullet list)
   - **Score Descriptors (1-6)** (six
   bullets, 1 is lowest, 6 is highest)
   - **Common pitfalls** (bullet list)
   - **Borderline decision notes** (1-3
   bullets)

C) **Edge Cases & Fairness**
   - Edge-case handling (bullet list)
   - Fairness/Accessibility notes (bullet
   list)
```

````
D) **Copy-paste rubric block
("RUBRIC_SPEC")**
   Provide a **self-contained,
   machine-readable JSON** between triple
   backticks that a grader prompt can
   ingest. Use exactly this schema (no
   extra fields, no weights):
   ```json
   {
     "rubric_title": "string",
     "task_prompt": "string",
     "student_context": "string",
     "criteria": [
       {
         "name": "string",
         "definition": "string",
         "what_to_look_for": ["string",
         "..."],
         "levels": {
           "1": "string",
           "2": "string",
           "3": "string",
           "4": "string",
           "5": "string",
           "6": "string"
         },
         "pitfalls": ["string", "..."],
         "borderline_notes": ["string",
         "..."]
       }
     ],
     "guardrails": {
       "edge_cases": ["string", "..."],
       "fairness_notes": ["string", "..."]
     }
   }
````

Figure 7: The prompt used for generating analytic scoring criteria.

Criterion-wise pairwise essay comparison (Step 2). Figure 8 shows the template used to collect criterion-wise pairwise evaluations. For each essay pair and each criterion, we ask an evaluator LLM to output a forced-choice decision (`"Essay 1"` vs. `"Essay 2"`), with ties disallowed. We require a single JSON object to make the outputs machine-parsable and to minimize ambiguity across heterogeneous model APIs. The instruction explicitly restricts the evaluation to the provided criterion to reduce unintended holistic spillover.

```
You are a decisive and analytical writing
judge. Your task is to compare two essays,
written in response to the same prompt,
based on a single, detailed evaluation
criterion. You must determine which essay
better fulfills this criterion.

**CRITICAL JUDGMENT INSTRUCTION:**
Your judgment MUST be based
**exclusively** on the single criterion
provided. You MUST ignore all other
aspects of the essays (such as grammar,
vocabulary, or overall persuasiveness)
**unless** they are directly relevant to
the specified criterion. Do not let
strengths or weaknesses in other areas
influence your decision.

**OUTPUT FORMAT:**
Your entire output MUST be a single, valid
JSON object. The JSON object must contain
a single key named "winner". The value for
this key must be one of these two exact
strings: "Essay 1" or "Essay 2".

Do not include any explanations or
introductory text outside of the JSON
object.

**Required Output Format Example:**
{
  "winner": "Essay 1"
}

---

**ESSAY PROMPT:**
<Assignment prompt is written here>

---

**CRITERION:**
<The criterion is written here>

---

**ESSAY 1:**
<The first essay is written here>

---

**ESSAY 2:**
<The second essay is written here>
```

Figure 8: The full prompt template used for pairwise comparison of essays based on a single criterion.

Pairwise criterion-importance comparison (Step 2). Figure 10 shows the template used to collect pairwise comparisons between criteria, which are used to estimate criterion weights w_c in Step 3. The prompt defines "more important" in terms of assignment alignment, discriminative power, observability, non-redundancy, and fairness for the target writer group. As with essay comparisons, we enforce a strict forced-choice JSON output to obtain consistent pairwise labels $y_{c,c',k}$ across evaluators.
```

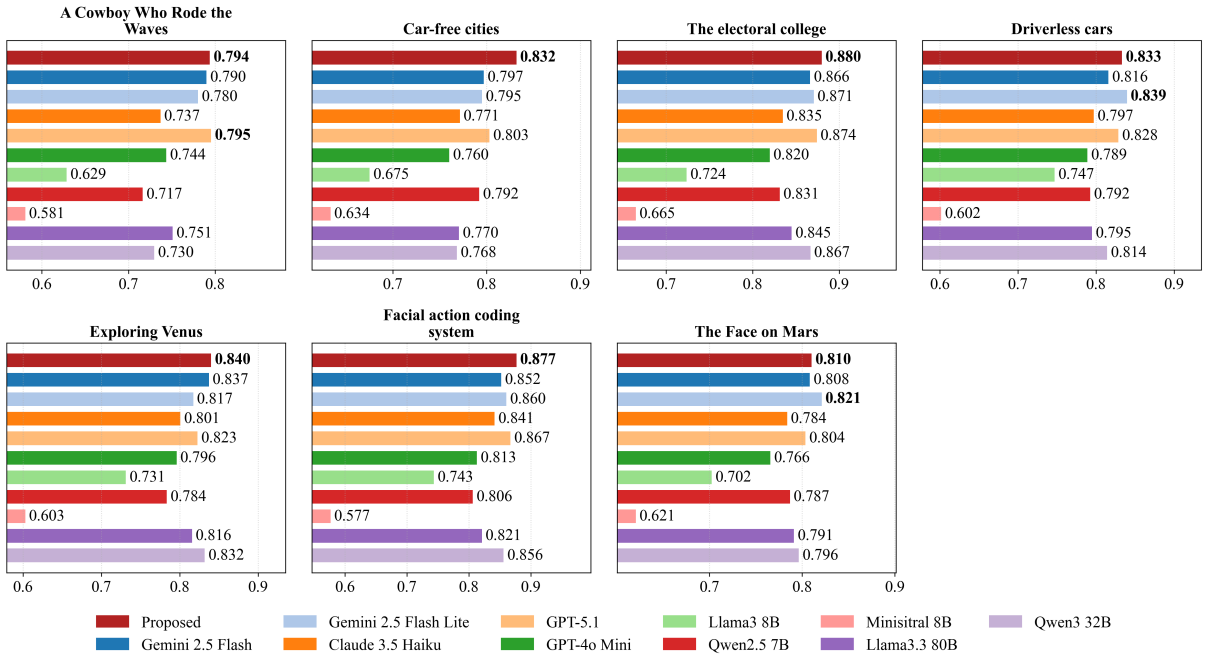

Figure 9: Performance comparison between the proposed framework and pairwise (single model), by theme. While individual models may outperform the proposed framework on specific themes, no single model consistently surpasses it across all themes, highlighting the benefit of reliability-aware aggregation under evaluator uncertainty.

Variant	MAPLE (Full)	Fixed-Reliability	Collapsed-Criteria	Fixed-Criterion Weights
A Cowboy Who Rode the Waves	**0.752** ± **0.073**	0.736 ± 0.062	0.731 ± 0.065	0.740 ± 0.058
Car-free cities	**0.803** ± **0.057**	0.781 ± 0.061	0.776 ± 0.061	0.782 ± 0.061
The electoral college	**0.859** ± **0.038**	0.824 ± 0.037	0.829 ± 0.049	0.842 ± 0.041
Driverless cars	**0.814** ± **0.047**	0.799 ± 0.058	0.801 ± 0.073	0.804 ± 0.073
Exploring Venus	**0.817** ± **0.051**	0.785 ± 0.048	0.806 ± 0.038	0.812 ± 0.038
Facial action coding system	**0.854** ± **0.027**	0.823 ± 0.039	0.836 ± 0.049	0.843 ± 0.041
The Face on Mars	**0.798** ± **0.047**	0.774 ± 0.056	0.769 ± 0.050	0.777 ± 0.047

Table 5: Ablation study results broken down by theme. We compare the full MAPLE framework against Fixed-Reliability, Fixed-Criterion Weights, and Collapsed-Criteria variants. MAPLE (Full) consistently outperforms all ablated variants across every theme.

```
ROLE:
You are an assessment designer. Decide
which of two rubric criteria should be
weighted as more important when grading a
set of essays responding to the same
assignment.

INPUTS:

Assignment Prompt (verbatim):
{{ASSIGNMENT_PROMPT}}

Criterion A:

Name: {{CRITERION_A_NAME}}

Description: {{CRITERION_A_DESCRIPTION}}

Criterion B:

Name: {{CRITERION_B_NAME}}

Description: {{CRITERION_B_DESCRIPTION}}

Writer Group: {{WRITER_GROUP}}

POLICY:

Use only the provided inputs; do not
import outside standards or knowledge.

Define “more important” as the criterion
that, relative to the other, best
satisfies these lenses:

Core-Demand Alignment — directness to the
assignment’s explicit requirements and
command verbs.

Discriminative Power — ability to separate
stronger vs. weaker essays for this writer
group.
```

```
Observability and Reliability — ease of
seeing and scoring it consistently in
typical responses.

Non-Redundancy — unique value not already
covered by the other criterion.

Fairness and Developmental Fit —
appropriate challenge for the specified
group.

Internally score each criterion 0–2 on
each lens and sum the scores (0–10). Do
not output scores or reasoning.

DECISION RULES (ties are not allowed):

Choose the criterion with the higher
total.

If totals are equal, compare the criteria
in this priority order and choose the one
that is stronger on the first lens that
differs:

Core-Demand Alignment

Observability and Reliability

Discriminative Power

Non-Redundancy

Fairness and Developmental Fit

If still equal, choose the criterion more
tightly tied to the assignment's command
verbs (e.g., evaluate, explain, argue,
analyze, use evidence).

If still indistinguishable, choose
Criterion A.

Never return a tie.

OUTPUT (result only):
Return exactly one of the following JSON
objects and nothing else:
{ "priority_criterion": "Criterion A" }
{ "priority_criterion": "Criterion B" }

CONSTRAINTS:

No preamble, no explanation, no reasoning,
no additional fields, no extra whitespace
or lines.

Do not quote or wrap the JSON; print it
raw.
```

Figure 10: The prompt used for determining the priority weighting between two criteria.

D Results

This section provides supplementary breakdowns for Experiments for RQ2 (Section 5.2) and the ablation study (Section 6). We report per-theme results and additional visualizations to illustrate variability across prompts and evaluators. All scores are measured by Concordance Index (CI) and are averaged over 10 resampling runs per theme unless noted otherwise.

E Computational Resources

Our experiments are inference-only. The total budget is dominated by LLM inference in Steps 1–2, while the Step-3 aggregation (optimization of the extended Crowd-BT model) runs on CPU and typically completes within a few minutes per run. For open-source models, parameter sizes are as indicated by their model names. For closed-source models accessed via official APIs, parameter counts and serving infrastructure are not disclosed by the providers; we therefore report API usage as a proxy for compute. In Experiments for RQ2, for each run with N essays, C criteria, and K evaluator models, Step 2 issues $K\left(C\binom{N}{2}+\binom{C}{2}\right)$ LLM requests for essay comparisons and criterion-importance comparisons (excluding occasional retries for invalid JSON responses).