LLMs Judge Themselves: A Game-Theoretic Framework for Human-Aligned Evaluation

Anonymous ACL submission

Abstract

Ideal or real, that is the question. In this work, we explore whether principles from game theory can be effectively applied to the evaluation of large language models (LLMs). This inquiry is motivated by the growing inadequacy of conventional evaluation practices, which often rely on fixed-format tasks with reference answers and struggle to capture the nuanced, subjective, and open-ended nature of modern LLM behavior. To address these challenges, we propose a novel alternative: automatic mutual evaluation, where LLMs assess each other's output through self-play and peer review. These peer assessments are then systematically compared with human voting behavior to evaluate their alignment with human judgment. Our framework incorporates game-theoretic voting algorithms to aggregate peer reviews, enabling a principled investigation into whether model-generated rankings reflect human preferences. Empirical results reveal both convergences and divergences between theoretical predictions and human evaluations, offering valuable insights into the promises and limitations of mutual evaluation. To the best of our knowledge, this is the first work to jointly integrate mutual evaluation, game-theoretic aggregation, and human-grounded validation for evaluating the capabilities of LLMs.

1 Introduction

As Large Language Models (LLMs) (Qiu et al., 2025; DeepSeek-AI et al., 2025; Jaech et al., 2024) are increasingly deployed in real-world applications, evaluation has shifted from isolated performance measurement to comparative model ranking (Liu et al., 2025; Chiang et al., 2024). However, existing evaluation methodologies (Lin and Chen, 2023) often simplify assessment to single-model judgments or scalar metrics, resulting in misalignment with human comparative judgments. Humans typically navigate such contexts by aggregating partial rankings from diverse perspectives, then synthesizing them into a globally optimized consensus (Yuan et al., 2025; Lyon and Pacuit, 2013). This perspective suggests that LLM evaluation is not only a prediction problem but also a collective decision problem, making it naturally amenable to a **game-theoretic** view. In particular, our work focuses on the part of this perspective that is most closely connected to *social choice theory*: how multiple judgments over competing model outputs can be aggregated into a ranking that best reflects human preferences. Figure 1 provides an overview of this setting.

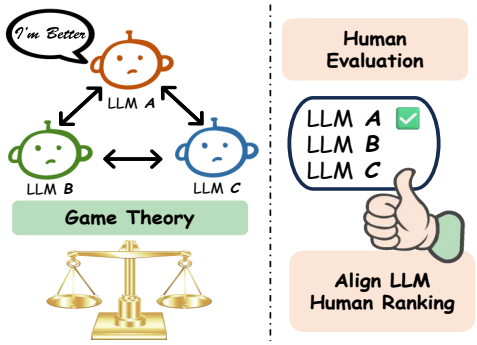

Figure 1: Illustration of game-theoretic peer evaluation for LLM performance ranking.

More broadly, game-theoretic approaches offer a useful language for reasoning about interaction, competition, and collective outcomes among multiple evaluators, while social choice theory provides the concrete aggregation tools needed to combine rankings into a consensus decision. Recent work has increasingly connected these collective decision mechanisms to LLM evaluation (Pasch, 2025; Jaech et al., 2024; Sun et al., 2025). Motivated by this connection, we ask: *Can game-theoretic evaluation frameworks, instantiated through social-choice-based aggregation methods, yield model rankings that align with human judgments when evaluating LLM capabilities?*

To validate the aforementioned hypothesis, we must address self-preference bias (Liu et al., 2024; Dietz et al., 2025), where models tend to favor their own outputs when acting as evaluators. To this end, we adopt an automated peer-evaluation setting (Chen et al., 2024; Chu et al., 2024b) and formalize it as a game-theoretic voting process. Specifically, we introduce a *decentralized peer ranking* framework in which each LLM serves as both evaluatee and evaluator, producing rankings over candidate outputs without relying on a central judge. We then aggregate these rankings using a suite of aggregation algorithms (Zhang et al., 2025; Charikar et al., 2024a,b; Xia and Zheng, 2022; Wood et al., 2024) to derive a consensus ranking and compare the resulting methods by their alignment with human judgments.

This decentralized peer ranking framework replaces centralized judgment with mutual model-based assessment, enabling collective evaluation without a single authoritative judge, while its reliability and fairness depend on several key factors. Before presenting the experimental design, it is important to clarify the core research objectives our study seeks to address.

During peer evaluation, how well do different aggregation methods align with human judgments when ranking LLMs? We find that aggregation-based methods consistently achieve stronger and more stable alignment with human judgments than single-model evaluations. Among them, consensus-oriented methods, especially Kemeny–Young (Kemeny, 1959), perform best across benchmarks. These methods also outperform rankings induced by standard accuracy-based benchmarks, suggesting that preference aggregation captures aspects of human evaluation that correctness alone does not.

Can game-theoretic ranking algorithms mitigate such self-bias and produce more reliable evaluation rankings? Consistent with prior studies (Liu et al., 2024; Dietz et al., 2025), our results confirm that LLMs exhibit systematic self-preference bias, with self-assigned rankings showing weak and unstable alignment with human judgment. We further find that game-theoretic aggregation substantially reduces the impact of this bias. In particular, consensus-based aggregation produces stable rankings that remain well aligned with human preferences even when self-assigned rankings are included, demonstrating the robustness of game-theoretic methods to self-preference bias.

Are there specific LLM capabilities where game-theoretic evaluation exhibits varying alignment with human judgment? LLMs possess diverse capabilities across a wide range of tasks, from logical reasoning to creative generation. Our results show that alignment with human judgment varies substantially across different capability domains. Game-theoretic aggregation achieves the strongest and most stable alignment on tasks with clear and objective correctness signals, such as mathematical reasoning tasks, while exhibiting weaker and more variable alignment on more subjective tasks, including creative writing and linguistically nuanced benchmarks. Nevertheless, even in such settings, aggregate rankings can still capture human consensus at the dataset level overall.

In summary, we propose a decentralized, game-theoretic peer ranking framework in which each LLM simultaneously serves as an evaluatee and an evaluator, with each evaluator producing a ranking over candidate models and a consensus ranking obtained by aggregating these individual rankings using principled aggregation algorithms. Using this framework, we conduct a systematic analysis of aggregation-based ranking methods across multiple benchmarks and evaluation settings, comparing their alignment with human judgments, robustness to self-preference bias, and consistency across capability domains. Our results show that consensus-based aggregation methods, particularly the Kemeny–Young method, consistently produce rankings that are more stable and more closely aligned with human preferences than both single-model evaluations and accuracy-based benchmarks. We further demonstrate that game-theoretic aggregation substantially reduces the impact of self-preference bias and remains reliable even when self-assigned rankings are included, while revealing clear differences in reliability across task types, with the strongest performance on tasks with clear and objective correctness signals. Taken together, these findings highlight the importance of treating LLM evaluation as a structured aggregation and decision-making problem, and suggest that future evaluation protocols should prioritize consensus-oriented and bias-aware aggregation mechanisms to achieve fair and human-aligned assessment.

2 Related Work

LLM evaluation. Evaluating LLMs remains an active area of study because their capabilities are

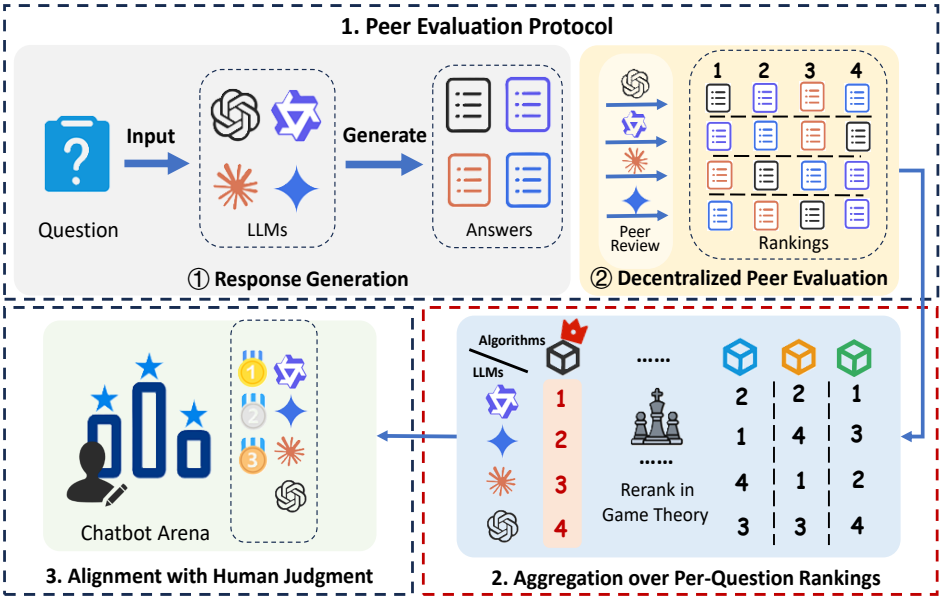

Figure 2: The proposed framework for game-theoretic evaluation of LLMs.

broad, rapidly evolving, and not always easy to capture with a single metric. Traditional benchmarks such as MMLU (Hendrycks et al., 2021) and GSM8K (Chiang et al., 2024) rely on curated datasets with reference answers and human or rule-based grading, providing consistency and reproducibility for model comparison. At the same time, benchmark saturation has prompted discussion about whether static test sets fully reflect generalizable performance (Fu et al., 2025; An et al., 2025). Another line of work uses LLMs as evaluators for open-ended or subjective tasks, often by appointing a single model as judge (Chu et al., 2024a). This design offers scalability, although it can also make evaluation outcomes sensitive to the choice of evaluator.

Aggregation-based evaluation. Related work has also explored comparative and aggregation-based evaluation frameworks. Some studies use game-inspired environments to analyze strategic behavior under explicit payoff structures (Duan et al., 2024; Cipolina-Kun et al., 2025), while other work draws on preference aggregation from social choice theory to derive consensus judgments from multiple comparisons. For example, Zhang et al. (2025) use an enhanced Borda count to rerank mathematical reasoning responses. Along similar lines, our work is related to this broader line of research, and focuses on decentralized peer evaluation in which all participating LLMs provide rankings, allowing aggregation rules to be examined with respect to their alignment with human judgments and their robustness to self-preference bias.

3 Methodology

3.1 Overview of Our Evaluation Framework

Conventional evaluation methods for LLMs often rely on comparisons against predefined reference answers, which emphasize correctness with respect to fixed outputs. While effective for many benchmark settings, such approaches do not explicitly model the comparative and preference-based nature of human judgment that arises in many real-world evaluation scenarios.

As shown in Figure 2, we propose an automated peer evaluation framework in which LLMs assess the outputs of all candidate models, including their own. Each model acts as both an evaluatee and an evaluator, producing rankings over peer-generated responses. These decentralized judgments are aggregated using game theoretic voting algorithms to yield a consensus ranking over all candidate models at each question, which supports analysis of alignment with human preferences. The framework consists of three core components: the peer evaluation protocol, aggregation of per question rankings, and alignment with human judgment.

3.2 Peer Evaluation Protocol

Let $\mathcal{M} = \{LLM_1, \ldots, LLM_M\}$ denote a collection of M models, and let $\mathcal{Q} = \{q_1, \ldots, q_N\}$ be a shared set of evaluation questions. For each

model $LLM_i \in \mathcal{M}$, we collect a set of responses $A_i^{\mathcal{Q}} = \{a_{ij}\}_{j=1}^{N}$, where a_{ij} denotes the response produced by LLM_i for question q_j.

As illustrated in the upper portion of Figure 2, the evaluation procedure consists of two stages.

Response Generation. Each model independently responds every question $q_j \in \mathcal{Q}$, producing a response a_{ij} for each $LLM_i \in \mathcal{M}$.

Decentralized Peer Evaluation. For each question q_j, every model LLM_i simultaneously serves as an evaluator and is presented with a randomly ordered and anonymized set of responses $A^{(j)} = \{a_{1j}, a_{2j}, \ldots, a_{Mj}\}$, including its own. The evaluator provides a complete ranking $\pi_i^{(j)} \in \mathcal{S}_M$ over the candidate responses based on their quality with respect to the given question, where $\mathcal{S}_M$ denotes the set of all permutations over the M candidate models. The ranking $\pi_i^{(j)}$ represents the model's subjective preference ordering for question q_j, analogous in form to a human-provided ranking. Collectively, the set $\Pi^{(j)} = \{\pi_i^{(j)}\}_{i=1}^{M}$ constitutes a decentralized preference profile for question q_j.

3.3 Aggregation over Per-Question Rankings

Given the decentralized preference profiles $\Pi^{(j)}$ obtained from peer evaluation for a given evaluation question q_j, the aggregation problem maps the preference profile $\Pi^{(j)}$ to a consensus ranking $\pi^{(j)} \in \mathcal{S}_M$, which reconciles potentially conflicting judgments from different evaluators.

Since rankings may vary across evaluators due to differences in model capabilities, evaluation noise, or systematic biases, the aggregation method must be robust and consistent. We consider multiple ranking aggregation algorithms, including Borda count, Copeland voting, Kemeny–Young and so on. In our experiments, aggregation-based methods consistently exhibit stronger alignment with human rankings than individual model judgments. Among them, the Kemeny–Young method achieves the highest and most stable alignment with human preferences and is therefore adopted as the default aggregation strategy. For each question q_j, Kemeny–Young identifies the ranking $\pi^{(j)}$ that minimizes the total Kendall–Tau distance to the input rankings in $\Pi^{(j)}$, equivalently minimizing total pairwise disagreement. Alternative aggregation strategies are discussed in Appendix D.

3.4 Alignment with Human Judgment

To evaluate our framework, we compare rankings produced by the game-theoretic method against human preference rankings. Specifically, we use rankings derived from Chatbot Arena (Chiang et al., 2024), a large-scale crowdsourced evaluation platform in which human annotators provide pairwise preferences over model outputs. Aggregating these votes across annotators and prompts yields a consensus ranking over LLMs. This ranking serves as our benchmark for alignment analysis. We compute several metrics to quantify the similarity between our rankings and those from Chatbot Arena, namely the **Pearson correlation coefficient**, which measures linear correlation between model scores, and **Kendall's** τ (Kendall, 1948), which measures rank correlation.

High alignment scores indicate that decentralized model-driven evaluation can effectively approximate aggregated human judgments, offering a scalable and label-free alternative to traditional human evaluation pipelines.

4 Experiments and Analysis

4.1 Experimental Setups

Datasets. We evaluate model reasoning across a diverse set of widely used benchmarks, including GSM8K (Chiang et al., 2024), MMLU (Hendrycks et al., 2021), GPQA (Rein et al., 2023), CEval (Huang et al., 2023), IFEval (Zhou et al., 2023a), MBPP (Austin et al., 2021), and the Creative Writing benchmark (Wu et al., 2025). All datasets are subsampled for tractability while preserving diversity across domains and reasoning types. To further mitigate the effects of benchmark saturation, we additionally prompt `GPT-4` to construct 3 synthetic evaluation sets, namely GenOverall, GenMath, and GenChinese, which are designed to cover broad reasoning abilities, mathematical reasoning, and Chinese language understanding, respectively. Details of the dataset construction process are provided in Appendix E.

Models. Our experiments include 6 recent LLMs from OpenAI (GPT-4o variants) and Anthropic (Claude-3 and 3.5 series), chosen for their consistently top-tier performance in recent benchmarks.

Human Preference Reference. Chatbot Arena (Chiang et al., 2024) serves as our human preference reference. Because it provides system-level rankings across multiple domains, we

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.296	0.156	0.311	0.313	0.169	0.024	0.300	0.313	0.342	0.302	-0.151	0.289	0.367	0.345
	Std.	0.494	0.511	0.511	0.499	0.504	0.493	0.530	0.521	0.536	0.526	0.475	0.517	0.540	0.539
	Min	-0.829	-0.829	-0.771	-0.771	-0.771	-0.714	-0.784	-0.771	-0.771	-0.771	-1.000	-0.771	-0.771	-0.771
	25%	0.014	-0.271	-0.043	0.029	-0.314	-0.257	-0.203	-0.214	-0.214	-0.214	-0.657	-0.214	-0.214	-0.214
	50%	0.343	0.257	0.371	0.343	0.343	-0.086	0.504	0.543	0.514	0.514	-0.143	0.457	0.571	0.486
	75%	0.671	0.486	0.671	0.671	0.500	0.371	0.683	0.714	0.771	0.671	0.200	0.671	0.786	0.771
	Max	1.000	0.943	1.000	1.000	0.943	0.943	0.968	1.000	1.000	1.000	0.714	0.943	1.000	1.000

Table 1: Distribution of **Pearson** correlation coefficients between automated rankings and human rankings on the `GenOverall` dataset. Correlations are computed at the micro level for each question. The left columns report individual LLMs, and the right columns report aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment, and color intensity reflects correlation strength.

match each benchmark to the closest leaderboard subset: GSM8K and GenMath to *math-specific*, MMLU, GPQA, and GenOverall to *overall*, CEval and GenChinese to *Chinese*, IFEval to *instruction-following*, and MBPP to *code-related* rankings. Alignment is quantified using Pearson and Kendall (Kendall, 1948) correlations between our aggregated rankings and those from Chatbot Arena.

Evaluation Levels. We evaluate alignment at two complementary levels: *micro-level* and *macro-level*. **Micro-level correlation** measures alignment for individual questions. For each question, we compute the correlation between rankings produced by an aggregation method and the corresponding human reference, then analyze the resulting distribution using summary statistics such as the mean, standard deviation, and percentiles. For example, a higher 25th percentile indicates that most questions exhibit strong alignment with human judgments. **Macro-level correlation** averages per-question rankings to obtain a single overall ranking for the dataset, then computes a single correlation score against the human-provided global ranking. This captures dataset-level alignment but not per-question variability.

Ranking Algorithms. Our framework uses nine classical aggregation methods from social choice theory and voting literature to produce a final ranking over models from decentralized peer evaluations: Average, Borda Count (Saari, 1985), Copeland (Saari and Merlin, 1996), Dodgson (Brandt, 2009), Instant-Runoff Voting (IRV) (Brandt, 2009), Kemeny-Young (Kemeny, 1959), Kendall (Kendall, 1948), and Spearman. These algorithms take as input ranking lists or pairwise preferences collected across evaluation questions and produce an overall consensus ranking of models. They cover a broad range of principles, including positional, pairwise, and consistency-based approaches. Further implementation details are provided in Appendix D.

		CEval	Writing	GSM8K	GPQA	MMLU
Prs	**Acc**	0.227	-	0.319	0.311	0.932
	Kem	0.714	0.914	0.941	0.907	0.919
Ken	**Acc**	0.138	-	0.086	0.298	0.894
	Kem	0.467	0.867	0.867	0.733	0.733

Table 2: Alignment between **macro-level** model rankings and human preferences. **Acc** denotes rankings induced by accuracy, and **Kem** denotes rankings derived from Kemeny–Young aggregation over per-question peer evaluations. Pearson (Prs) and Kendall (Ken) coefficients are reported, with higher values indicating stronger alignment with Chatbot Arena rankings.

4.2 Q1: Can Game-Theoretic Evaluation Align with Human Judgment?

To evaluate alignment with human judgment, we compare rankings produced by different aggregation algorithms with the Chatbot Arena leaderboard (Chiang et al., 2024) using Pearson and Kendall correlations. To reduce the impact of benchmark saturation, we primarily report Pearson results on `GenOverall` in the main text (Table 1), while Kendall results and additional datasets are provided in Appendix F.

Re-ranking Methods Achieve Higher Alignment with Human Judgment than Single-Model Evaluations. As shown in Table 1, most rank aggregation algorithms achieve stronger alignment with human judgment than individual models, except for IRV. This advantage is reflected in both higher average Pearson correlations and greater stability. For example, the `Kemeny` method achieve a median Pearson correlation of 0.571, substan-

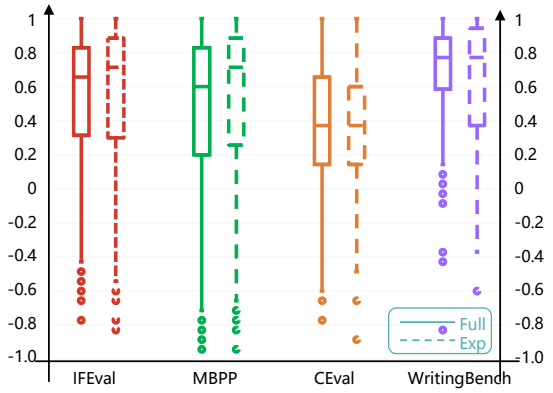

Figure 3: Pearson correlation distributions between human rankings and Kemeny–Young aggregated rankings under the sparse expander-graph setting. Solid boxes correspond to the full comparison graph, while dashed boxes correspond to the expander-based sparse graph.

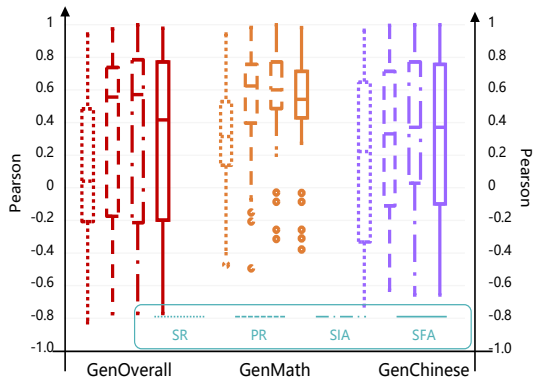

Figure 4: Pearson correlation distributions between model rankings and human preferences under SR, PR, SIA, and SFA. Higher values and narrower distributions indicate stronger and more stable alignment. Additional results are in the Appendix G.

tially outperforming the strongest individual model, `gpt-4o-20241120`, which attains a median of 0.343 with greater variance. Kendall results in Appendix F show the same pattern. These results indicate that aggregation methods improve alignment with human preferences while yielding more consistent outcomes across questions.

Kemeny Aggregation Delivers the Strongest and Most Stable Human Alignment. On the aggregation side, `Kemeny-Young` achieves the highest median correlations with human preferences and relatively narrow interquartile ranges, indicating both strong alignment and low variance. Theoretically, this advantage stems from the Kemeny–Young objective, which minimizes aggregate Kendall–Tau distance and therefore directly optimizes agreement over pairwise model preferences. From a social-choice perspective, it is therefore a maximum-consensus rule that preserves shared pairwise preferences across evaluators. This is particularly suitable for LLM evaluation, where individual model judgments are noisy and biased, but repeated pairwise preferences can still reveal a stable human-aligned ordering. By contrast, Borda-style scoring can lose pairwise structure, while `IRV` is sensitive to early eliminations. Across Tables 6-13, `Kendall` and `Kemeny-Young` perform similarly in most settings because both are based on rank-distance minimization. However, `Kemeny-Young` generalizes more consistently across benchmarks, especially when rankings are noisy, heterogeneous, or partially sparse, making it the most robust aggregation method overall.

To assess whether accuracy reflects human preference, we compare accuracy-induced rankings with game-theoretic rankings using **macro-level correlation**. As shown in Table 2, rankings derived from `Kemeny-Young` consistently achieve higher alignment with human judgments than accuracy-based rankings. Notably, on the `Creative Writing` dataset, where no definitive reference answers are available, game-theoretic aggregation still shows strong correlation with human assessments, highlighting its robustness in subjective settings. A similar pattern appears when comparing aggregation methods with individual models: `gpt-4o-20240513` performs best on GENOVERALL (Table 1), while `gpt-4o-20241120` leads on the remaining datasets (Tables 6–13). In contrast, models such as `claude-3.5-sonnet` and `claude-3.5-haiku` exhibit lower correlations and higher variability. Overall, these results highlight the benefit of game-theoretic rank aggregation, which improves alignment with human preferences and yields more consistent rankings across diverse evaluation settings.

Scalable Evaluation with Expander Graphs. To improve scalability, we adopt an expander-graph extension (Deac et al., 2022), which provides sparse yet well-connected comparison structures. We consider two settings that differ in comparison density. In the **Full** setting, each evaluator produces a complete ranking over all M candidate LLMs, yielding $O(M^2)$ total comparisons. In the **Exp** setting, each evaluator ranks only a small subset of K models (we use $K = 3$), sampled according to

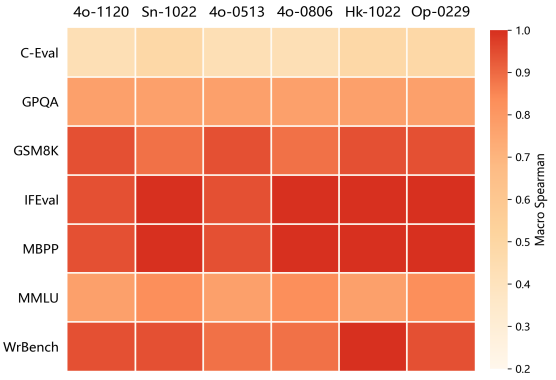

Figure 5: Macro-level Spearman consistency between incremental rankings and human rankings across datasets and newly added models.

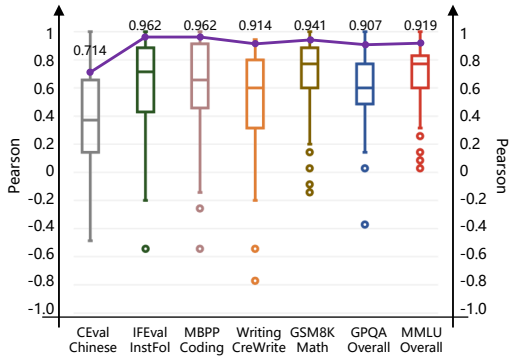

Figure 6: Micro-level Pearson correlation distributions between human rankings and Kemeny-Young aggregated rankings across benchmarks. Purple dots indicate macro-level correlations. 'CreWrite' and 'InstFol' denote creative writing and instruction following.

an expander graph construction, reducing complexity to $O(KM)$ with $K \ll M$. Figure 3 compares Pearson correlations between human rankings and Kemeny–Young aggregated rankings under the two settings. Despite the reduced comparison load, the aggregated rankings remain highly consistent with those from the full graph. Across all benchmarks, the expander-based results closely match the full-setting distributions, indicating that substantial sparsification does not degrade alignment with human judgment. These results show that the proposed peer-ranking framework can scale to larger model pools while preserving ranking stability and human alignment. Additional details are provided in Appendix A.

Incremental Ranking for Newly Added Models. We further examine whether a newly added model can be ranked efficiently after a consensus ranking over an existing model pool has already been established. To simulate this setting, we treat each model in turn as the newly added model, combine it with all $\binom{5}{3} = 10$ subsets of three existing models, obtain local rankings with `Kemeny-Young`, and then aggregate these partial rankings into a final six-model ranking. Figure 5 summarizes the macro-level Spearman consistency between these incremental rankings and human rankings across datasets and held-out models. The results show that this subset-based update strategy remains strongly aligned with human judgment on most benchmarks, while avoiding full reranking over all models.

4.3 Q2: Can Game-Theoretic Re-ranking Mitigate Self-Preference Bias in LLMs?

Recent research has demonstrated that LLMs commonly exhibit **self-preference bias** when evaluating their own outputs (Liu et al., 2024; Dietz et al., 2025). In such cases, models tend to assign favorable scores to their own generations, compromising the objectivity of evaluation outcomes. This phenomenon raises concerns about the reliability of LLMs as autonomous evaluators, particularly in decentralized or multi-agent settings where self-assessment can distort collective judgment.

To investigate self preference bias, we compare 4 evaluation protocols that differ in whether self rankings are included and whether rankings are aggregated. In both Self Ranking **(SR)** and Peer Ranking **(PR)**, each model produces a complete ranking over all models, and scores are derived directly without consensus aggregation. Under **SR**, a model is scored by the rank it assigns to itself, reflecting pure self assessment. Under **PR**, a model is scored by averaging ranks assigned by peer models, excluding self judgments. Under Self Inclusive Aggregation **(SIA)**, all rankings, including self evaluations, are aggregated using `Kemeny-Young` to produce a consensus ranking. Under Self Free Aggregation **(SFA)**, self assigned ranks are removed and treated as missing before applying the same `Kemeny-Young` procedure, yielding a consensus ranking based only on peer assessments.

Game-Theoretic Aggregation Mitigates Self-Preference Bias. As shown in Table 3, SR consistently yields more favorable rankings than PR across benchmarks, corroborating prior findings on LLM self-evaluation bias (Liu et al., 2024; Dietz et al., 2025). On GenOverall, for instance, `gpt-4o-20241120` shifts from 2.521 under SR to

Dataset	Method	4o-1120	Sn-1022	4o-0513	4o-0806	Hk-1022	Op-0229
GenOL	SR	2.521	2.313	3.625	4.146	3.021	3.208
	PR	2.929(+0.408)↑	2.433(+0.121)↑	3.733(+0.108)↑	4.342(+0.196)↑	3.925(+0.904)↑	4.071(+0.863)↑
	SIA	2.396	2.333	3.479	4.604	3.979	4.208
	SFA	2.708(+0.313)↑	2.354(+0.021)↑	3.604(+0.125)↑	4.521(-0.083)	4.021(+0.042)↑	4.125(-0.083)
GenMT	SR	2.082	3.224	3.633	3.653	2.878	4.020
	PR	2.739(+0.657)↑	2.551(-0.673)	3.482(-0.151)	3.739(+0.086)↑	3.727(+0.849)↑	5.065(+1.045)↑
	SIA	2.245	2.367	3.469	3.857	3.694	5.367
	SFA	2.510(+0.265)↑	2.327(-0.041)	3.551(+0.082)↑	3.776(-0.082)	3.694(+0.)	5.286(-0.082)
GenCN	SR	2.280	1.820	4.140	4.340	2.740	3.620
	PR	2.876(+0.596)↑	2.008(+0.188)↑	4.200(+0.060)↑	4.424(+0.084)↑	3.440(+0.700)↑	4.464(+0.844)
	SIA	2.400	1.820	4.140	4.440	3.500	4.700
	SFA	2.660(+0.260)↑	1.900(+0.080)↑	4.280(+0.140)↑	4.500(+0.060)↑	3.580(+0.080)↑	4.600(-0.100)

Table 3: **Model Rankings** under Different Evaluation Protocols. This table reports model rankings under four evaluation protocols (SR, PR, SIA, and SFA) on the GenOL (GenOverall), GenMT (GenMath) and GenCN(GenChinese), where lower values indicate better performance (1 = best, 6 = worst).

2.929 under PR, a gain of 0.408 ranking positions under peer evaluation. Similar trends appear on GenMath and GenChinese, and additional results on other benchmarks are reported in Table 15. By contrast, the gap between SIA and SFA under `Kemeny-Young` is consistently small, suggesting that aggregation substantially reduces the influence of self preference bias and yields more balanced rankings. Figure 4 further shows that SR has the weakest and most variable alignment with human judgment, whereas SIA achieves the strongest overall alignment and SFA remains comparable, indicating that game theoretic aggregation mitigates self preference bias even when self evaluations are included. Due to space limitations, we report 3 representative benchmarks in Table 3 and Figure 4, with full results in Appendix G.

4.4 Q3: Which Abilities of LLMs Can Be Reliably Evaluated Through Game Theory?

Game-Theoretic Evaluation Yields the Most Human-Aligned Rankings in Math Tasks. Figure 6 compares how Kemeny–Young aggregation aligns with human preferences across different LLM capabilities. Alignment is strongest on GSM8K, which achieves the highest macro-level Pearson correlation (0.941) together with a relatively tight micro-level distribution. This pattern is expected because math problems admit clearer correctness signals, so peer rankings are more likely to reflect a shared ordering. By contrast, benchmarks such as CEval and Creative Writing show lower and more variable correlations, suggesting that when judgments depend more on style, nuance, or task interpretation, local ranking signals become less stable. Overall, these results indicate that peer-based aggregation is most reliable when evaluation criteria are relatively objective, but it remains useful in more subjective settings after averaging across many questions. This helps explain why `Creative Writing` still attains high macro-level alignment (0.914) despite larger micro-level variance: individual judgments are noisy, yet the aggregate ranking continues to capture broad human consensus.

5 Conclusion

We study LLM evaluation from a game-theoretic perspective and argue that assessing model capabilities is fundamentally a ranking aggregation problem rather than a purely score-based comparison. We propose a decentralized peer ranking framework in which LLMs act as both evaluators and evaluatees, and individual rankings are aggregated using principled aggregation methods. Across diverse benchmarks, we show that aggregation-based approaches, particularly consensus-based methods such as Kemeny–Young, achieve stronger and more stable alignment with human judgments than single-model or accuracy-based evaluations, while substantially mitigating self-preference bias. Importantly, these advantages persist under sparse comparison regimes, enabling scalable evaluation without modifying the aggregation procedure. While alignment varies across task types, being strongest for domains with clear correctness signals and weaker for more subjective tasks such as creative writing, aggregate rankings nevertheless capture human consensus at the dataset level.

6 Limitations.

Our evaluation framework relies on model-generated judgments as a proxy for human preferences. While game-theoretic aggregation mitigates individual biases at the aggregate level, it assumes that models have sufficient evaluative capacity to produce meaningful global rankings. In addition, our experiments are conducted on a limited set of models and benchmarks, and the scalability extension based on expander graphs is explored only in a preliminary setting. Finally, we observe that alignment is more stable for tasks with clearer evaluation criteria, such as mathematical reasoning, while performance on more linguistically nuanced or subjective tasks remains more variable. A more comprehensive investigation of these factors, including larger model pools and a broader range of task types, is left for future work.

References

Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and Shuang Zhou. 2025. Amo-bench: Large language models still struggle in high school math competitions. *CoRR*, abs/2510.26768.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *CoRR*, abs/2108.07732.

Felix Brandt. 2009. Some remarks on dodgson's voting rule. *Mathematical Logic Quarterly*, 55(4):460–463.

Moses Charikar, Alexandra Lassota, Prasanna Ramakrishnan, Adrian Vetta, and Kangning Wang. 2024a. Six candidates suffice to win a voter majority. *CoRR*, abs/2411.03390.

Moses Charikar, Prasanna Ramakrishnan, Kangning Wang, and Hongxun Wu. 2024b. Breaking the metric voting distortion barrier. *J. ACM*, 71(6):42:1–42:33.

Junjie Chen, Weihang Su, Zhumin Chu, Haitao Li, Qinyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024. An automatic and cost-efficient peer-review framework for language generation evaluation. *CoRR*, abs/2410.12265.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024a. Automatic large language model evaluation via peer review. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 384–393. ACM.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024b. PRE: A peer review based large language model evaluator. *CoRR*, abs/2401.15641.

Lucia Cipolina-Kun, Marianna Nezhurina, and Jenia Jitsev. 2025. Game reasoning arena: A framework and benchmark for assessing reasoning capabilities of large language models via game play. *CoRR*, abs/2508.03368.

Andreea Deac, Marc Lackenby, and Petar Veličković. 2022. Expander graph propagation. In *Learning on Graphs Conference*, pages 38–1. PMLR.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Llm-evaluation tropes: Perspectives on the validity of llm-evaluations. *Preprint*, arXiv:2504.19076.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *CoRR*, abs/2402.12348.

Lingyue Fu, Xin Ding, Yaoming Zhu, Shao Zhang, Lin Qiu, Weiwen Liu, Weinan Zhang, Xuezhi Cao, Xunliang Cai, Jiaxin Ding, and Yong Yu. 2025. Catarena: Evaluation of LLM agents through iterative tournament competitions. *CoRR*, abs/2510.26852.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 80 others. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

John G Kemeny. 1959. Mathematics without numbers. *Daedalus*, 88(4):577–591.

Maurice George Kendall. 1948. Rank correlation methods.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *Preprint*, arXiv:2305.13711.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2025. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *Preprint*, arXiv:2403.16950.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. Llms as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12688–12701. Association for Computational Linguistics.

Aidan Lyon and Eric Pacuit. 2013. The wisdom of crowds: Methods of human judgement aggregation. In *Handbook of human computation*, pages 599–614. Springer.

Stefan Pasch. 2025. Ai vs. human judgment of content moderation: Llm-as-a-judge and ethics-based response refusals. *Preprint*, arXiv:2505.15365.

Ruizhong Qiu, Weiliang Will Zeng, James Ezick, Christopher Lott, and Hanghang Tong. 2025. How efficient is LLM-generated code? a rigorous & high-standard benchmark. In *The Thirteenth International Conference on Learning Representations*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Donald G Saari. 1985. The optimal ranking method is the borda count. Technical report, Discussion paper.

Donald G Saari and Vincent R Merlin. 1996. The copeland method: I.: Relationships and the dictionary. *Economic theory*, 8(1):51–76.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. 2025. Game theory meets large language models: A systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 10669–10677. ijcai.org.

Kenan Wood, Hammurabi Mendes, and Jonad Pulaj. 2024. Distributed agreement in the arrovian framework. In *28th International Conference on Principles of Distributed Systems, OPODIS 2024, December 11-13, 2024, Lucca, Italy*, volume 324 of *LIPIcs*, pages 32:1–32:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244.

Lirong Xia and Weiqiang Zheng. 2022. Beyond the worst case: Semi-random complexity analysis of winner determination. In *Web and Internet Economics - 18th International Conference, WINE 2022, Troy, NY, USA, December 12-15, 2022, Proceedings*, volume 13778 of *Lecture Notes in Computer Science*, pages 330–347. Springer.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *CoRR*, abs/2404.18824.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-llm consensus and human review. *Preprint*, arXiv:2503.17620.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang, and Dongzhan Zhou. 2025. Llama-berry: Pairwise optimization for olympiad-level mathematical reasoning via o1-like monte carlo tree search. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7315–7337. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. Don't make your LLM an evaluation benchmark cheater. *CoRR*, abs/2311.01964.

A Expander-Graph-Based Sparse Evaluation

To further examine the scalability of our evaluation framework, we conduct additional experiments using expander-graph-based sparse comparison structures. Instead of requiring each model to rank the outputs of all others, the expander graph restricts comparisons to a small, well-connected subset per model while preserving global connectivity. This design reduces the evaluation complexity from $O(M^2)$ to $O(KM)$, where $K \ll M$ denotes the average number of comparisons per model.

Experiments are conducted on the IFEVAL, CEVAL, MBPP, and WRITINGBENCH benchmarks, using 6 language models: QWEN2.5-32B-INSTRUCT, QWEN3-32B, DEEPSEEK-V3.1, DEEPSEEK-V3.2-INSTRUCT, DEEPSEEK-V3.2, and DEEPSEEK-V3-250324. We consider two comparison settings. In the *full* setting, each model evaluates the outputs of all other models, resulting in a fully connected comparison graph over the 6 models. In the *exp* setting, each model evaluates only three other models, corresponding to an average comparison degree of $K = 3$ under the expander-graph-based sparse structure, substantially reducing the comparison budget.

Figure 3 reports micro-level Pearson correlation distributions between human rankings and Kemeny–Young aggregated rankings under both settings. Despite the significantly reduced number of comparisons in the *exp* setting, the resulting rankings remain largely consistent with those obtained from the full comparison graph, indicating that the expander-based design provides a practical trade-off between evaluation cost and ranking fidelity.

B Kendall–Tau Distance and Coefficient

The Kendall–Tau distance between two permutations π and σ is defined as the number of discordant pairs:

$$D(\pi, \sigma) = \sum_{i<j} \mathbf{1}\left[\text{sign}(\pi(i) - \pi(j)) \neq \text{sign}(\sigma(i) - \sigma(j))\right] \tag{1}$$

where $\text{sign}(\cdot)$ returns $+1$ or -1 for strict orderings.

The Kendall–Tau **coefficient** measures normalized alignment between two rankings:

$$\tau(\pi, \sigma) = \frac{C - D}{\binom{n}{2}},$$

where C and D denote the numbers of concordant and discordant pairs, respectively.

Example. For $\pi = [A, B, C]$ and $\sigma = [B, A, C]$, the only discordant pair is (A, B), yielding $D = 1$ and $\tau = \frac{2-1}{3} = \frac{1}{3}$.

C Kemeny–Young Rank Aggregation

Given a set of input rankings $\{\pi_k\}_{k=1}^m$, the Kemeny–Young method seeks a consensus ranking σ^* that minimizes the total Kendall–Tau distance to all inputs:

$$\sigma^* = \arg\min_{\sigma} \sum_{k=1}^{m} D(\sigma, \pi_k).$$

This optimization problem is NP-hard in general, though effective heuristics are available in practice.

Procedure.

1. **Pairwise Comparison:** For each pair of candidates (i, j), count how many rankings place i above j.

2. **Preference Graph Construction:** Build a directed graph where each edge $(i \to j)$ is weighted by the number of times i is preferred over j.

3. **Feedback Arc Set Removal:** Remove a minimum-weight set of edges to make the graph acyclic; a topological ordering of the resulting graph yields the consensus ranking.

Example. Given rankings $[A, B, C]$, $[B, A, C]$, and $[C, A, B]$, the resulting pairwise preferences induce conflicts among candidates. The ranking that minimizes the total pairwise disagreement is $[B, A, C]$.

In our implementation, we adapt this formulation by maximizing the total Kendall–Tau coefficient $\sum_k \tau(\sigma, \pi_k)$, which is equivalent to minimizing the total Kendall–Tau distance.

C.1 Theoretical Properties of Kemeny–Young Aggregation

The strong empirical performance of the Kemeny–Young method can be understood from the structure of its objective, which explicitly optimizes alignment across all pairwise comparisons.

Maximizing Aggregate Pairwise Alignment. Minimizing the sum of Kendall–Tau distances is equivalent to minimizing the total number of pairwise inversions between the consensus ranking and the input rankings. Each inversion represents a local disagreement between two candidates. By globally minimizing such disagreements, the Kemeny–Young solution preserves the largest possible fraction of pairwise preferences expressed by the evaluators, making it well suited to settings with noisy or partially inconsistent rankings.

Interpretation as Social Welfare Maximization. The Kemeny–Young objective admits a natural social choice interpretation. If each evaluator assigns unit utility to each pairwise ordering it agrees with, then the total Kendall–Tau alignment corresponds to aggregate utility across evaluators. Under this view, the Kemeny–Young ranking maximizes total group utility, providing a principled justification for its robustness in collective ranking scenarios.

Robustness to Conflicting and Biased Rankings. Unlike methods based on local elimination or score accumulation, Kemeny–Young optimizes a global objective defined over all pairwise comparisons simultaneously. This global optimization property reduces sensitivity to anomalous or biased individual rankings, including self-preference bias, as no single evaluator can disproportionately influence the outcome. As a result, the method tends to produce stable and human-aligned rankings even when individual inputs are noisy or partially unreliable.

Together, these properties help explain why Kemeny–Young aggregation consistently achieves strong alignment with human judgment in our experiments, particularly in settings characterized by evaluator disagreement and heterogeneous preferences.

D Ranking Aggregation Methods

Table 4 summarizes the ranking aggregation methods evaluated in this work and their core aggregation principles.

Average Rank Aggregation. The average rank method assigns each candidate a score equal to its mean ranking position across evaluators. Formally, the aggregated score of candidate i is

$$s_i = \frac{1}{m} \sum_{k=1}^{m} \pi_k(i),$$

and candidates are ordered in ascending order of s_i, with lower scores indicating better ranks.

Borda Count. Borda count assigns each candidate a score based on its position in each ranking, with higher-ranked positions receiving more points. For n candidates, the Borda score of candidate i is

$$s_i = \sum_{k=1}^{m} (n - \pi_k(i)),$$

and candidates are ranked by decreasing s_i.

Copeland Method. The Copeland method aggregates rankings via pairwise comparisons. For each pair of candidates (i, j), let

$$w(i, j) = \sum_{k=1}^{m} \mathbf{1}[\pi_k \text{ ranks } i \text{ above } j].$$

Candidate i receives one point for each pairwise win $w(i, j) > w(j, i)$, and the final ranking is determined by the net number of pairwise wins.

Dodgson Method. The Dodgson method seeks the ranking that is closest to being a Condorcet winner. For each candidate, it computes the minimum number of adjacent swaps in the input rankings required to make that candidate beat all others in pairwise comparisons. Candidates are ranked by increasing swap distance.

Instant-Runoff Voting (IRV). IRV iteratively eliminates the candidate with the fewest first-place votes. At each round, first-choice votes are counted, the candidate with the minimum count is removed, and ballots are redistributed according to the next preferred candidate. The elimination order induces the final ranking.

Kendall–Tau Aggregation. Kendall–Tau aggregation selects the ranking that minimizes the total Kendall–Tau distance to the input rankings:

$$\sigma^* = \arg\min_{\sigma \in \mathcal{S}_n} \sum_{k=1}^{m} D(\sigma, \pi_k),$$

where $D(\cdot, \cdot)$ denotes the Kendall–Tau distance. This formulation operates directly on complete rankings and typically assumes that all input rankings are total orderings.

Method	Aggregation Principle
Average	Computes the average ranking position of each model across all rankings and orders models by their mean rank.
Borda Count (Saari, 1985)	Assigns each model a score based on its position in each ranking, with higher-ranked positions receiving more points; models are ranked by total score.
Copeland Method (Saari and Merlin, 1996)	Aggregates rankings by counting pairwise wins and losses between models and ranking them by net win count.
Dodgson Method (Brandt, 2009)	Measures the minimum number of adjacent swaps required to make a model a Condorcet winner.
IRV (Instant-Runoff Voting) (Brandt, 2009)	Iteratively eliminates the model with the fewest first-choice votes and redistributes votes until a complete ranking is obtained.
Kendall Aggregation (Kendall, 1948)	Selects the ranking that maximizes the total Kendall–Tau correlation with the input rankings, operating directly in the ranking space and typically assuming complete rankings.
Spearman Aggregation	Selects the ranking that maximizes the total Spearman correlation with the input rankings.
Kemeny–Young (Kemeny, 1959)	Finds the ranking that maximizes alignment with pairwise majority preferences induced by the input rankings, equivalently minimizing total pairwise disagreement; this formulation naturally accommodates partial rankings.

Table 4: Summary of ranking aggregation methods evaluated in this work.

Spearman Aggregation. Spearman aggregation selects the ranking that maximizes total Spearman rank correlation with the input rankings, equivalently minimizing squared rank deviations:

$$\sigma^* = \arg\min_{\sigma \in \mathcal{S}_n} \sum_{k=1}^{m} \sum_{i=1}^{n} \left(\sigma(i) - \pi_k(i)\right)^2 .$$

Kendall–Tau Aggregation vs. Kemeny–Young under Partial Rankings. Kendall–Tau aggregation and the Kemeny–Young method are closely related but not identical in general. When all input rankings are complete and strictly ordered, maximizing the total Kendall–Tau correlation is equivalent to the Kemeny–Young objective of minimizing total pairwise disagreement (Kemeny, 1959). However, this equivalence breaks down in the presence of partial or missing rankings, where the two methods differ fundamentally in how they utilize available information.

To illustrate this difference, consider a setting with 6 candidate models and 6 evaluators.

Their rankings are given by

	model A	model B	model C	model D	model E	model F
π_1	5	1	2	4	6	3
π_2	2	1	3	4	5	6
π_3	2	3	1	4	5	6
π_4	3	2	1	4	5	6
π_5	1	3	2	4	$\perp$	$\perp$
π_6	2	3	1	4	6	5

where $\perp$ denotes missing ranks. Applying Kendall–Tau aggregation yields the consensus ranking $[3, 2, 1, 4, 5, 6]$, whereas the Kemeny–Young method produces $[2, 3, 1, 4, 5, 6]$.

The discrepancy arises from the fundamentally different treatment of incomplete information. Kendall–Tau aggregation operates on complete rankings and therefore discards any evaluator with missing values, effectively ignoring the partial ranking π_5. In contrast, the Kemeny–Young method aggregates pairwise majority preferences and can incorporate all available pairwise comparisons. Although π_5 does not specify a full ordering, it provides additional pairwise evidence favoring *model A* over *model B*, which alters the majority relation between these two models and leads to a different consensus outcome. This example illustrates that Kendall–Tau aggregation and Kemeny–Young coincide under complete rankings, while Kemeny–Young is more robust to sparse or incomplete evaluations, a property that is particularly

Dataset	Domain Focus	Sampled
CEval	Standardized Exams (Chinese)	100
IFEval	Instruction Following	100
MBPP	Program Synthesis	100
GSM8K	Elementary Math Reasoning	100
MMLU	Interdisciplinary Knowledge	114
GPQA	Complex Problem Solving	50
Creative Writing	Open-ended Generation	100

Table 5: Key Properties of Evaluation Datasets

important in decentralized peer evaluation settings.

E Dataset Statistics and Selection Criteria

To comprehensively evaluate the reasoning capabilities of LLMs, we adopt a multi-level, multi-dimensional evaluation framework. We carefully select three representative benchmark datasets—GSM8K, MMLU, and GPQA—that span tasks of varying difficulty levels and reasoning demands, as summarized in Table 5.

These benchmarks are widely recognized for their effectiveness in probing high-level cognitive capabilities in LLMs, each emphasizing distinct dimensions of reasoning:

- **GSM8K** (Chiang et al., 2024) focuses on basic mathematical reasoning. We sample 100 elementary-level arithmetic word problems to assess the model's capabilities in computation, basic logic, and semantic understanding.

- **MMLU** (Hendrycks et al., 2021) extends the evaluation to interdisciplinary knowledge integration. We select 114 multiple-choice questions across 57 subjects, spanning STEM fields (e.g., mathematics, physics) as well as humanities and social sciences (e.g., history, law), to evaluate the model's cross-domain reasoning and conceptual abstraction skills.

- **GPQA** (Rein et al., 2023) is a high-level benchmark specifically introduced in this study. It comprises 50 open-ended generative questions that require comprehensive problem solving, often involving multimodal information fusion, multi-constraint optimization, and counterfactual reasoning. Models are required to produce full natural language solutions, posing greater demands on deep reasoning, creative thinking, and knowledge transfer.

- **CEval** (Huang et al., 2023) is a Chinese-language exam-style benchmark that evaluates model performance on national-level standardized test questions across subjects such as law, medicine, and finance. We use a 100-example subset drawn from its development set, covering both humanities and STEM domains.

- **IFEval** (Zhou et al., 2023a) evaluates instruction-following capabilities of large language models. It includes a diverse range of tasks involving goal specification, constraint satisfaction, and multi-step procedural reasoning. We sample 100 instances across different task categories to assess general instruction comprehension.

- **MBPP** (Austin et al., 2021) targets program synthesis. Each instance consists of a short programming task described in natural language along with a set of unit tests. We sample 100 problems to evaluate coding proficiency and functional reasoning ability.

- **Creative Writing** (Wu et al., 2025) consists of open-ended generation tasks such as story continuation, character development, and stylistic rewriting. We sample 100 prompts to assess creativity, coherence, and narrative fluency under subjective evaluation.

- **GenMath**. GenMath is a newly constructed generative evaluation set designed to assess mathematical ability. All questions are generated using GPT-4 and are not observed during model pretraining. The prompt templates used for data generation are provided in Section H.

- **GenChinese**. GenChinese is a generative evaluation set targeting Chinese language understanding and generation. It consists of newly generated questions covering reading

comprehension, semantic reasoning, and instruction understanding in Chinese. All instances are generated using GPT-4 and are not observed during model pretraining. The prompt templates used for data generation are provided in Section H.

- **GenOverall**. GenOverall is designed to evaluate general comprehension and reasoning ability across diverse domains. The dataset includes open-ended questions spanning everyday knowledge, abstract reasoning, and cross-domain understanding. All questions are generated using GPT-4 and are not observed during model pretraining. The prompt templates used for data generation are provided in Section H.

F Supplementary experiment on Q1

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.645	0.291	0.616	0.555	0.426	0.262	0.664	0.659	0.665	0.670	0.569	0.666	0.695	0.678
	Std.	0.222	0.427	0.233	0.280	0.293	0.381	0.250	0.266	0.278	0.264	0.402	0.264	0.277	0.295
	Min	-0.086	-0.829	0.029	-0.486	-0.600	-0.429	-0.127	-0.143	-0.143	-0.143	-0.371	-0.143	-0.143	-0.200
	25%	0.486	0.029	0.429	0.371	0.257	0.043	0.562	0.543	0.557	0.600	0.143	0.600	0.600	0.614
	50%	0.714	0.314	0.771	0.657	0.486	0.257	0.764	0.714	0.743	0.771	0.429	0.771	0.771	0.771
	75%	0.771	0.600	0.771	0.771	0.657	0.543	0.809	0.871	0.871	0.829	0.543	0.829	0.886	0.886
	Max	0.943	0.943	0.886	0.943	0.829	0.943	0.975	1.000	1.000	1.000	0.734	1.000	1.000	1.000
Kendall	**Mean**	0.517	0.243	0.496	0.440	0.333	0.232	0.535	0.560	0.568	0.563	0.464	0.555	0.597	0.584
	Std.	0.200	0.350	0.189	0.229	0.235	0.301	0.236	0.231	0.253	0.236	0.402	0.236	0.243	0.252
	Min	-0.067	-0.600	0.067	-0.333	-0.467	-0.333	-0.067	-0.067	-0.067	-0.067	-0.333	-0.067	-0.067	-0.067
	25%	0.367	0.067	0.333	0.333	0.200	0.067	0.414	0.467	0.467	0.467	0.067	0.467	0.467	0.467
	50%	0.600	0.200	0.600	0.467	0.333	0.200	0.574	0.600	0.600	0.600	0.200	0.600	0.600	0.600
	75%	0.600	0.467	0.600	0.600	0.467	0.467	0.690	0.733	0.733	0.733	0.333	0.733	0.733	0.733
	Max	0.867	0.867	0.733	0.867	0.733	0.867	1.000	1.000	0.467	1.000	1.000	1.000	1.000	1.000

Table 6: Distribution of **Pearson and Kendall** correlation coefficients between model-generated rankings and human preferences on the **GSM8K** dataset. Each value reflects **micro-level alignment**, computed as the correlation between a model-generated ranking and the corresponding human ranking per question. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.376	0.165	0.367	0.315	0.166	-0.025	0.317	0.351	0.326	0.321	-0.083	0.301	0.365	0.335
	Std.	0.348	0.334	0.343	0.360	0.287	0.362	0.396	0.385	0.377	0.367	0.468	0.372	0.396	0.385
	Min	-0.371	-0.486	-0.486	-0.486	-0.543	-0.771	-0.576	-0.486	-0.486	-0.486	-0.829	-0.486	-0.486	-0.486
	25%	0.257	-0.086	0.257	0.043	0.029	-0.329	0.160	0.143	0.143	0.143	-0.486	0.086	0.143	0.129
	50%	0.371	0.200	0.371	0.371	0.200	0.029	0.316	0.371	0.343	0.343	-0.143	0.286	0.371	0.371
	75%	0.600	0.357	0.643	0.600	0.371	0.257	0.597	0.657	0.657	0.600	0.371	0.600	0.657	0.600
	Max	0.943	0.829	1.000	0.943	0.771	0.714	0.968	1.000	0.943	1.000	0.886	1.000	1.000	1.000
Kendall	**Mean**	0.328	0.141	0.307	0.275	0.142	-0.019	0.239	0.312	0.282	0.270	-0.124	0.255	0.339	0.306
	Std.	0.264	0.250	0.275	0.297	0.202	0.257	0.304	0.329	0.303	0.310	0.353	0.307	0.332	0.322
	Min	-0.200	-0.333	-0.333	-0.333	-0.333	-0.600	-0.333	-0.333	-0.333	-0.333	-0.733	-0.333	-0.333	-0.333
	25%	0.200	-0.067	0.200	0.067	0.067	-0.200	0.067	0.067	0.167	0.067	-0.367	0.067	0.200	0.067
	50%	0.333	0.067	0.333	0.333	0.133	0.000	0.200	0.333	0.200	0.267	-0.133	0.200	0.333	0.333
	75%	0.467	0.300	0.467	0.467	0.300	0.200	0.354	0.600	0.467	0.467	0.067	0.467	0.600	0.467
	Max	0.867	0.733	1.000	0.867	0.600	0.467	0.966	1.000	0.867	1.000	0.733	1.000	1.000	1.000

Table 7: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **CEval** dataset. Each value reflects micro-level alignment computed per individual problem. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.466	0.419	0.557	0.443	0.323	0.335	0.562	0.551	0.556	0.559	0.110	0.544	0.592	0.569
	Std.	0.355	0.378	0.319	0.368	0.388	0.376	0.304	0.318	0.288	0.299	0.392	0.323	0.285	0.310
	Min	-0.371	-0.543	-0.429	-0.371	-0.714	-0.543	-0.328	-0.543	-0.371	-0.143	-0.829	-0.371	-0.371	-0.371
	25%	0.257	0.143	0.386	0.143	0.100	0.029	0.376	0.371	0.371	0.371	-0.143	0.371	0.486	0.429
	50%	0.486	0.400	0.629	0.486	0.314	0.371	0.633	0.600	0.600	0.600	0.143	0.600	0.600	0.600
	75%	0.714	0.757	0.771	0.771	0.586	0.600	0.787	0.771	0.771	0.771	0.429	0.771	0.771	0.771
	Max	1.000	1.000	1.000	1.000	1.000	0.943	0.945	1.000	1.000	1.000	0.771	1.000	1.000	1.000
Kendall	**Mean**	0.371	0.347	0.459	0.344	0.264	0.263	0.420	0.450	0.448	0.448	-0.045	0.437	0.497	0.472
	Std.	0.330	0.343	0.282	0.326	0.337	0.288	0.290	0.296	0.282	0.284	0.318	0.297	0.279	0.292
	Min	-0.333	-0.333	-0.333	-0.333	-0.467	-0.467	-0.276	-0.467	-0.333	-0.200	-0.733	-0.200	-0.333	-0.333
	25%	0.200	0.067	0.333	0.067	0.067	0.067	0.200	0.333	0.333	0.200	-0.333	0.200	0.333	0.333
	50%	0.333	0.333	0.467	0.333	0.200	0.333	0.414	0.467	0.467	0.467	0.067	0.467	0.467	0.467
	75%	0.600	0.600	0.600	0.600	0.467	0.467	0.600	0.600	0.600	0.600	0.200	0.600	0.600	0.600
	Max	1.000	1.000	1.000	1.000	1.000	0.867	0.867	1.000	1.000	1.000	0.600	1.000	1.000	1.000

Table 8: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **GPQA** dataset. Each value reflects micro-level alignment computed per individual problem. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.550	0.241	0.600	0.493	0.110	-0.048	0.535	0.600	0.570	0.565	-0.015	0.543	0.621	0.587
	Std.	0.350	0.460	0.319	0.342	0.520	0.440	0.396	0.362	0.358	0.370	0.431	0.379	0.366	0.374
	Min	-0.371	-0.829	-0.314	-0.600	-0.829	-0.943	-0.743	-0.600	-0.600	-0.657	-1.000	-0.657	-0.543	-0.543
	25%	0.386	-0.029	0.429	0.271	-0.314	-0.371	0.338	0.429	0.371	0.371	-0.314	0.314	0.429	0.371
	50%	0.657	0.314	0.714	0.571	0.143	-0.029	0.638	0.714	0.657	0.600	0.086	0.600	0.714	0.714
	75%	0.829	0.586	0.829	0.714	0.543	0.300	0.835	0.829	0.829	0.829	0.257	0.829	0.886	0.829
	Max	0.943	0.943	0.943	0.943	0.886	0.829	0.959	1.000	1.000	1.000	0.771	1.000	1.000	1.000
Kendall	**Mean**	0.453	0.203	0.501	0.397	0.094	-0.045	0.450	0.518	0.486	0.483	-0.080	0.459	0.548	0.505
	Std.	0.306	0.379	0.289	0.297	0.412	0.359	0.347	0.331	0.330	0.345	0.354	0.356	0.336	0.337
	Min	-0.333	-0.733	-0.200	-0.467	-0.733	-0.867	-0.600	-0.467	-0.467	-0.600	-1.000	-0.600	-0.333	-0.333
	25%	0.333	-0.067	0.333	0.200	-0.200	-0.333	0.215	0.333	0.333	0.333	-0.333	0.200	0.333	0.333
	50%	0.533	0.267	0.600	0.467	0.067	-0.067	0.467	0.600	0.600	0.467	-0.067	0.467	0.600	0.600
	75%	0.733	0.467	0.733	0.600	0.467	0.200	0.690	0.733	0.733	0.733	0.200	0.733	0.733	0.733
	Max	0.867	0.867	0.867	0.867	0.733	0.733	1.000	1.000	1.000	1.000	0.600	1.000	1.000	1.000

Table 9: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **IFEval** dataset. Each value reflects micro-level alignment computed per individual problem. TThe left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.555	0.209	0.515	0.380	0.261	-0.016	0.555	0.613	0.551	0.540	0.086	0.527	0.596	0.574
	Std.	0.303	0.340	0.316	0.397	0.364	0.439	0.382	0.348	0.373	0.366	0.397	0.368	0.343	0.348
	Min	-0.543	-0.714	-0.600	-0.543	-0.714	-0.771	-0.696	-0.543	-0.543	-0.543	-1.000	-0.543	-0.543	-0.543
	25%	0.429	-0.029	0.371	0.200	0.029	-0.300	0.440	0.457	0.371	0.371	-0.143	0.371	0.457	0.429
	50%	0.629	0.257	0.514	0.371	0.314	0.086	0.596	0.657	0.600	0.543	0.200	0.543	0.657	0.600
	75%	0.771	0.471	0.757	0.714	0.529	0.257	0.866	0.886	0.857	0.829	0.371	0.829	0.914	0.829
	Max	0.886	0.771	0.943	0.886	0.886	0.886	0.990	1.000	1.000	1.000	0.714	1.000	1.000	1.000
Kendall	**Mean**	0.447	0.157	0.419	0.320	0.219	0.003	0.436	0.549	0.489	0.464	0.004	0.450	0.538	0.518
	Std.	0.246	0.273	0.281	0.321	0.296	0.339	0.326	0.301	0.326	0.316	0.301	0.309	0.312	0.312
	Min	-0.333	-0.600	-0.467	-0.467	-0.600	-0.600	-0.414	-0.333	-0.333	-0.333	-1.000	-0.333	-0.333	-0.333
	25%	0.333	-0.067	0.233	0.200	0.067	-0.200	0.200	0.400	0.333	0.200	-0.200	0.200	0.333	0.333
	50%	0.467	0.200	0.467	0.333	0.200	0.067	0.414	0.600	0.467	0.467	0.067	0.467	0.600	0.467
	75%	0.600	0.333	0.600	0.600	0.467	0.200	0.690	0.800	0.733	0.733	0.200	0.733	0.800	0.733
	Max	0.733	0.600	0.867	0.733	0.733	0.733	1.000	1.000	1.000	1.000	0.467	1.000	1.000	1.000

Table 10: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **MBPP** dataset. Each value reflects micro-level alignment computed per individual problem. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.656	0.389	0.618	0.503	0.345	0.198	0.648	0.681	0.664	0.662	0.050	0.644	0.707	0.675
	Std.	0.267	0.359	0.265	0.305	0.362	0.368	0.243	0.237	0.252	0.249	0.406	0.258	0.236	0.253
	Min	-0.200	-0.829	-0.371	-0.200	-0.829	-0.543	-0.138	-0.429	-0.200	-0.086	-1.000	-0.257	0.029	-0.086
	25%	0.543	0.143	0.429	0.257	0.143	-0.029	0.557	0.600	0.600	0.543	-0.143	0.543	0.600	0.600
	50%	0.771	0.429	0.657	0.600	0.371	0.229	0.687	0.771	0.771	0.771	0.029	0.714	0.771	0.771
	75%	0.829	0.643	0.829	0.771	0.600	0.471	0.824	0.829	0.829	0.829	0.371	0.829	0.829	0.829
	Max	1.000	1.000	1.000	1.000	1.000	0.829	0.978	1.000	1.000	1.000	0.771	1.000	1.000	1.000
Kendall	**Mean**	0.533	0.302	0.496	0.399	0.274	0.164	0.528	0.574	0.566	0.556	-0.043	0.536	0.607	0.574
	Std.	0.259	0.284	0.264	0.272	0.289	0.280	0.241	0.221	0.240	0.239	0.329	0.242	0.235	0.246
	Min	-0.067	-0.600	-0.333	-0.067	-0.600	-0.467	-0.138	-0.200	-0.200	-0.067	-1.000	-0.200	-0.067	-0.200
	25%	0.333	0.067	0.333	0.200	0.067	-0.067	0.414	0.467	0.467	0.467	-0.333	0.467	0.467	0.467
	50%	0.600	0.333	0.533	0.467	0.200	0.200	0.552	0.600	0.600	0.600	-0.067	0.600	0.600	0.600
	75%	0.733	0.467	0.733	0.600	0.467	0.333	0.690	0.733	0.733	0.733	0.200	0.733	0.733	0.733
	Max	1.000	1.000	1.000	1.000	1.000	0.733	1.000	1.000	1.000	1.000	0.600	1.000	1.000	1.000

Table 11: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **MMLU** dataset. Each value reflects micro-level alignment computed per individual problem. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Pearson	**Mean**	0.375	0.135	0.465	0.303	0.197	0.105	0.410	0.482	0.445	0.406	-0.113	0.386	0.486	0.449
	Std.	0.413	0.475	0.328	0.469	0.377	0.422	0.490	0.469	0.448	0.489	0.442	0.509	0.447	0.474
	Min	-0.486	-0.829	-0.543	-0.600	-0.600	-0.771	-0.894	-0.886	-0.886	-0.886	-1.000	-0.886	-0.771	-0.771
	25%	0.029	-0.200	0.257	-0.071	-0.086	-0.200	0.283	0.371	0.286	0.286	-0.400	0.257	0.314	0.257
	50%	0.371	0.086	0.429	0.429	0.114	0.114	0.643	0.657	0.600	0.543	-0.086	0.543	0.600	0.600
	75%	0.714	0.543	0.700	0.714	0.543	0.371	0.741	0.829	0.743	0.686	0.200	0.686	0.800	0.800
	Max	0.943	0.943	0.943	1.000	0.943	0.886	0.876	0.943	0.943	0.943	0.714	0.943	0.943	0.943
Kendall	**Mean**	0.299	0.105	0.368	0.259	0.133	0.117	0.296	0.406	0.368	0.338	-0.153	0.312	0.424	0.381
	Std.	0.347	0.385	0.301	0.391	0.297	0.312	0.433	0.401	0.386	0.412	0.357	0.426	0.385	0.415
	Min	-0.333	-0.733	-0.467	-0.467	-0.467	-0.600	-0.828	-0.733	-0.733	-0.733	-1.000	-0.733	-0.600	-0.600
	25%	0.067	-0.200	0.200	-0.067	-0.067	-0.067	0.200	0.267	0.200	0.200	-0.333	0.200	0.267	0.200
	50%	0.333	0.067	0.333	0.333	0.067	0.067	0.333	0.467	0.467	0.467	-0.067	0.333	0.467	0.467
	75%	0.467	0.333	0.567	0.600	0.333	0.333	0.600	0.733	0.600	0.600	0.067	0.600	0.667	0.667
	Max	0.867	0.867	0.867	1.000	0.867	0.733	0.828	0.867	0.867	0.867	0.467	0.867	0.867	0.867

Table 12: Distribution of **Pearson** and **Kendall** correlation coefficients between automated rankings and human ranking on the **Creative Writing** dataset. Each value reflects micro-level alignment computed per individual problem. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Spm	Kem	Ken
GenOverall	**Mean**	0.307	0.163	0.371	0.331	0.167	0.116	0.384	0.452	0.401	0.414	0.390	0.436	0.425
	Std.	0.449	0.446	0.445	0.449	0.427	0.439	0.453	0.389	0.418	0.432	0.446	0.426	0.431
	Min	-0.771	-0.771	-0.657	-0.657	-0.657	-0.829	-0.683	-0.657	-0.657	-0.657	-0.657	-0.657	-0.657
	25%	-0.029	-0.143	-0.029	-0.029	-0.143	-0.200	0.166	0.143	0.143	0.143	0.114	0.114	0.086
	50%	0.486	0.229	0.486	0.429	0.257	0.086	0.532	0.600	0.543	0.543	0.486	0.600	0.600
	75%	0.643	0.543	0.714	0.686	0.543	0.457	0.736	0.743	0.714	0.743	0.714	0.743	0.743
	Max	0.943	0.829	1.000	1.000	0.829	0.943	0.983	1.000	0.943	1.000	1.000	1.000	1.000
GenMath	**Mean**	0.553	0.217	0.506	0.239	0.203	0.320	0.486	0.562	0.513	0.528	0.502	0.563	0.536
	Std.	0.294	0.441	0.324	0.415	0.426	0.427	0.393	0.378	0.391	0.366	0.384	0.342	0.337
	Min	-0.200	-0.943	-0.429	-0.714	-0.886	-0.714	-0.713	-0.714	-0.714	-0.714	-0.714	-0.314	-0.314
	25%	0.429	-0.129	0.314	-0.029	-0.129	0.043	0.302	0.429	0.371	0.371	0.257	0.371	0.371
	50%	0.600	0.257	0.543	0.200	0.257	0.371	0.559	0.657	0.657	0.600	0.600	0.657	0.600
	75%	0.771	0.600	0.829	0.586	0.586	0.643	0.772	0.829	0.771	0.829	0.829	0.829	0.829
	Max	0.943	0.943	0.943	0.943	0.943	0.943	0.962	1.000	1.000	1.000	0.943	1.000	1.000
GenChinese	**Mean**	0.458	0.227	0.431	0.425	0.274	0.178	0.466	0.535	0.491	0.482	0.475	0.514	0.489
	Std.	0.415	0.417	0.412	0.428	0.400	0.437	0.448	0.409	0.415	0.440	0.442	0.423	0.436
	Min	-0.600	-0.543	-0.771	-0.771	-0.771	-0.543	-0.903	-0.714	-0.714	-0.886	-0.943	-0.714	-0.714
	25%	0.286	-0.086	0.286	0.243	-0.086	-0.300	0.271	0.386	0.157	0.286	0.286	0.200	0.157
	50%	0.571	0.371	0.543	0.543	0.429	0.257	0.563	0.657	0.600	0.600	0.600	0.600	0.600
	75%	0.771	0.600	0.700	0.771	0.600	0.543	0.813	0.829	0.829	0.829	0.829	0.829	0.814
	Max	1.000	0.829	1.000	0.943	0.886	0.943	0.968	1.000	1.000	1.000	1.000	1.000	1.000

Table 13: **Distribution of Pearson correlation coefficients** between model-generated rankings and human preferences on the `GenOverall`, `GenMath` and `GenChinese` dataset. Correlations are computed at the micro level as the aglinment between model rankings and human rankings for each question. The left columns show the results for individual LLMs, while the right columns report outcomes after applying different aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment. Color intensity reflects correlation strength. Abbreviations: Avg(average voting), Dod(Dodgson), Cop(Copeland), Bor(Borda), Irv(instant-runoff voting), Spm(Spearman), Kem(Kemeny-Young), Ken(Kendall). Models: 4o11(gpt-4o-20241120), 4o05(gpt-4o-20240513), 4o08(gpt-4o-20240806), Sn10(claude-3.5-Sonnet-20241022), Hk10(claude-3.5-haiku-20241022), Op02(claude-3-opus-20240229).

Traditional evaluation of LLMs often depends on *static datasets*, which are fixed sets of benchmark questions accompanied by reference answers, such as MMLU (Hendrycks et al., 2021) and GSM8K (Chiang et al., 2024). While these datasets consisten consistent and reproducible comparisons, they suffer from inherent limitations. In particular, many static benchmarks have become publicly available and widely used, increasing the risk of data leakage (Zhou et al., 2023b) and benchmark saturation (Xu et al., 2024), leading to inflated performance estimates.

As shown in Table 13, despite being constructed synthetically, the generated datasets yield alignment results highly consistent with those observed in Experiment 4.2 **Q1**. Our game-theoretic evaluation framework continues to produce rankings that strongly correlate with human preferences across both Pearson and Kendall metrics. Notably, methods such as Kemeny-Young (`Kem`) again demonstrate the highest median and upper-quartile correlations, reaffirming their robustness. These findings demonstrate that our peer-based aggregation approach is robust to potential data leakage in the benchmark. Furthermore, it maintains strong alignment with human preferences even on entirely novel, LLM-generated datasets, highlighting its reliability across a wide range of evaluation conditions.

		4o11	Sn10	4o05	4o08	Op02	Hk10	Avg	Dod	Cop	Bor	Irv	Spm	Kem	Ken
Kendall	**Mean**	0.245	0.131	0.291	0.261	0.139	0.092	0.322	0.370	0.322	0.348	-0.237	0.328	0.362	0.350
	Std.	0.336	0.326	0.343	0.357	0.317	0.345	0.359	0.340	0.328	0.358	0.386	0.368	0.355	0.353
	Min	-0.600	-0.600	-0.467	-0.467	-0.467	-0.733	-0.467	-0.467	-0.467	-0.467	-0.867	-0.467	-0.467	-0.467
	25%	0.067	-0.067	0.067	0.067	-0.067	-0.133	0.067	0.133	0.133	0.067	-0.467	0.067	0.133	0.067
	50%	0.333	0.133	0.333	0.200	0.200	0.067	0.414	0.467	0.467	0.467	-0.333	0.333	0.467	0.467
	75%	0.467	0.333	0.467	0.467	0.433	0.333	0.552	0.600	0.600	0.600	0.067	0.600	0.600	0.600
	Max	0.867	0.733	1.000	1.000	0.733	0.867	1.000	1.000	0.867	1.000	0.600	1.000	1.000	1.000

Table 14: Distribution of **Kendall** correlation coefficients between automated rankings and human rankings on the `GenOverall` dataset. Correlations are computed at the micro level for each question. The left columns report individual LLMs, and the right columns report aggregation algorithms. Summary statistics include the mean, standard deviation, minimum, quartiles, and maximum. Higher values indicate stronger alignment with human judgment, and color intensity reflects correlation strength.

G Supplementary experiment on Q2

Dataset	Method	**4o-1120**	**Sn-1022**	**4o-0513**	**4o-0806**	**Hk-1022**	**Op-0229**
GSM8K	SR	2.000	3.000	3.420	3.44	3.860	5.000
	PR	2.44(+0.44) ↑	2.764(-0.236)	3.188(-0.232)	3.256(-0.184)	4.352(+0.492) ↑	5.056(+0.056) ↑
	SIA	1.74	2.520	3.08	3.48	4.68	5.500
	SFA	1.920(+0.180) ↑	2.700(+0.180) ↑	3.060(-0.020)	3.320(-0.160)	4.720(+0.04) ↑	5.460(-0.04)
GPQA	SR	1.900	3.74	3.060	4.460	3.714	4.020
	PR	1.897(-0.003)	3.451(-0.289)	2.927(-0.133)	4.552(+0.092) ↑	3.952(+0.238) ↑	4.244(+0.224) ↑
	SIA	1.48	3.44	2.760	4.900	4.020	4.4
	SFA	1.520(+0.04) ↑	3.520(+0.08) ↑	2.820(+0.060) ↑	4.78(-0.120) ↑	4.08(+0.060) ↑	4.24(-0.160)
MMLU	SR	1.64	3.500	2.167	4.632	3.886	4.265
	PR	2.061(+0.420) ↑	3.381(-0.119)	2.881(+0.714) ↑	4.461(-0.170)	3.958(+0.072) ↑	4.437(+0.171) ↑
	SIA	1.386	3.325	2.272	5.000	4.211	4.807
	SFA	1.649(+0.263) ↑	3.377(+0.053) ↑	2.535(+0.263) ↑	4.702(-0.298)	4.211(0.0)	4.895(+0.088) ↑
CEval	SR	3.000	3.100	2.820	3.160	4.375	3.674
	PR	3.433(+0.433) ↑	2.887(-0.213)	3.489(+0.669) ↑	2.828(-0.332)	4.932(+0.557) ↑	3.604(-0.070)
	SIA	3.260	2.760	2.980	2.840	5.280	3.880
	SFA	3.500(+0.240) ↑	2.620(-0.140)	3.220(+0.240) ↑	2.680(-0.160)	5.120(-0.160)	3.740(-0.140)
IFEval	SR	1.840	3.580	3.420	3.360	3.580	3.265
	PR	2.756(+0.916) ↑	3.160(-0.420)	3.423(+0.003) ↑	3.512(+0.152) ↑	4.843(+1.263) ↑	3.692(+0.427) ↑
	SIA	1.760	2.880	3.380	3.560	5.360	4.060
	SFA	2.180(+0.420) ↑	3.040(+0.160) ↑	3.500(+0.120) ↑	3.460(-0.100)	5.300(-0.060)	4.000(-0.060)
MBPP	SR	1.729	3.320	3.560	3.449	3.900	3.460
	PR	2.557(+0.828) ↑	3.528(+0.208) ↑	3.273(-0.287)	3.590(+0.141) ↑	4.486(+0.586) ↑	3.872(+0.412) ↑
	SIA	1.580	3.120	3.360	3.800	4.960	4.180
	SFA	1.780(+0.200) ↑	3.380(+0.260) ↑	3.260(-0.100)	3.760(-0.040)	4.820(-0.140)	4.300(+0.120) ↑
Writing	SR	2.420	3.314	3.200	3.740	3.460	3.825
	PR	2.665(+0.245) ↑	3.215(-0.099)	3.482(+0.282) ↑	3.552(-0.188)	4.385(+0.925) ↑	3.952(+0.127) ↑
	SIA	1.960	3.040	3.220	3.600	4.900	4.280
	SFA	2.220(+0.260) ↑	3.040(0.0)	3.300(+0.08) ↑	3.480(-0.120)	4.880(-0.020)	4.280(0.0)

Table 15: **Model Rankings under Different Evaluation Protocols.** This table presents model rankings on four representative benchmarks (GSM8K, GPQA, MMLU, CEval, IFEval, MBPP, WritingBench) under four evaluation protocols: **SR**, **PR**, **SIA**, and **SFA**, where lower values indicate better rankings (1 = best, 6 = worst). The SIA and SFA rankings are computed via Kemeny-Young aggregation, with and without the target model's own votes, respectively. Comparing these values allows us to assess the impact of self-preference and the robustness of aggregation. Models: **4o-1120** = gpt-4o-20241120, **4o-0513** = gpt-4o-20240513, **4o-0806** = gpt-4o-20240806, **Sn-1022** = Claude-3.5-Sonnet-20241022, **Hk-1022** = Claude-3.5-Haiku-20241022, **Op-0229** = Claude-3-Opus-20240229. Due to space constraints, results on additional benchmarks are provided in the appendix.

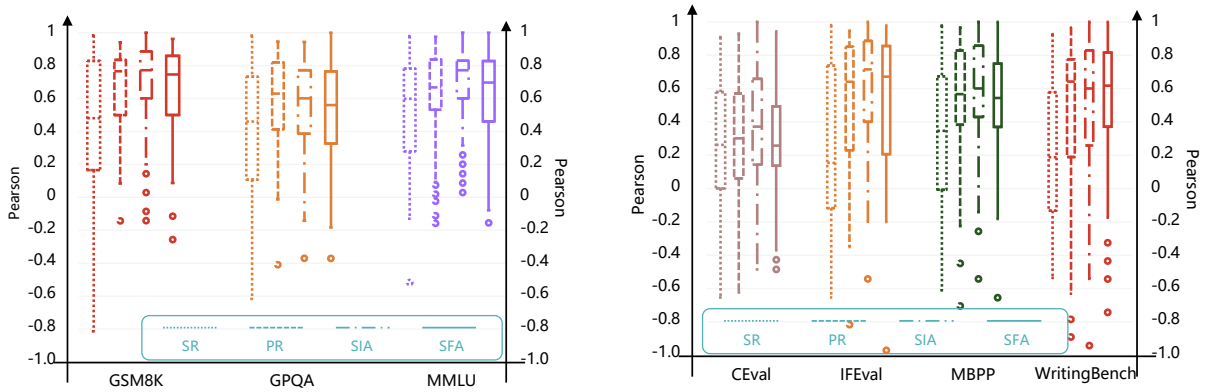

Figure 7: **Alignment with Human Judgments under Different Evaluation Protocols.** This figure reports the distribution of Pearson correlation coefficients between model-generated rankings and human preferences (from Chatbot Arena) across seven benchmarks under four evaluation protocols: **SR** (Self Ranking), **PR** (Peer Ranking), **SIA** (Self-Inclusive Aggregation), and **SFA** (Self-Free Aggregation). In each boxplot, a **higher** box indicates stronger alignment with human rankings, while a **shorter** box implies lower variance and thus more stable alignment. Compared to SR and PR, the SIA and SFA protocols yield both higher and tighter boxes, suggesting that game-theoretic aggregation not only enhances ranking accuracy but also improves robustness against self-bias. Additional results are provided in the appendix due to space constraints.

H Generation Prompts for Capability-Specific Benchmarks

High Difficulty Math Problems Generation Named GenMath

Please generate **50 high-difficulty math problems** covering the following fields:

- Advanced Algebra
- Mathematical Analysis
- Number Theory
- Combinatorics
- Geometry
- Differential Equations
- Mathematical Logic and Set Theory
- Linear Algebra
- Probability and Statistics
- Elementary Topology

Requirements:

1. Output format should be a JSON array in the following structure:

```
[
  {"id": id, "question": question}
]
```

2. Mathematical expressions and formulas should be written using Markdown math syntax, enclosed in `$...$` for inline math or `$$...$$` for display equations.
3. All questions should be written in **English**, with clear and precise language.

High-quality Chinese Question Generation Named GenChinese

Please generate **50 Chinese language-related tasks** that cover a comprehensive range of linguistic dimensions. These tasks should be suitable for applications such as:

- Phonetics and Phonology
- Vocabulary and Word Formation
- Grammar and Syntax
- Rhetoric and Stylistics
- Semantics and Pragmatics
- Language Knowledge and Usage
- Reading Comprehension

- Language Expression and Writing

Requirements:

1. Output format should be a JSON array in the following structure:

```
[
  {"id": id, "question": question}
]
```

2. All questions should be written in **Chinese**, with clarity and appropriateness for use in linguistics research, teaching, test design, or LLM training.

Comprehensive Evaluation of LLM Capabilities Problem Generation Named GenOverall

Please generate **50 evaluation questions** designed to comprehensively assess the capabilities of large language models (LLMs). The questions should span a wide range of skills and reasoning dimensions.

Requirements:

1. Output format should be a JSON array in the following structure:

```
[
  {"id": id, "question": question}
]
```

2. All questions should be written in **English**, using clear, precise, and instruction-oriented language.

I Prompts

Answer Ranking Prompt Design for Overall

You are a reviewer assigned to rank multiple solutions to a given question. Your evaluation must be based solely on the following three criteria:

- **Accuracy**: How correct and relevant is the information?
- **Logical Consistency**: How coherent and well-reasoned is the explanation?
- **Fluency**: How clear and natural is the language?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}
[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}
Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.

Answer Ranking Prompt Design for Mathematical Problem

You are a reviewer assigned to rank multiple solutions to the same math problem. Your evaluation must be based solely on the following three criteria:

- **Accuracy**: Is the mathematical reasoning correct, and does the solution produce the correct answer?
- **Logical Rigor**: Is the problem-solving process well-structured, justified, and logically sound at each step?
- **Clarity of Explanation**: Is the reasoning clearly explained, using appropriate notation and terminology?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}
[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}
Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.

Answer Ranking Prompt Design for Chinese

You are a reviewer assigned to rank multiple answers written in Chinese. Your evaluation must be based solely on the following three criteria:

- **Linguistic Accuracy**: Are the grammar, vocabulary, and expressions consistent with standard modern Chinese?
- **Clarity of Expression**: Is the language smooth, natural, and easy to understand? Is the logic clearly conveyed?
- **Contextual Appropriateness**: Does the response match the intended tone, audience, and context of the prompt?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}

[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}
Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.

Answer Ranking Prompt Design for Instruction Following

You are a reviewer assigned to rank multiple responses to the same instruction. Your evaluation must be based solely on the following three criteria:

- **Task Completion**: Does the response fully and accurately follow all aspects of the given instruction?
- **Interpretation Accuracy**: Does the response show a correct understanding of the instruction's intent?
- **Relevance and Focus**: Is the content tightly aligned with the instruction, without going off-topic or omitting key parts?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}
[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}

Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.

Answer Ranking Prompt Design for Code Implementation

You are a reviewer assigned to rank multiple code implementations. Your evaluation must be based solely on the following three criteria:

- **Correctness**: Does the code run successfully and meet all functional requirements of the prompt?
- **Logical Clarity**: Is the code logically structured and easy to follow?
- **Readability**: Are variable names meaningful, comments helpful, and formatting clean and maintainable?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}
[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}
Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.

Answer Ranking Prompt Design for Creative Writing

You are a reviewer assigned to rank multiple creative writing pieces. Your evaluation must be based solely on the following three criteria:

- **Originality**: Is the content imaginative and unique? Does it offer a fresh perspective or concept?
- **Structural Coherence**: Is the narrative or composition logically organized and well-developed from beginning to end?
- **Expressive Quality**: Is the language vivid, engaging, and emotionally resonant?

Please strictly follow the format below:
Here is the question and the options:
[Question]
{{question.strip()}}
{% for choice in choices %}
{{choice}}
{% endfor %}
—
[Solution - 1]
{{resps[0]}}
[Solution - 2]
{{resps[1]}}
[Solution - 3]
{{resps[2]}}
[Solution - 4]
{{resps[3]}}
[Solution - 5]
{{resps[4]}}
[Solution - 6]
{{resps[5]}}
Output Format (Rank from best to worst):
1. Solution x
2. Solution y
3. Solution z
...
You must rank **all six solutions**, without skipping or tying any of them. **Do not add any comments or explanations.** Only return the final ordered list by solution number.