2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

LLM Evaluation Panel Selection Based on Cross Assessment and Similarity Matrix

Zhixiang Yang, Rongduo Han[0009-0007-5157-4660], Xiaoteng Pan, Liming Kang, Meiping Wang, Nan Gao, Haining Zhang(✉)

College of Software, Nankai University, No. 23 Hongda Street, Tianjin Economic-Technological Development Area, China
zhanghaining@nankai.edu.cn

Abstract. The evaluation of Large Language Models (LLMs) has become increasingly crucial as these models continue to advance in capabilities and complexity. The current evaluation methods primarily rely on lexical metrics and single-model scoring systems, falling short on comprehensively and accurately assessing the capabilities and performance of LLMs in semantic understanding and logical reasoning, which presents a significant challenge in developing reliable and trustworthy assessment frameworks. The contributions of the study are as follows. First, it introduces an automated approach that combines cross-model evaluation mechanisms with similarity analysis to systematically select members for the multi-model evaluation panel. Second, it validate the effectiveness of it's methodology using expert-annotated evaluation data. Experimental results demonstrate that the multi-model evaluation panel approach achieves noticeable improvement in scoring consistency versus human evaluation as compared to single-model approach.

Keywords: Large language models, Model performance evaluation, Multi-model scoring, Cross Assessment, Similarity Matrix.

1 Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) with their vast parameters and diverse training data, showcasing exceptional capabilities in tasks like understanding and reasoning [1]. However, evaluating these advanced models comprehensively and precisely remains a significant challenge[2].

Current evaluation methods for LLMs primarily rely on traditional lexical metrics (e.g., BLEU [3], ROUGE [4]) and single-model scoring approaches (typically using GPT-4 or GPT-4o as evaluator [5]). Traditional metrics, which focus on lexical comparison between generated and reference texts, prove insufficient in assessing LLMs' capabilities in semantic understanding and logical reasoning tasks. Additionally, the prevalent practice of employing a single model as evaluator, while offering some level of automation, introduces inherent biases from its training data and architecture [6]. These limitations affect the accuracy and credibility of evaluation outcomes,

particularly in assessing complex capabilities such as creative text understanding, generation and reasoning. The LLM evaluation requires more comprehensive assessment beyond simple similarity metrics [7].

Multi-Model evaluation panel offers better accuracy and fairness of LLM performance evaluation [8]. But the challenge is how to develop a more effective and systematic approach for selecting LLM evaluation panel members. The selection of appropriate evaluator models requires careful consideration of multiple factors, including model diversity, evaluation capability, and complementarity among different models. A robust selection method must not only account for individual model performance but also consider the collective evaluation quality of the entire panel, ensuring comprehensive and unbiased assessment of LLM capabilities and performance [9].

The approach of the study leverages the collective wisdom of multiple models as evaluators. The automated cross-model evaluation process is carefully designed to select diverse and quality models for the evaluation panel. Through the integration of cross-model evaluation mechanisms and similarity matrix analysis, it establishes a robust LLM evaluation panel that demonstrates improved accuracy and fairness in LLM evaluation, as validated via Chinese language comprehension and generation tasks.

The key innovations of the study are twofold. First, it introduces an automated method for constructing multi-model evaluation panels, establishing a systematic approach on evaluator selection and panel formation. Second, it develops an integrated framework that combines cross-model evaluation mechanisms with similarity matrix analysis to select panel members, demonstrating significantly improved fairness and accuracy versus single-model panel based evaluations. The multi-model panel evaluation result shows higher correlation with that of human expert evaluation.

2 Related Works

The capability and performance evaluation of LLMs has evolved significantly alongside the advancement of Natural Language Processing technologies. This section reviews current approaches in model evaluation, focusing on specialized metrics-based methods and general-purpose evaluation approaches, highlighting their respective strengths and limitations.

2.1 Specialized Metrics-based Evaluation

Specialized metrics such as BLEU and ROUGE are commonly employed in fields like machine translation and text summarization to assess lexical overlap and semantic similarity. These metrics deliver quantifiable outcomes but are limited in evaluating modern LLMs' broader capabilities, such as reasoning and contextual understanding, due to their focus on surface-level features [10]. Consequently, they fall short of providing a comprehensive assessment of model performance.

2.2 General-purpose Model-based Evaluation

The advent of advanced LLMs has ushered in a new evaluation approach where a single model, often GPT-4, assesses other LLMs' outputs [11, 12]. This method is valued for its flexibility and human-like judgments across tasks. However, it is hindered by biases in the evaluator model [13], inconsistent performance across domains, and a lack of diverse perspectives that multiple evaluators could offer.

2.3 Contribution

This research addresses these limitations by proposing a framework that systematically constructs a multi-model evaluation panel through cross-model evaluation and similarity analysis. This automated selection process ensures diversity and complementarity among evaluators, enhancing both the fairness and accuracy of LLM assessments. Experimental results demonstrate that it's multi-model panel not only outperforms single-model evaluation methods but also exhibits closer alignment with human expert judgments compared to other multi-model panel combinations, as validated through ablation studies. This approach provides a reliable and comprehensive methodology for assessing the capabilities of LLMs.

3 Method

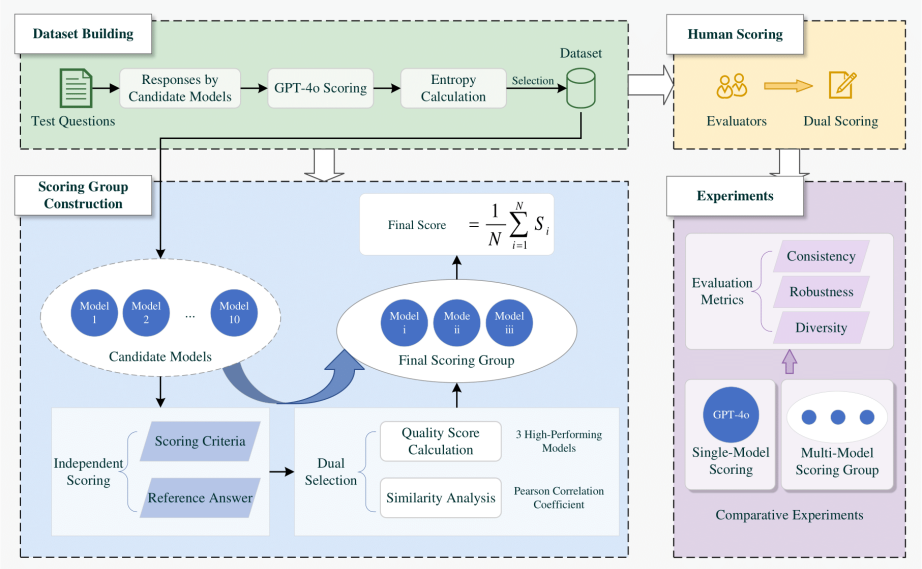

Fig. 1. Overview of the framework

This study proposes a systematic framework for constructing a reliable multi-model evaluator panel in LLM assessment. The proposed method consists of three stages:

entropy-based test case selection, expert scoring collection, and evaluator panel formation through cross-model evaluation and similarity analysis.

As shown in Fig. 1. Overview of the framework, the framework operates as follows. First, discriminative test cases are selected from GPT-4o evaluated questions using entropy-based filtering. These test cases are then scored by human experts to establish ground truth. Subsequently, each candidate evaluator model assesses these test cases, with its scoring results validated by other candidates. Finally, similarity analysis is employed to select candidate panel members and validate their collective performance against human experts.

And in this section, it also introduces how Hunyuan Large, GPT-4o, and Claude 3.5 Sonnet were selected as members of the evaluation panel through these methods.

3.1 Dataset Construction

This subsection delineates the dataset construction process, a pivotal step in evaluating large language model (LLM) combinations. The process is bifurcated into two integral phases: entropy-based question selection and human expert scoring. The former leverages entropy to quantify question discriminability, while the latter establishes authoritative ground truth scores through human expertise.

Entropy-Based Question Selection. The initial phase centers on curating questions with high discriminative power from an existing pool. The study utilizes questions from the Language Understanding and Text Generation tasks of an anonymous leaderboard, where responses from various LLMs have been previously generated and scored by GPT-4o. For each question, a prompt is crafted based on reference answers and grading rules, enabling GPT-4o to assign scores to the LLM responses. Given that multiple LLMs answer each question, a distribution of scores emerges, reflecting varying performance levels.

To quantify the discriminative power of each question, the entropy is computed based on the score distribution across LLMs, using the formula:

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

Higher entropy values signify greater score dispersion, indicating a question's ability to differentiate LLM capabilities [14]. Questions with entropy values exceeding the 75th percentile are selected to form the evaluation dataset.

The initial dataset had 1000 questions, and through entropy filtering, 250 questions with higher entropy values were selected as the final dataset.

Human Expert Scoring. The second phase establishes ground truth scores through human evaluation, enhancing the dataset's reliability. For each question selected via entropy filtering, three human experts independently assign scores, adhering strictly to predefined grading rules and reference answers. This dual-scoring approach mitigates individual bias and ensures consistency. The ground truth score for each question is then computed as the arithmetic mean of the three expert scores:

$$S_{\text{ground truth}} = \frac{S_1+S_2+S_3}{3} \tag{2}$$

These ground truth scores serve as the authoritative benchmark for evaluating the accuracy of LLM combinations in subsequent experiments. By averaging expert judgments, this process not only reinforces fairness but also provides a stable reference against which automated scoring mechanisms can be assessed.

3.2 Panel Member Selection Rules

Algorithm 1 Multi-Model Scoring Panel Construction

Input:

$M = \{m_1, m_2, \ldots, m_n\}$: Candidate models
T: Test set with 120 questions
$S = \{s_1, s_2, s_3\}$: Strong evaluation models

Output:

G: Final scoring panel
Initialize quality scores $Q \leftarrow \emptyset$
for each model $m_i \in M$ **do**
    Score test set T using m_i
    Evaluate m_i's scores using models in S
    $Q_i \leftarrow$ average of evaluations from S
    Store (m_i, Q_i) in Q
end for
Sort Q by quality scores in descending order
$m_b \leftarrow$ highest-ranked model
Compute correlation matrix C between m_b and others
Select two additional models with:
    High quality scores (Q_i)
    Low correlation with m_b (based on C)
$G \leftarrow \{m_b, m_1', m_2'\}$ **return** G

Algorithm 1 is adopted to construct an effective and diverse scoring panel.

Candidate Model Selection. Candidate models are selected based on performance and cost efficiency. Ten top-performing LLMs from mainstream leaderboards are chosen, namely GPT-4o, Claude 3.5 Sonnet, ERNIE-4.0 , Gemini 1.5 Pro, Gemini 1.5 Flash, Qwen Max, Hunyuan Large, Doubao Pro 32K, Qwen2.5 72B Instruct, Yi 1.5 34B Chat 16K. The selection criteria include demonstrated benchmark performance, architectural diversity, and proven capabilities in multiple NLP tasks. All models are accessed via public APIs with versions fixed as of January 15, 2025, to ensure fairness and consistency.

Quality Assessment. A quality assessment framework centered on language understanding and text generation is implemented. An entropy-based screening mechanism is employed to identify models capable of producing nuanced and reliable evaluations. The assessment involves performance evaluation on curated test cases representing varying complexity, analysis of scoring consistency across different language tasks, and evaluation of the ability to generate detailed, well-reasoned outputs. Three high-performing models (GPT-4o, Claude 3.5 Sonnet, and Qwen Max) perform cross-evaluation to assign quality scores that reflect both scoring accuracy and evaluation rationales.

Diversity Optimization. The similarity analysis reveals the similarity and difference in LLM behavior patterns. For instance, a high correlation (0.89) is found between Hunyuan Large and Doubao Pro 32K's scoring patterns, indicating a strong correlation between the scores given by them as evaluators. However, in order to ensure that the selected evaluation panel can accurately score on diverse questions, it is necessary to choose models with relatively low similarity between them. Therefore, when making a choice, it is necessary to avoid choosing this combination as much as possible.

This analysis leads to the prioritization of LLMs with distinct evaluation characteristics. The final panel composition is optimized through an iterative process that balances individual LLM performance with overall panel diversity.

3.3 Panel Member Selection Implementation

To construct a scoring panel with high scoring quality and diverse scoring behavior, a dual-selection strategy based on $quality\ score$ and $similarity\ analysis$ was employed. The higher the quality score, the better quality. The lower the similarity score, the better diversity. The detailed process was implemented as follows.

Quality Score Calculation and Initial Screening. A dual-selection strategy based on quality score and similarity analysis was employed to construct a scoring panel with high scoring quality and diverse evaluation behavior. Higher quality scores indicate better performance, and lower similarity scores imply greater diversity. Quality scores of candidate LLMs were calculated using a cross-evaluation mechanism in which three pre-selected high-performing models from the leaderboard (e.g. SuperCLUE [15], OpenCompass) evaluated the scoring quality of each candidate. The following table presents the detailed cross-evaluation results of the candidate LLMs.

Table 1 presents the detailed cross-evaluation results, from which Hunyuan Large, Qwen Max, and Doubao Pro 32K emerged as top candidates. However, further screening based on rating similarity was performed because high similarity among ratings indicates reduced diversity, potentially compromising scoring robustness when addressing various question types.

Table 1. Cross-Evaluation Results and Quality Score Rankings of Candidate LLMs.

Model Name	GPT-4o	Qwen Max	Claude 3.5 Sonnet	Average
Hunyuan Large	87.27	81.69	87.01	85.34
Claude 3.5 Sonnet	86.07	81.92	86.87	84.95
Doubao Pro 32K	87.08	81.36	84.76	84.40
GPT-4o	87.21	79.74	85.59	84.18
Gemini 1.5 Flash	85.94	79.38	85.79	83.71
Qwen2.5 72B Instruct	84.63	80.19	85.97	83.60
Qwen Max	84.67	80.05	85.02	83.25
Gemini 1.5 Pro	84.50	78.83	83.79	82.37
ERNIE-4.0	80.50	73.60	80.61	78.24
Yi 1.5 34B Chat 16K	77.56	64.17	78.67	73.47

Similarity Analysis and Optimization. A comprehensive similarity analysis was then conducted based on the Pearson correlation coefficients between the scoring patterns of candidate LLMs, as illustrated in Fig. 2. This analysis revealed patterns in LLM behavior, such as the high correlation (0.89) between Hunyuan Large and Doubao Pro 32K, suggesting potential redundancy. On this basis, candidates with high quality and low inter-model similarity were prioritized. Quantitative metrics designate Hunyuan Large, with a quality score of 85.34, as the primary panel member, while GPT-4o and Claude 3.5 Sonnet, exhibiting similarity scores of 0.86 and 0.87 respectively, are selected to ensure complementary diversity. Consequently, the final panel achieves both high individual performance and significant overall diversity.

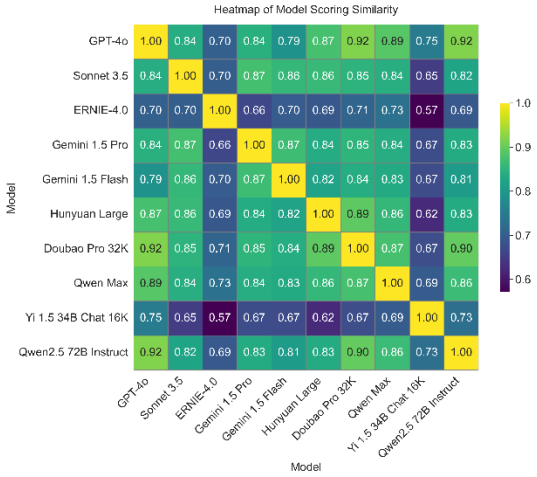

Fig. 2. Similarity Heatmap of Candidate LLMs' Scoring Results

3.4 Analysis of Scoring Results

Score Integration. The scoring process consists of two main stages: quality score calculation and final response score integration.

In the first stage, for each candidate model, its quality score is calculated by averaging the assessments from n high-performing models. For a candidate model's responses to j questions, each high-performing model first generates an average quality score across all questions, then these n scores are averaged to obtain the final quality score:

$$Q_c = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{j}\sum_{k=1}^{j} q_{i,k}\right) \tag{3}$$

where $\mathrm{Q_c}$ represents the final quality score for the candidate model, $\mathrm{q_{i,k}}$ denotes the quality score given by the i-th powerful model for the k-th question.

After selecting the scoring panel members based on quality scores, the final score for each response is calculated by averaging the scores from all panel members:

$$\text{Final Score} = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{4}$$

where $\mathrm{S_i}$ represented the score by the i-th panel member, and N is the total number of scoring panel.

This averaging approach in both stages helps ensure stability and reliability in the evaluation process, while mitigating potential biases from individual models.

Statistical Analysis. Three independent experiments were conducted to validate the stability and accuracy of the evaluation approach. The statistical analysis consisted of the following steps.

Friedman test to examine significant differences across the three experiments. Wilcoxon signed-rank test for pairwise comparisons between experiments. Stability analysis using Mean Absolute Error (MAE) between LLM scores and human expert judgements:

$$\text{Stability} = \sigma(\mathrm{MAE}_A, \mathrm{MAE}_B, \mathrm{MAE}_C) \tag{5}$$

where $\sigma(\cdot)$ represented the standard deviation of MAE values across three experiments, with lower values indicating higher stability.

4 Experiments and Results

The multi-model assessment framework was validated through comprehensive experiments and compared against a single-model assessment method. The results showed that the proposed method is more effective in terms of accuracy, stability, and relevance to human judgment. Moreover, ablation studies were conducted to further demonstrate the superiority of the method.

4.1 Experimental Results

Comparative Analysis. The initial experiments compared the performance of single-model scoring using GPT-4o against that of the proposed multi-model scoring approach. The single-model evaluation, while showing reasonable performance with a Pearson correlation coefficient of 0.75 and Spearman's rank correlation coefficient of 0.74, exhibited notable limitations. In particular, significant fluctuations in scoring consistency were observed when dealing with high-entropy questions. The single-model's scoring often deviated from human judgments to a greater extent.

In contrast, the multi-model scoring panel demonstrated superior performance across all evaluation metrics. This approach achieved Pearson's and Spearman's rank correlation coefficients of 0.78 and 0.78, respectively, representing a significant improvement over single-model evaluation. More importantly, the selected multi-model panel showed remarkable stability in handling high-entropy questions, maintaining better alignment with human scoring patterns even in challenging cases.

Extended Validation. To further validate the findings, the study conducted an extended experiment with an expanded test set [16, 17], increasing the number of evaluation questions from 100 to 250. This expansion encompassed a broader range of task scenarios and difficulty levels. The results from this extended evaluation were particularly encouraging. Compared to the single-model scoring method using GPT-4o, which achieved correlation coefficients of Pearson: 0.79 and Spearman: 0.78, the multi-model scoring panel selected using the proposed approach demonstrated superior performance, achieving higher correlation coefficients (Pearson: 0.85, Spearman: 0.84). Additionally, the selected multi-model panel exhibited improved stability across the broader test set, highlighting its robustness in diverse and challenging evaluation scenarios.

4.2 Ablation Studies

To validate the effectiveness of the approach, two key ablation studies were conducted. The experimental group consisted of Hunyuan Large, GPT-4o, and Claude 3.5 Sonnet, which were selected through the LLM cross-evaluation mechanism.

Impact of LLM Cross-evaluation Mechanism. The proposed scoring panel selection method was compared with a baseline approach that randomly selected three models.

As shown in Fig. **3**, the Pearson correlation comparison demonstrated that the proposed method achieved consistently higher correlation with human evaluation across various test cases. Similarly, Fig. **4** showed the Spearman correlation results, further confirming the superiority of the cross-evaluation based selection over random selection.

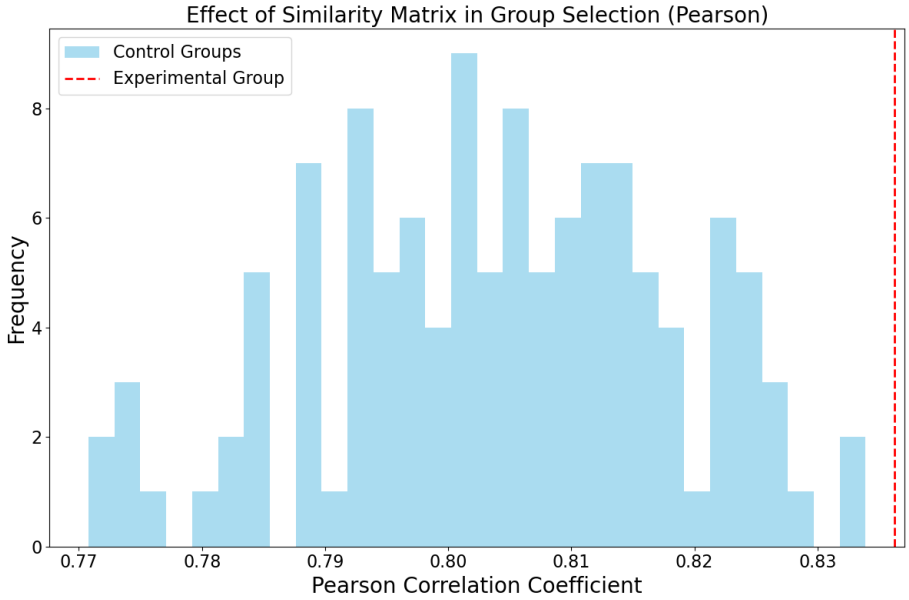

Fig. 3. Pearson Correlation Comparison in Model Selection Methods

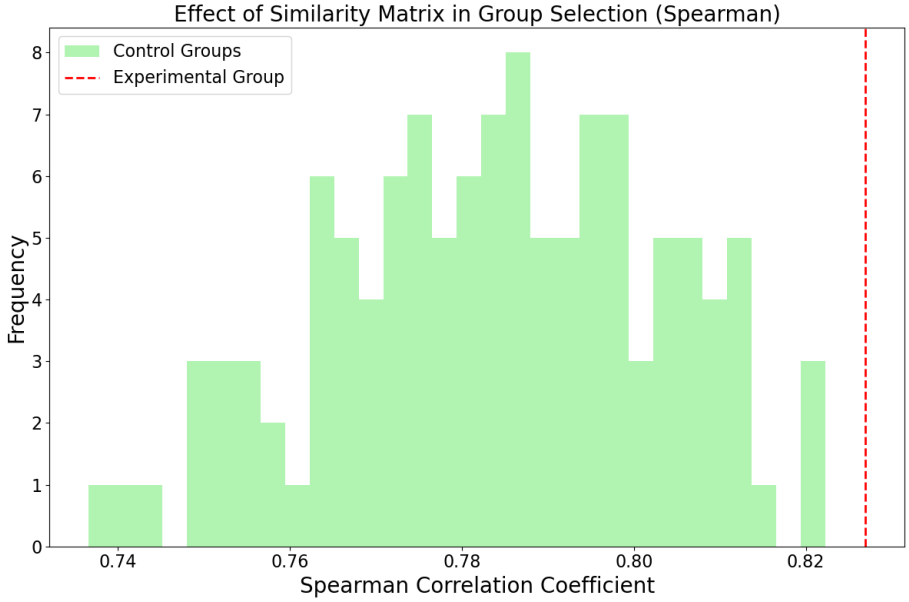

Fig. 4. Spearman Correlation Comparison in Model Selection Methods

Effect of Scoring Panel Size. The study also investigated how the number of LLMs in the scoring panel affected the LLM evaluation performance.

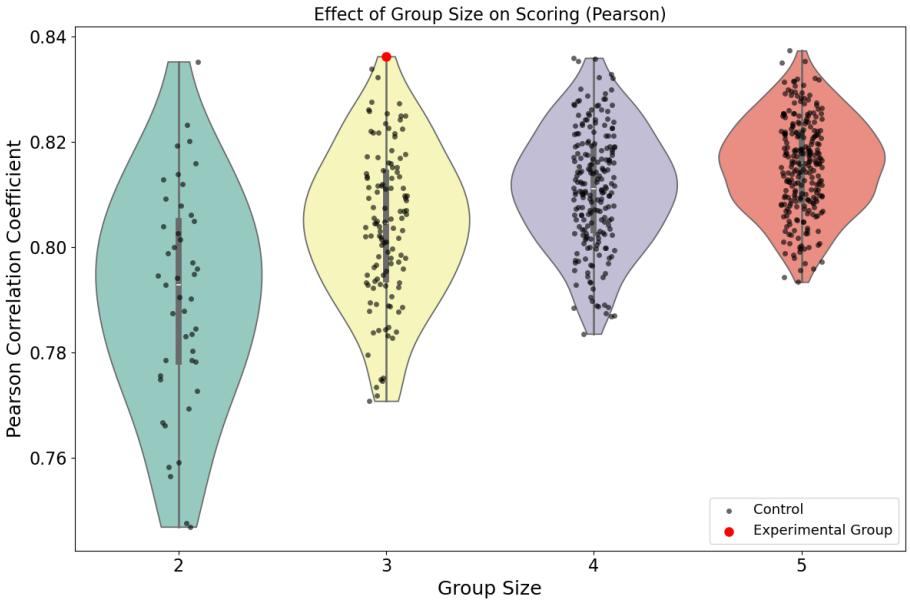

Fig. 5. Impact of Panel Size on Pearson Correlation

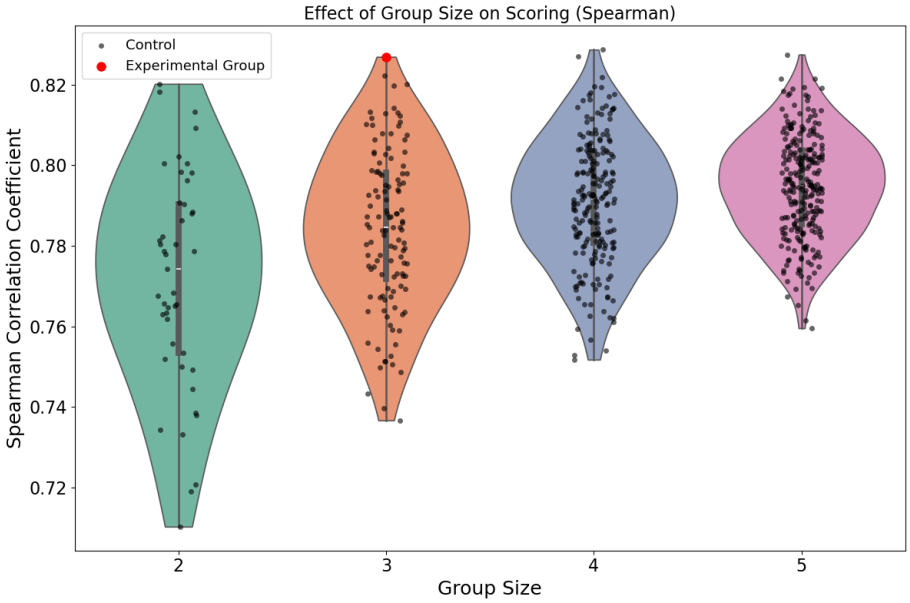

Fig. 6. Impact of Panel Size on Spearman Correlation

The impact on Pearson correlation is shown in Fig. **5**, while Fig. **6** presents the Spearman correlation results. Both metrics indicate that a panel size of three achieves the optimal balance between evaluation quality and computational efficiency, with larger panels showing decreasing benefits.

These ablation studies provided strong empirical evidence for both the effectiveness of the scoring panel selection strategy and the optimal balance of the scoring panel. The results consistently showed that the proposed method outperformed random selected multi-model panel and single-model evaluation, and that a carefully selected panel of three LLMs provided the best trade-off between performance and efficiency.

4.3 Significance and Stability Analysis

In this part, three experiments (A, B, C) were conducted based on the extended dataset, and a thorough analysis of the three experimental results was performed. The experiments consisted of three models: Hunyuan Large, GPT-4o, and Claude 3.5 Sonnet. The aim was to prove that the performance of the 3-member scoring panel was better than that of the single model (GPT-4o).

All evaluations were referenced to the expert-annotated data.

Significance Tests. Three independent trials (A, B, C) were performed, and the error (LLM score versus human score) was computed for each question. Table 2 presents the Friedman p-values across trials A/B/C and the Wilcoxon signed-rank p-values in pairwise comparisons:

Table 2. Friedman (p) and pairwise Wilcoxon signed-rank (p) values for each combo in trials A, B, C.

Combo	Friedman_p	Wilcoxon_AB	Wilcoxon_AC	Wilcoxon_BC
GPT-4o	0.653	0.808	0.857	0.690
Experimental Group	0.313	0.533	0.101	0.141

Overall, neither GPT-4o nor the 3-member LLM panel exhibited a strong statistical difference among trials (all $p > 0.05$). However, the selected LLM panel yielded slightly lower p-values in Wilcoxon_AC and Wilcoxon_BC, suggesting modest variation across trials that did not reach the significance thresholds.

Stability and Error Distribution. Table 3 summarized the mean absolute error (MAE) for trials A, B, and C, and its standard deviation (MAE_std). While GPT-4o and the selected LLM panel obtained relatively close MAEs, the selected LLM panel achieved a substantially lower MAE_std (0.016 vs. 0.041), indicating more consistent performance across repeated runs.

Fig. 7 illustrates the error distribution (model score minus reference) in the three trials, highlighting the selected panel's tighter clustering around zero. Such distributions corroborated the stability data, suggesting that the selected LLM panel not only

slightly improves error rates but also remains more robust across different experimental runs.

Table 3. Mean absolute errors (MAE) in trials A, B, C and their standard deviation(MAE_std).

Combo	MAE_A	MAE_B	MAE_C	MAE_std
GPT-4o	0.653	0.808	0.857	0.690
Experimental Group	0.313	0.533	0.101	0.141

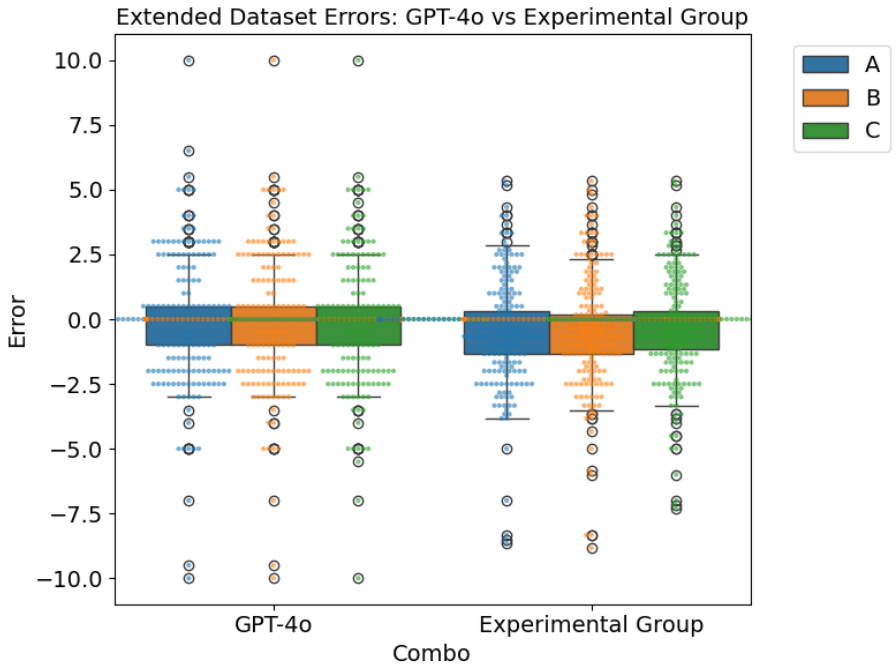

Fig. 7. Comparison of GPT-4o vs. Experimental Group on the extended dataset across trials A, B, and C. Experimental Group's errors exhibit a smaller spread around zero, reflecting more stable performance.

In conclusion, although the Friedman and Wilcoxon tests do not reveal strong statistical differences among the three trials for either approach, The selected LLM panel yields lower overall MAE and notably reduced variability (MAE_std) compared to GPT-4o, indicating that the selected multi-model scoring panel is superior in both accuracy and consistency.

4.4 Analysis and Discussion

The multi-model scoring approach outperforms others because the diverse LLM architectures and training backgrounds in the panel reduce individual biases. By carefully selecting LLMs with complementary strengths and low similarities, comprehensive and accurate evaluations are ensured. The cross-evaluation mechanism is vital for maintaining scoring quality, allowing continuous monitoring and optimization. Balancing quality and diversity through quality scores and similarity analysis is key to achieving robust results.

However, limitations exist. Using mainstream LLMs may miss some evaluation perspectives, and shared biases among top LLMs in cross-evaluation could reduce objectivity. Although the test set is expanded, it may not cover all scenarios, and human scoring's subjectivity can affect accuracy. These findings suggest future research should focus on including more diverse model types, improving cross-evaluation mechanisms, expanding datasets, and developing objective scoring criteria while exploring automation.

5 Discussion and Conclusion

This study introduces a novel method for selecting a multi-model panel for LLM evaluation [2, 18]. Extensive experiments show it significantly improves evaluation, with the panel outperforming single-model methods and aligning closely with human judgments. A 3-member panel proves particularly effective, balancing quality and efficiency, highlighting the importance of selection over panel size.

While the proposed method is effective, certain limitations warrant further investigation. The current reliance on mainstream LLMs may not capture all possible evaluation perspectives, and the computational demands of multi-model evaluation present ongoing challenges for scalability. These limitations point to several promising directions for future research, including the development of more efficient model selection strategies using reinforcement learning techniques, expansion of evaluation coverage to more diverse task types and domains, and exploration of automated generation of scoring standards. Additionally, the integration of advanced cross-validation mechanisms could further enhance evaluation reliability.

Looking ahead, the principles and methodologies developed in the study can be extended and refined to create even more comprehensive and reliable evaluation frameworks. As large language models continue to evolve in complexity and capability, the importance of robust and multi-faceted evaluation approaches will only grow.

References

1. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models (2024), https://arxiv.org/abs/2307.06435

2. Liang, P., Bommasani, R., Lee, T., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
4. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
5. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023), https://arxiv.org/abs/2306.05685
6. Hackl, V., Müller, A.E., Granitzer, M., Sailer, M.: Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. Frontiers in Education 8 (Dec 2023). https://doi.org/10.3389/feduc.2023.1272229, http://dx.doi.org/10.3389/feduc.2023.1272229
7. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A survey on evaluation of large language models (2023), https://arxiv.org/abs/2307.03109
8. Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., Lewis, P.: Replacing judges with juries: Evaluating llm generations with a panel of diverse models (2024), https://arxiv.org/abs/2404.18796
9. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms as-judges: A comprehensive survey on llm-based evaluation methods (2024), https://arxiv.org/abs/2412.05579
10. Laskar, M.T.R., Alqahtani, S., Bari, M.S., Rahman, M., Khan, M.A.M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C.W., Parvez, M.R., Hoque, E., Joty, S., Huang, J.: A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 13785–13816. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp main.764, https://aclanthology.org/2024.emnlp-main.764/
11. Bohnet, B., Tran, V.Q., Verga, P., Aharoni, R., Andor, D., Soares, L.B., Ciaramita, M., Eisenstein, J., Ganchev, K., Herzig, J., Hui, K., Kwiatkowski, T., Ma, J., Ni, J., Saralegui, L.S., Schuster, T., Cohen, W.W., Collins, M., Das, D., Metzler, D., Petrov, S., Webster, K.: Attributed question answering: Evaluation and modeling for attributed large language models (2023), https://arxiv.org/abs/2212.08037
12. Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.C., Lai, Y., Tao, C., Ma, S.: Leveraging large language models for NLG evaluation: Advances and challenges. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 16028–16045. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.896, https://aclanthology.org/2024.emnlp-main.896/
13. Panickssery, A., Bowman, S.R., Feng, S.: Llm evaluators recognize and favor their own generations (2024), https://arxiv.org/abs/2404.13076

14. Bein, B.: Entropy. Best Practice Research Clinical Anaesthesiology 20(1), 101–109 (2006). https://doi.org/https://doi.org/10.1016/j.bpa.2005.07.009, https://www.sciencedirect.com/science/article/pii/S1521689605000558, monitoring Consciousness
15. Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., Lan, Z.: CLUE: A Chinese language understanding evaluation benchmark. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 4762–4772. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling main.419, https://aclanthology.org/2020.coling-main.419/
16. Yu, L., Jiang, W., Shi, H., et al.: Metamath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:2309.12284 (2023)
17. Zhong, W., Cui, R., Guo, Y., et al.: Agieval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364 (2023)
18. Sun, K., Wang, R., Søgaard, A.: Comprehensive reassessment of large-scale evaluation outcomes in llms: A multifaceted statistical approach. arXiv preprint arXiv:2403.15250 (2024)