GRP: Goal-Reversed Prompting for Zero-Shot Evaluation with LLMs

Anonymous ACL submission

Abstract

Using Large Language Models (LLMs) to evaluate and compare two answers from different models typically involves having LLM-based judges select the better answer. However, humans often approach problem-solving from a reverse perspective, for instance, by choosing the worse option instead of the better one in a pairwise comparison. Generally, this kind of reverse thinking plays a crucial role in human reasoning and decision-making and can further test the difference between original and reverse thought processes simultaneously. To address the above issue, in this paper, we propose a Goal-Reversed Prompting (GRP) approach for pairwise evaluation that shifts the original task from selecting the better answer to choosing the worse one. We encourage LLMs to think in reverse by prompting LLMs to identify the worse response. Experiments on closed-source models demonstrate that GRP significantly enhances evaluation capabilities, outperforming the prompt template with the original goal.

1 Introduction

Large Language Models (LLMs) have been achieving extraordinary feats in recent years and evolving astonishingly. With the debut of more advanced AI models each month, a pivotal challenge surfaces: *How can we effectively evaluate, compare, and oversee these models?* Human judgments have long been the gold standard for assessing and supervising language models. However, the process of collecting these judgments is often expensive and time-consuming. To address the above limitation, using LLM-based judges has emerged as an effective paradigm (Zheng et al., 2023; Tan et al., 2024), increasingly adopted for evaluating and ranking models. These AI judges have become integral to enhancing model capabilities. This shift not only streamlines evaluation but also accelerates the development of more sophisticated language models.

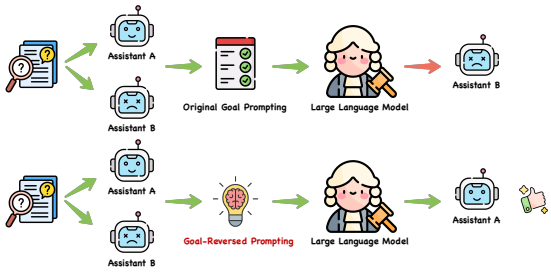

Figure 1: The pipeline of pairwise comparison with different goals (the second row illustrates the reverse).

In the prior studies (Zheng et al., 2023; Wang et al., 2023), there are mainly two types: one involves scoring a single answer and comparing pairs of answers. The former requires directly scoring one answer to a given question, with the score possibly ranging from 1 to 10. The latter usually has a broader range of applications; it requires evaluating two answers (from two different models) to a given question and determining which is better. The pipeline of the latter one is shown in Figure 1.

Prior work (Tan et al., 2024) has carefully analyzed the question: "*What makes one answer objectively better than another?*" They believe that LLM judges should follow three guiding principles when selecting the better answer: (1) The answer must faithfully follow human instructions; (2) It should provide a factually and logically correct response; (3) Its style should align with human preferences. Therefore, a strong LLM judge must first determine whether the answer follows the instructions, evaluate its factual and logical accuracy, and consider whether its style aligns with human preferences. For example, suppose the question is, "What is the capital of Spain?" The answer "1+1=2" is always factually correct, but it should not be favoured over the answer "Barcelona," which attempts to answer the question but is incorrect. Once principle (1) is satisfied (both answers follow the instructions), the correct answer should be preferred over the incor-

Directly Prompting
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is **worse** and must output only one of the following choices as your final verdict with a label: 1. Assistant A is **worse**: [[B>A]] 2. Assistant B is **worse**: [[A>B]] Example output: "My final verdict is Assistant A is **worse**: [[B>A]]".
CoT Prompting
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is **worse** and must output only one of the following choices as your final verdict with a label: 1. Assistant A is **worse**: [[B>A]] 2. Assistant B is **worse**: [[A>B]] Example output: "My final verdict is Assistant A is **worse**: [[B>A]]". Let's think step by step.
Prompting via SOPs
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is **worse**. Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers. When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information. Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive. Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt. After providing your explanation, you must output only one of the following choices as your final verdict with a label: 1. Assistant A is **worse**: [[B>A]] 2. Assistant B is **worse**: [[A>B]] Example output: "My final verdict is Assistant A is **worse**: [[B>A]]".
<|User Prompt|> {User-Prompt} <|The Start of Assistant A's Answer|> {Answer-A} <|The End of Assistant A's Answer|> <|The Start of Assistant B's Answer|> {Answer-B} <|The End of Assistant B's Answer|>

Table 1: Prompt template of Arena-Hard Judge with Reverse Goal are used in this paper.

rect one. Only when both (1) and (2) are met do style differences affect the judgment. Based on the above, they proposed the JudgeBench evaluation set to test a model's ability to evaluate which of two answers is better.

Reverse thinking plays a crucial role in human reasoning process (Branchini et al., 2021; Chen et al., 2024b). For example, an effective way to improve scores in a math test is to reason both forward and backward. In forward reasoning, we begin with the question and work step by step toward an answer. Conversely, reverse thinking starts from the predicted answer and works backward to the original question. This two-way approach allows us to verify the accuracy of the solution and identify potential errors . Inspired by this method, we propose GRP, a goal-reversed prompting method in pairwise evaluation, to shift the ultimate goal of a standard operating procedures based prompt template from choosing the better answer to selecting the worse one, as shown in Figure 1. By forcing LLMs to pick the worse answer, we prompt LLMs to think and solve problems differently from the original goal. Experiments on several closed-source models have demonstrated that the proposed GRP can effectively enhance the models' evaluation capabilities, significantly outperforming the prompt template with the original goal.

2 Methodology

In this paper, we introduce a straightforward and effective method for evaluation scenarios involving large language models: we transform the task of selecting which answer is better into identifying which answer is worse. This inversion simplifies the evaluation process and can lead to more robust and reliable assessments by focusing on eliminating inferior responses.

Furthermore, considering that prompt templates based on the Chain-of-Thought (CoT) paradigm

LLMs	*Version*	*Knowledge*	*Reasoning*	*Math*	*Coding*	***Overall***
SOPs Prompting						
GPT-4o-mini†	*2024-07-18*	46.75	42.86	66.07	42.86	48.29
GPT-4o†	*2024-05-13*	46.10	48.98	62.50	52.38	50.29
GPT-4o	*2024-11-20*	61.04	56.12	73.21	61.90	61.71
Gemini-1.5-flash†	*001*	26.62	20.41	30.36	9.52	23.43
Gemini-1.5-pro†	*001*	45.45	37.76	55.36	9.52	40.57
Gemini-1.5-pro	*002*	57.79	52.04	71.43	47.62	57.14
Claude-3-Haiku†	*2024-03-07*	20.78	14.29	16.07	2.38	16.00
Claude-3.5-Sonnet†	*2024-06-20*	59.74	62.24	60.71	54.76	60.00
SOPs Prompting via Reverse Goal						
GPT-4o	*2024-11-20*	**68.18**	61.22	71.43	**63.81**	**66.23**
Gemini-1.5-Pro	*002*	56.49	59.18	**76.79**	47.62	59.43
Claude-3.5-Sonnet	*2024-06-20*	62.99	**66.32**	**76.79**	61.90	66.00

Table 2: Evaluation results on JudgeBench with **GPT-4o** generated pairs. Among them, † indicate results directly re-evaluated from those in Tan et al. (2024). In the experimental section of this study, we adopted a more stringent strategy by excluding the scenario where A equals B.

generally perform less effectively than those based on Standard Operating Procedures (SOPs) for tasks of this nature, we have modified the existing evaluation prompt templates by reversing the final objective. Thus, this modification results in a new prompt—the inverted objective prompt—which we present in detail in Table 1.

Our approach streamlines the evaluation process and enhances the performance of large models by leveraging the strengths of SOP-based prompts. We can more effectively discriminate between potential answers by identifying weaker responses, thereby improving the overall evaluation mechanism. This method offers a simple yet robust alternative for enhancing evaluation strategies in the context of large language models.

3 Experimental Setup

In this paper, we utilize the JudgeBench (Tan et al., 2024) dataset as our test set, which consists of a total of 350 questions: 154 in Knowledge, 98 in Reasoning, 56 in Mathematics, and 42 in Coding. Additionally, we employ another version of the dataset where responses are generated using Claude-3.5-Sonnet instead of GPT-4o. This alternative version yields 270 question-response pairs: 154 in Knowledge, 51 in Reasoning, 34 in Mathematics, and 31 in Coding.

LLM-based judges are known to exhibit positional bias (Zheng et al., 2023; Wang et al., 2023), where the order in which response pairs are presented can influence their decisions. Evaluating the judge using a single order introduces this bias into the evaluation process. To mitigate positional bias, we assess the LLM-based judge twice, swapping the order of the response pairs in the second trial.

LLMs	*Version*	***Overall***
SOPs Prompting		
GPT-4o	*2024-11-20*	52.22
Gemini-1.5-Pro	*002*	52.57
Claude-3.5-Sonnet†	*2024-06-20*	42.22
SOPs Prompting via Reverse Goal		
GPT-4o	*2024-11-20*	52.96
Gemini-1.5-Pro	*002*	58.57
Claude-3.5-Sonnet	*2024-06-20*	44.07

Table 3: Evaluation results on JudgeBench with **Claude-3.5-Sonnet** generated pairs.

Different from the prior work (Tan et al., 2024), we strictly consider only the cases where the judge consistently and correctly identifies $A > B$ or $B > A$ in both trials. If the judge provides inconsistent decisions (e.g., $A > B$ in one trial and $B > A$ in the other) or includes tie judgments ($A = B$), we deem the decision incorrect, as it indicates the judge is either guessing or unable to distinguish between the responses reliably. This approach ensures a more accurate measurement of the judges' ability by focusing solely on clear and consistent judgments of superiority.

We also consider the prompt template used in Arena-Hard Judge (Tan et al., 2024) as the SOPs

for evaluation, which prompts the LLM to first generate its own reference answer. It then analyzes both the reference answer and the initial response before delivering a final verdict. To ensure reproducibility, we implemented a significant change across all judges by employing greedy decoding (temperature = 0).

4 Results and Analysis

4.1 Main Results

As shown in Table 2, it can be observed that the goal-reversed prompting template leads to performance improvements in each LLM, demonstrating the effectiveness of GRP. Compared to GPT-4O and GEMINI-1.5-PRO, GPT-4O achieve the best results, with a 4.52% improvement over the forward objective. This phenomenon suggests that models with stronger reasoning abilities are better at thinking from different perspectives and, therefore, can achieve better results when using reverse objectives. At the same time, CLAUDE-3.5-SONNET achieves the highest improvement, with an overall increase of 6%. Furthermore, we find that when adopting inverse thinking, the evaluation accuracy on the math task in JudgeBench significantly improved.

4.2 Additional Analysis

To empirically investigate the evaluation capabilities, we conduct an ablation study using Claude-3.5-Sonnet instead to generate response pairs. As illustrated in Table 3, compared to the prompt template with the original goal, each LLM shows improvement under the prompts of the reverse goal. Furthermore, combined with Table, we find that when LLMs evaluate their own results, the bias phenomenon is more pronounced, leading to a deterioration in evaluation outcomes.

Based on Table 4, we observe that the evaluation performance of the models can be improved even by directly using the reverse goal without complex standard operating procedures in the prompt template. This further demonstrates the effectiveness of goal-reversed prompting in zero-shot LLM evaluation.

5 Related Work

The emergence of advanced language models has fundamentally transformed human-AI interaction paradigms through their unprecedented text generation capacities (Zhao et al., 2023; OpenAI, 2023).

LLMs	*Version*	***Overall***
Directly Prompting		
GPT-4O	*2024-11-20*	47.14
GEMINI-1.5-PRO	*002*	47.78
CLAUDE-3.5-SONNET	*2024-06-20*	48.57
Directly Prompting via Reverse Goal		
GPT-4O	*2024-11-20*	51.43
GEMINI-1.5-PRO	*002*	52.57
CLAUDE-3.5-SONNET	*2024-06-20*	53.14

Table 4: Evaluation results on JudgeBench with **GPT-4o** generated pairs with different prompts.

As these artificial intelligence systems achieve increasingly sophisticated performance, rigorous assessment of their instruction-following competence becomes crucial for responsible deployment. Recent methodological innovations have witnessed a paradoxical trend where researchers employ the subject models themselves as evaluation instruments (Koo et al., 2023; Liusie et al., 2023; Liu et al., 2023; Zhu et al., 2023; Lu et al., 2023; Fu et al., 2023; Zheng et al., 2023; Wang et al., 2023; Chen et al., 2024a; Song et al., 2025), creating a self-referential assessment framework that has sparked considerable academic debate. While initial implementations of this auto-evaluation approach demonstrated practical utility Zheng et al. (2023), subsequent critical analyses have exposed inherent limitations in reliability (Wang et al., 2023; Chen et al., 2024a). Empirical investigations reveal that language model-based assessments frequently manifest confounding variables, including positional bias and semantic preference distortions (Chen et al., 2024a), introducing non-deterministic fluctuations in evaluation outcomes. This work proposes a goal-reversed prompting template from a new perspective, which effectively improves the evaluation accuracy of large language models.

6 Conclusion

We propose a goal-reversed prompting method for pairwise evaluation that shifts the original task from selecting the better answer to choosing the worse one. By prompting LLMs to identify the inferior response, we encourage alternative problem-solving approaches. Experiments on closed-source models demonstrate that GRP significantly enhances evaluation capabilities, outperforming the prompt template with the original goal.

7 Limitations

Due to limited resources, this paper tested closed-source models only by making API calls. A more important research focus in future work will be exploring the evaluation effects of other open-source models under goal-reversed prompting.

References

Erika Branchini, Elena Capitani, Roberto Burro, Ugo Savardi, and Ivana Bianchi. 2021. Opposites in reasoning processes: Do we use them more than we think, but less than we could? *Frontiers in Psychology*, 12.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? A study on judgement biases. *CoRR*, abs/2402.10669.

Justin Chih-Yao Chen, Zifeng Wang, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long T. Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee, and Tomas Pfister. 2024b. Reverse thinking makes llms stronger reasoners. *CoRR*, abs/2411.19865.

Xue-Yong Fu, Md. Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are large language models reliable judges? A study on the factuality evaluation capabilities of llms. *CoRR*, abs/2311.00681.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *CoRR*, abs/2309.17012.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023. Llms as narcissistic evaluators: When ego inflates evaluation scores. *CoRR*, abs/2311.09766.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2023. Zero-shot NLG evaluation through pairware comparisons with llms. *CoRR*, abs/2307.07889.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025. Can many-shot in-context learning help LLMs as evaluators? a preliminary empirical study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8232–8241, Abu Dhabi, UAE. Association for Computational Linguistics.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *CoRR*, abs/2410.12784.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

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

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *CoRR*, abs/2310.17631.