GLaPE: Gold Label-agnostic Prompt Evaluation for Large Language Models

Anonymous ACL submission

Abstract

Despite the rapid progress of large language models (LLMs), their task performance remains sensitive to prompt design. Recent studies have explored leveraging the LLM itself as an optimizer to identify optimal prompts that maximize task accuracy. However, when evaluating prompts, such approaches heavily rely on elusive manually annotated gold labels to calculate task accuracy for each candidate prompt, which hinders its generality. To overcome the limitation, this work proposes GLaPE, a gold label-agnostic prompt evaluation method to alleviate dependence on gold labels. GLaPE is composed of two critical aspects: self-consistency evaluation of a single prompt and mutual-consistency refinement across multiple prompts. Experimental results on 8 widely-recognized reasoning tasks demonstrate that GLaPE can produce more effective prompts, achieving performance comparable to those derived from manually annotated gold labels. Analysis shows that GLaPE provides reliable evaluations aligned with accuracy, even in the absence of gold labels. Code is publicly available at `Anonymous`.

1 Introduction

As the integration of large language models (LLMs) into natural language processing tasks has become imperative in recent years (Achiam et al., 2023; Scao et al., 2023; Chowdhery et al., 2022; Touvron et al., 2023), the sensitivity of the performance of LLMs to prompts has garnered significant attention (Pezeshkpour and Hruschka, 2023; Loya et al., 2023). While traditional soft prompt tuning methods (Li and Liang, 2021; Liu et al., 2022; Lester et al., 2021; Qin and Eisner, 2021) demonstrate effectiveness in guiding the LLM to perform desired tasks, they encounter limitations when applied to private LLMs, such as GPT-4 (OpenAI, 2023). This situation necessitates the exploration of effective strategies for optimizing prompts without requiring gradient updates.

Recent studies (Yang et al., 2023; Zhou et al., 2022) have unveiled a noteworthy strategy, where the LLM itself acts as the optimizer to seek the prompt that maximizes task accuracy. Specifically, OPRO (Yang et al., 2023) provides an intriguing avenue for prompt optimization based on a gold label evaluation recipe (Figure 1a). The optimization commences with an initial prompt, then iteratively evaluates existing prompts and generates novel prompts based on prior assessments. However, a significant caveat emerges as these studies heavily rely on manually annotated gold labels. Concretely, the gold label, representing the ideal output, serves as a crucial ingredient for evaluating and refining prompts. Nevertheless, the acquisition of such gold labels poses a formidable obstacle (Huang et al., 2023; Stechly et al., 2023), introducing complexity and hindering the widespread implementation and generality of these optimization techniques. Therefore, exploring alternative methodologies becomes mandatory to address these challenges and improve the efficiency of prompt evaluation and optimization for LLMs.

To address the limitations, this work proposes a gold label-agnostic prompt evaluation (GLaPE) method to identify prompts that facilitate consistent and accurate answers. Instead of relying on gold labels, GLaPE evaluates prompts based on two critical aspects: self-consistency evaluation and mutual-consistency refinement. Inspired by Wang et al. (2022), we first consider a naive solution by utilizing self-consistency (SC) as the evaluation metric instead of accuracy, as correct answers generally exhibit higher SC than incorrect ones. However, we will show that SC alone may not always yield accurate evaluations, since SC does not always align well with accuracy and can overestimate prompts that produce incorrect but

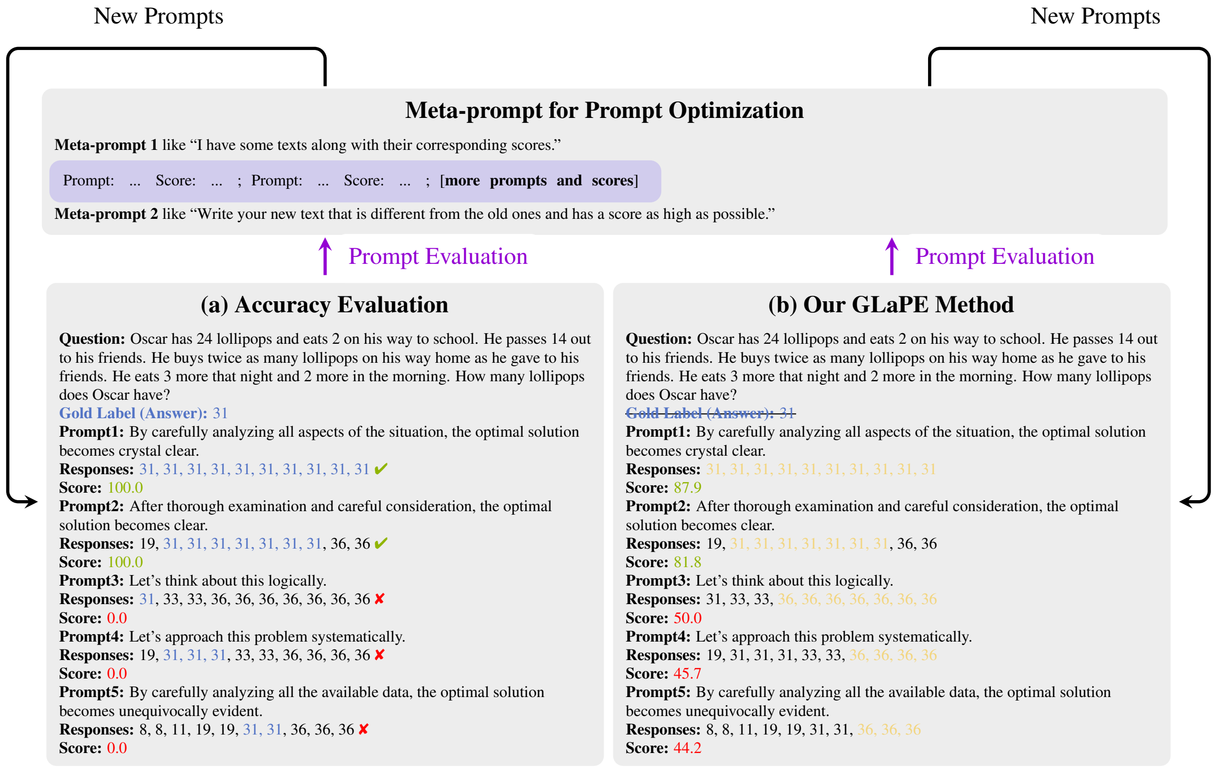


Figure 1: Sketch of prompt optimization utilizing the LLM as an optimizer (Yang et al., 2023), featuring distinct prompt evaluation metrics based on: (a) accuracy or (b) our proposed GLaPE. The texts are favorably read in colors. Blue: gold label, Yellow: most frequent answer, Green: high score, Red: low score, Purple: prompt evaluation.

consistent answers. To mitigate this, we then propose a complementary approach named mutual-consistency refinement across multiple prompts. This approach penalizes inconsistent scores based on SC across prompts that produce the same answers. By doing so, the refinement process effectively identifies prompts that demonstrate high SC but result in incorrect answers, leading to more reliable evaluation scores. Figure 2 illustrates our GLaPE method.

Building on our GLaPE evaluation strategy, we then develop a gold label-agnostic prompt optimization method. Specifically, we substitute the accuracy evaluation method in OPRO with our GLaPE method (Figure 1b). Experimental results on 8 widely-recognized reasoning tasks demonstrate that GLaPE can produce more effective prompts, achieving performance comparable to those derived from manually annotated gold labels.

Our key contributions are as follows:

(i) This work studies a gold label-agnostic prompt evaluation method to alleviate dependence on gold labels, which allows prompt evaluation in more realistic scenarios when human-annotated dataset is unavailable. To the best of our knowledge, this work is the first to study gold label-agnostic prompt evaluation for LLMs.

(ii) We propose a novel prompt evaluation approach named GLaPE, which consists of self-consistency evaluation of a single prompt and mutual-consistency refinement across multiple prompts. GLaPE helps LLMs optimize effective prompts that are comparable with those derived from manually annotated gold labels.

(iii) We elicit the analysis of why the widely-used SC approach fails at our evaluation task and figure out an effective mutual-consistency refinement approach to mitigate the challenge.

2 Related Work

Prompt Optimization In the domain of LLMs (Achiam et al., 2023; Scao et al., 2023; Chowdhery et al., 2022; Touvron et al., 2023), prompt engineering plays a crucial role in guiding models to generate desired outputs across diverse tasks (Pezeshkpour and Hruschka, 2023; Loya et al., 2023). Consequently, optimizing prompts becomes paramount for enhancing the performance and efficiency of LLMs. Various soft prompt tuning methods (Li and Liang, 2021; Liu et al., 2022; Lester et al., 2021; Qin and Eisner, 2021) have been explored in previous research to optimize prompts for open-source LLMs. However, these methods encounter challenges when applied to private LLMs, where accessing gradients is infeasible. Consequently, diverse gradient-free prompt optimization techniques (Zhou et al., 2022; Pan et al., 2023; Ye et al., 2023) have been explored. Recent works (Yang et al., 2023) have embraced an iterative process for gradient-free prompt optimization, commencing from an initial prompt and iteratively assessing existing prompts while generating new ones based on prior evaluations. Nevertheless, these iterative prompt optimization methods heavily depend on gold labels for prompt evaluation. In our work, we propose a novel gold label-agnostic prompt evaluation method and subsequently present a unique approach to optimize prompts for LLMs without the constraints associated with conventional gold label reliance.

Prompt Selection Prompt selection tasks aim to identify the optimal prompt among candidates for a given task, representing an alternative approach to prompt optimization. Recent studies have delved into probability-based evaluation methods, utilizing diverse metrics such as mutual information (Sorensen et al., 2022), entropy (Lu et al., 2021), and perplexity (Gonen et al., 2022). In contrast to these probability-centric assessments, our proposed evaluation approach exclusively relies on the output, making it applicable to private LLMs where only the output is accessible.

3 Background

3.1 Task Formulation

Existing studies on prompt design (Yang et al., 2023; Zhou et al., 2022) generally adhere to a two-stage paradigm in an iterative manner: (i) evaluate the prompt, analogous to calculating the loss function and gradient in soft prompt tuning; (ii) optimize the prompt, analogous to the gradient descent process in soft prompt tuning.

We formulate the two stages on top of the widely-used question-answering (QA) task defined by QA pairs (Q, A), where each pair comprises an input Q and its corresponding expected output A. We introduce the prompted model as $\mathcal{M}$ and an evaluation function f. Our objective is to determine the optimal natural language instruction prompt.

To begin with, we define the *meta-prompt* as the input to for prompt optimization. As the upper block shown in Figure 1, a meta-prompt contains three parts. The first part is a problem description. The second part is an optimization trajectory, includes past solutions and their evaluation scores. The third part is the optimization instruction for generating new candidate prompts.

Then, we describe the process of obtaining the optimization trajectory. In each iteration, the LLM generates a candidate prompt ρ to the QA task. We concatenate each question Q with the candidate prompt ρ to form the prompted input $[Q; \rho]$. Then, the prompted input is feed to the model to obtain the response $\mathcal{M}([Q; \rho])$. We evaluate the goodness of candidate prompt ρ based on the evaluation function f, e.g., calculating the accuracy between each pair of $\mathcal{M}([Q; \rho])$ and the labeled answer A in previous studies. Then the candidate prompt along with the evaluation score is added to the trajectory for the next iteration.

The optimization process terminates when the LLM is unable to propose new prompts with better evaluation scores, or a maximum number of optimization steps has reached.

3.2 Self-consistency

Here, we adopt the definition of self-consistency proposed by Wang et al. (2022). We sample n responses $(r_1, \cdots, r_n)$ from the LLM using the same prompt. The final answer is determined by a voting mechanism, where the most frequent response a is selected as the answer. Self-consistency is the frequency of a in all n responses, which can be formulated as:

$$\mathrm{SC} = \frac{\sum_{i=1}^{n} \mathbb{1}_{a=r_i}}{n}. \tag{1}$$

4 Investigating Gold Label-agnostic Prompt Evaluation

According to Section 3.1, the evaluation function f in existing studies measures the goodness of the prompt candidate ρ by maximizing the task accuracy. However, in real-world tasks, obtaining gold labels poses a considerable challenge, limiting the generalization of existing prompt optimization methods. Furthermore, we ultimately expect LLMs to solve problems for which answers are not already known. Therefore, when optimizing prompts to enhance performance, gold labels are not readily available. Thus, it is imperative to find a gold label-agnostic prompt evaluation method.

In this section, we will investigate the challenge of gold label-agnostic prompt evaluation and study

	AddSub	AQuA	Big-Bench Date	GSM8K	MultiArith	SVAMP	StrategyQA	MATH
Correct Answers (%)	96.0	79.0	83.4	82.1	97.5	90.1	95.4	70.2
Incorrect Answers (%)	73.4	67.1	67.8	49.3	54.2	57.5	90.6	35.9
Overall Answers (%)	92.8	71.8	79.2	73.8	96.6	84.9	91.9	44.6

Table 1: The average self-consistency of correct, incorrect, and overall answers generated by the LLM that prompted with "Let's think step by step." on multiple datasets.

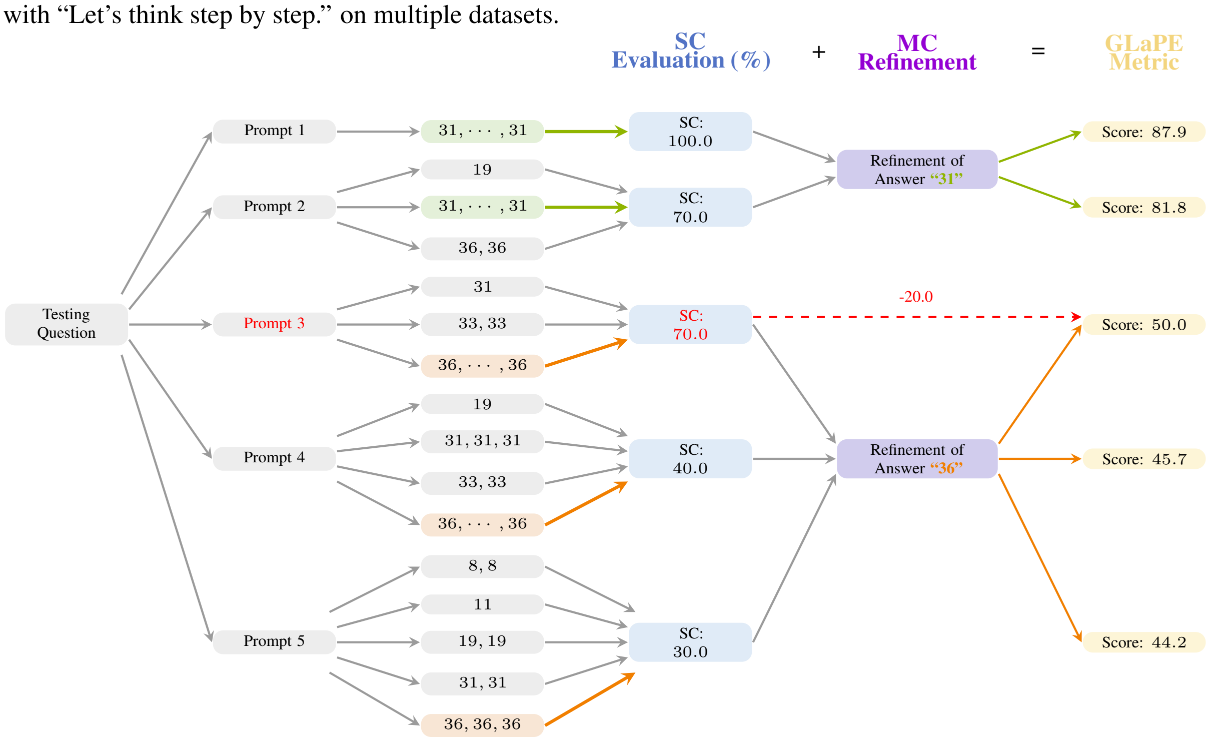


Figure 2: The schematic representation of our GLaPE method integrating self-consistency (SC) evaluation and mutual-consistency (MC) refinement. This sketch illustrates how our method assesses the prompts in Figure 1; computation details are provided in Appendix A.2. Notably, we observed that prompt3, as indicated by the red marker, produces an incorrect answer with high self-consistency (70%). Through the mutual-consistency refinement, our GLaPE score experiences a decrease of 20.0, rendering it more discernible when compared to prompt1 and prompt2. The texts are favorably read in colors of background. Blue: self-consistency, Purple: mutual-consistency refinement, Green: answer "31" (gold label), Orange: answer "36", Yellow: GLaPE metric.

how to design an effective approach to overcome the challenge.

4.1 SC Fails Due to Overestimating Prompts

For a gold label-agnostic prompt evaluation method, it is essential to rely exclusively on the responses and identify patterns within them. Building on the findings of Wang et al. (2022), which demonstrate that selecting the most frequently generated response enhances accuracy, we aim to investigate whether SC correlates with accuracy.

To this end, we experiment by utilizing the prompt "Let's think step by step." proposed by Kojima et al. (2022). We calculated the average self-consistency of correct, incorrect, and overall answers and presented the results in Table 1. We observe a significant superiority in the average self-consistency of correct answers compared to incorrect ones. A more specific example is shown in Figure 2. We see that the average SC of correct answers (answer "31") significantly surpasses that of incorrect ones. This observation indicates that the self-consistency of responses may reflect accuracy. Thus, it is possible to evaluate prompts based on the SC of the responses and incorporate this method in prompt optimization.

However, we also find that there exists disparity between SC and accuracy when using SC as the sole evaluation metric. This disparity happens to Prompt 3 as shown in Figure 2. Concretely, Prompt 3 yields an incorrect answer (answer "36") but has a high SC of 70.0. By taking the GSM8K dataset as the testbed, we computed both the self-consistency and accuracy for a group of prompts. Consequently, we draw each prompt as a point in Figure 3.

Evaluation Metric	AddSub	AQuA	Big-Bench Date	GSM8K	MultiArith	SVAMP	StrategyQA	MATH
GLaPE	0.44	0.04	0.88	0.49	0.88	0.69	0.18	0.67
SC evaluation	0.36	-0.13	0.75	0.40	0.29	0.31	0.14	0.33

Table 2: Spearman correlation coefficients (↑) between accuracy and SC / GLaPE across diverse datasets.

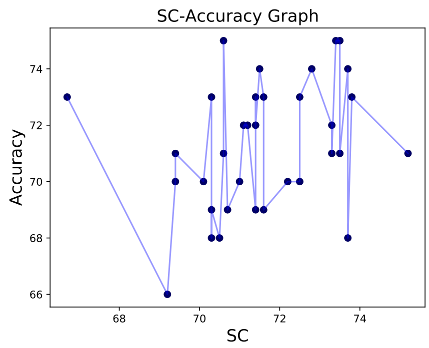


Figure 3: SC-Accuracy Graph for Prompts. Each prompt is represented as a point on the graph, where the x-coordinate signifies self-consistency and the y-coordinate signifies accuracy.

Given the observed fluctuations in the line, it is apparent that self-consistency does not align rigorously with accuracy. Additionally, we find that the Spearman correlation coefficient between SC and accuracy is relatively low, as shown in the first line of Table 2. Therefore, relying on self-consistency alone proves insufficient in offering a comprehensive representation of accuracy in prompt evaluation and optimization.

So far, we show that SC alone may not always yield accurate evaluations, since SC does not always align well with accuracy and can overestimate prompts that produce incorrect but consistent answers. Therefore, it deserves a more in-depth investigation to mitigate the side effects of the overestimated prompts by SC. Beyond examining individual prompt responses, we can analyze relationships between different prompts.

4.2 Mitigating the Challenge with Mutual-consistency (MC) Refinement

Although the performance of a single prompt is only related to its responses, we leverage other prompts for better evaluation in the absence of a gold label.

Specifically, we infer the gold label from other prompts and then refine the SC evaluation of the single prompt. Table 1 shows that correct answers exhibit higher self-consistency (SC), allowing us to predict answer correctness by analyzing the average SC of all prompts producing it. In Figure 2, we can predict that the answer "31" is more likely to be correct, while the answer "36" is not, as the average SC of "31" is 87.5, whereas that of "36" is 46.7. This prediction further aids in refining evaluation of each prompt. For an incorrect answer, we should lower the evaluation score of prompts with elevated SC, towards the average. In Figure 2, since the average SC of answer "36" is 46.7 while prompt 3 has an elevated SC of 70.0, the evaluation score of prompt 3 should be lowered. This refinement mitigate the SC evaluation of overestimated prompts.

In summary, we predict the correctness of an answer by its average SC and refine each SC towards this average. This aligns the evaluation of prompts producing the same answer.

Based on our pivot study above, we find that combining SC and MC is effective for achieving gold label-agnostic prompt evaluation.

5 GLaPE

In light of the discussions in Section 4, we propose GLaPE, a gold label-agnostic prompt evaluation approach. GLaPE is composed of two critical aspects: self-consistency evaluation of a single prompt and mutual-consistency refinement across multiple prompts. The overall procedure is illustrated in Figure as depicted in Figure 2.

For formal description purposes, we assume there are N different prompts and denote the evaluation score for each prompt ρ_i as f_i. Among multiple samplings of $\mathcal{M}$ prompted with $([Q;\rho_i])$, the answer is a_i and the self-consistency is c_i, as defined in Section 3.2.

Self-consistency Evaluation: We evaluate prompts based on the self-consistency of their answers by minimizing the loss function:

$$L_{\text{self}} = \sum_{i=1}^{N} (f_i - c_i)^2. \quad (2)$$

Mutual-consistency Refinement: Additionally, we propose L_{refine} as a corrective measure for SC evaluation. It measures and penalizes the mutual inconsistency of evaluation scores (f_i) for prompts sharing the same answer:

$$L_{\text{refine}} = \sum_{1 \le i < j \le N} \mathbb{1}_{a_i = a_j} (f_i - f_j)^2. \quad (3)$$

The overall loss function L_{total} is determined by balancing the loss functions of these two aspects:

$$L_{\text{total}} = \alpha \cdot L_{\text{self}} + (1 - \alpha) \cdot L_{\text{refine}}, \quad (4)$$

where α weights the contribution of self-consistency evaluation and mutual-consistency refinement in the evaluation process. Based on preliminary experiments (detailed in Appendix A.1), we set $\alpha = 0.5$.

We obtain the ultimate evaluations $f_1, \cdots, f_N$ by minimizing the loss function L_{total}. We initialize f_i with c_i for simplicity and utilize the default gradient descent method to find the optimal solution with a learning rate of 0.05.

6 Experiment

6.1 Experiment Setup

Datasets. Our experiments were conducted on 8 benchmark datasets to evaluate the performance of our gold label-agnostic prompt evaluation and optimization method. We selected five datasets specifically focused on arithmetic reasoning: AddSub (Hosseini et al., 2014), AQuA (Ling et al., 2017), GSM8K (Cobbe et al., 2021), MultiArith (Roy and Roth, 2015), and SVAMP (Patel et al., 2021). Additionally, we included the MATH dataset (Hendrycks et al., 2021), which is extremely challenging and comprehensive, to test our method's efficacy on particularly difficult benchmarks. Furthermore, we expanded our evaluation to commonsense reasoning benchmarks, such as Big-Bench Date (bench authors, 2023) and StrategyQA (Geva et al., 2021), to assess the performance of GLaPE in varied contexts.

Prompt Optimization. We implemented the OPRO method proposed by Yang et al. (2023). This technique utilizes an LLM to evaluate existing prompts, generating improved prompts based on the obtained evaluation scores. We chose this approach due to its adaptability; alternative metrics can easily replace evaluation scores in the meta-prompt of optimization. This flexibility facilitates the seamless execution of our gold label-agnostic prompt optimization experiments.

LLM Backbone. In both the evaluation and optimization phases, we employed gpt-3.5-turbo. For prompt evaluation, we empirically set the temperature to 0.7 and generated 10 outputs using chain-of-thought prompting (Wei et al., 2023). For prompt optimization, default hyperparameters and meta-prompt from Yang et al. (2023) were applied.

6.2 Main Results

Table 3 shows the main results on the 8 benchmark datasets. GLaPE is able to produce effective prompts, achieving performance comparable to those derived from manually annotated gold labels such as OPRO. The results suggests that our GLaPE can function as a robust metric, akin to accuracy. We also compared our method with other recent prompt optimization methods for private LLMs; these results are detailed in Appendix A.3, providing additional evidence to verify the generality of GLaPE.

6.3 Ablation Study

In this section, we conduct ablation studies to enhance our understanding of the GLaPE method, with a specific focus on the impact of the mutual-consistency refinement approach.

Initially, on the GSM8K dataset, we compared prompt optimization outcomes using two distinct evaluation methods: self-consistency assessment and GLaPE. As shown in Table 4, GLaPE-based prompt optimization results in a superior prompt compared to that obtained through confidence assessment. This observation suggests that incorporating mutual-consistency refinement to rectify confidence evaluation enhances the efficacy of prompt optimization.

Furthermore, we incorporated the Spearman correlation coefficient[1] into our study, wherein a higher coefficient signifies a stronger correlation between variables. This quantitative assessment was employed to juxtapose GLaPE with the solely SC-based evaluation regarding the correlation with accuracy. Our analysis concentrated on prompts within the optimization trajectory in the experiment in Section 6.2, to mitigate unnecessary computational costs. As delineated in Table 2, the Spearman coefficient between GLaPE and accuracy exceeds that of self-consistency across all datasets.

Additionally, we utilized the visualization method introduced in Section 4.1 to depict the prompts of the optimization trajectory in a graph (Figure 4). In Figure 4a, we observe a fluctuating line, whereas in Figure 4b, a consistently increasing line is evident. Both of the scrutiny indicate that our mutual-consistency refinement method significantly mitigates the disparity between self-consistency and accuracy.

[1] `https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient`

Dataset	Method	Prompt	Accuracy (%)
Addsub	Baseline (Wang et al., 2022)	Let's think step by step.	85.8
	OPRO (Yang et al., 2023)	Let's meticulously scrutinize every detail.	89.4
	GLaPE-based (Ours)	Let's carefully consider each step.	87.6
AQuA	Baseline (Wang et al., 2022)	Let's think step by step.	39.4
	OPRO (Yang et al., 2023)	After careful consideration and analysis, the optimal solution is revealed.	41.7
	GLaPE-based (Ours)	Through a meticulous analysis of all available data and a strategic approach to problem-solving, a definitive and optimal solution will undoubtedly arise.	43.7
Big-Bench Date	Baseline (Wang et al., 2022)	Let's think step by step.	72.4
	OPRO (Yang et al., 2023)	Using a systematic approach and thorough examination, the unequivocal and optimal solution becomes unmistakably evident.	72.1
	GLaPE-based (Ours)	Let's analyze this situation thoroughly and explore all possible solutions.	71.9
GSM8K	Baseline (Wang et al., 2022)	Let's think step by step.	74.8
	OPRO (Yang et al., 2023)	After careful analysis, the optimal solution becomes clear.	76.6
	GLaPE-based (Ours)	After careful analysis, the conclusion is evident.	77.7
MultiArith	Baseline (Wang et al., 2022)	Let's think step by step.	98.0
	OPRO (Yang et al., 2023)	Let's approach this problem systematically and strategically, step by step, with logical thinking and methodical planning.	99.6
	GLaPE-based (Ours)	Let's approach this problem strategically, methodically, and innovatively, exploring groundbreaking solutions.	99.3
SVAMP	Baseline (Wang et al., 2022)	Let's think step by step.	83.9
	OPRO (Yang et al., 2023)	Let's approach this problem with an innovative and revolutionary mindset, breaking the barriers of conventional thinking and achieving unprecedented results.	88.9
	GLaPE-based (Ours)	Let's approach this problem with an innovative, revolutionary, and groundbreaking solution.	88.7
StrategyQA	Baseline (Wang et al., 2022)	Let's think step by step.	66.1
	OPRO (Yang et al., 2023)	Let's tackle this problem with groundbreaking approaches and unparalleled creativity.	69.4
	GLaPE-based (Ours)	Let's explore all the possibilities.	70.2
MATH	Baseline (Wang et al., 2022)	Let's think step by step.	21.4
	OPRO (Yang et al., 2023)	Analyzing the data thoroughly can lead to valuable insights.	26.4
	GLaPE-based (Ours)	Let's approach this problem with an innovative, revolutionary, and groundbreaking solution.	25.9

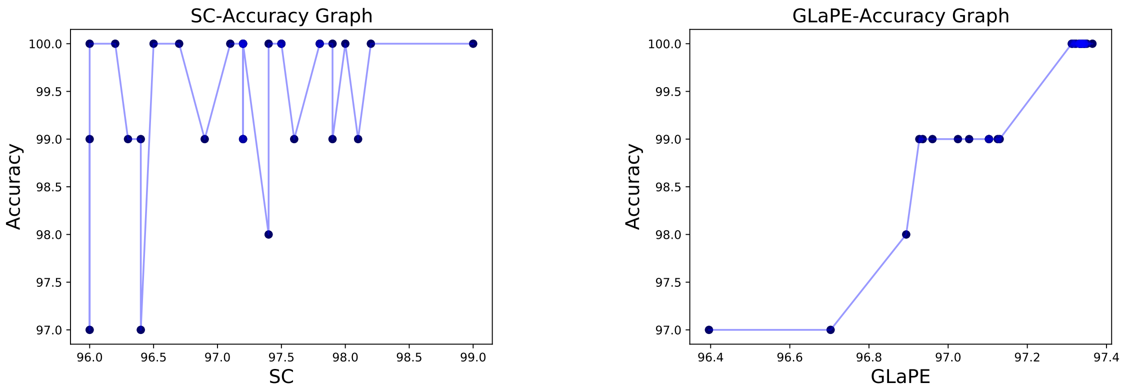
Table 3: Optimization results (optimal prompt and corresponding accuracy) of our GLaPE-based prompt optimization method and OPRO (Yang et al., 2023) across various datasets. Notably, Our optimal prompt is determined by selecting the prompt with the highest GLaPE score.

7 Rethink on Gold Label-agnostic Prompt Optimization

Our amalgamation of self-consistency evaluation and mutual-consistency refinement facilitates the identification of prompts leading to correct answers. However, we also observe a diminished Spearman correlation coefficient between our GLaPE and accuracy on the AQuA dataset and StrategyQA dataset, as depicted in Table 2. Given the suboptimal performance, we shift to reflect on the intrinsic restriction posed by the LLM. As stated in Section 4.1, in scenarios where all prompts result in consistent but inaccurate answers, our evaluation may fail to identify the error. Without access to external resources, discerning the consistent errors becomes challenging. We illustrate some example questions in the Strategy dataset in Figure 5, where

Evaluation Metric	Optimal prompt	Accuracy (%)
GLaPE	After careful analysis, the conclusion is evident.	77.7
SC evaluation	Let's break it down step by step.	75.1

Table 4: Comparison of prompt optimization based on self-consistency and our GLaPE.



(a) SC-Accuracy Graph.

(b) GLaPE-Accuracy Graph.

Figure 4: SC-Accuracy Graph v.s. GLaPE-Accuracy Graph for Prompts on Multiarith.

almost all prompts lead to a incorrect answer with extremely high self-consistency during our prompt optimization. By excluding these problematic questions, we observe a significant improvement in the Spearman correlation coefficient, as detailed in Appendix A.4.

Notably, Our GLaPE evaluation method has been shown effective in challenging datasets, too. Our GLaPE evaluation method aligns well with accuracy on the rigorous MATH dataset, as shown in Table 2. This alignment can be attributed to the significant disparity in self-consistency between correct and incorrect answers, as depicted in Table 1. Conversely, on the StrategyQA dataset, the self-consistency levels of correct and incorrect answers are remarkably similar. This similarity arises from the inherent errors within the LLM itself, which precludes effective distinction between these answers without access to the gold standard labels.

8 Conclusion

This work presents an innovative gold label-agnostic prompt evaluation methodology that operates in the absence of gold labels. By integrating self-consistency evaluation and mutual-consistency refinement, our evaluation demonstrates a strong correlation with accuracy. Subsequently, we incorporate our metric into prompt optimization, yielding prompts comparable to those optimized based on accuracy metrics across various tasks.

Example 1: Was Iggy Pop named after his father?
Label: Yes.
LLM Answer & Explanation: No. Iggy Pop, whose real name is James Newell Osterberg Jr., was not named after his father. His father's name was James Newell Osterberg Sr., so they shared the same first name but not the same nickname. Iggy Pop got his stage name from his first band, The Iguanas, and later shortened it to Iggy.
Example 2: Is Great Pyramid of Giza the last wonder of its kind?
Label: Yes.
LLM Answer & Explanation: No. The Great Pyramid of Giza is one of the Seven Wonders of the Ancient World. It is the only wonder from that list that still exists today. However, there have been other lists of wonders created throughout history. For example, the New7Wonders Foundation conducted a global poll in 2007 to determine the New Seven Wonders of the World.

Figure 5: Some examples in StrategyQA dataset where the LLM consistently provides inaccurate responses.

9 Limitations

First, in Section 7, we outlined the challenges faced by our GLaPE method in accurately assessing the inherent error of LLM itself. In future research, innovative approaches could be explored to identify the consistent mistakes. Another limitation in our current evaluation methodology is that we utilize a singular digital score as the assessment, which fails to furnish comprehensive information regarding the prompts. Consequently, future research could augment the granularity of prompt evaluations, incorporating other assessments, like natural language feedback, to address this shortfall.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

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

Rui Pan, Shuo Xing, Shizhe Diao, Xiang Liu, Kashun Shum, Jipeng Zhang, and Tong Zhang. 2023. Plum: Prompt learning using metaheuristic.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. 2023. Prompt engineering a prompt engineer.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Appendix

A.1 Preliminary Experiments

In this section, we discuss two crucial hyperparameters used in our experiments.

The first is the balance weight α, which balances SC evaluation and MC refinement as described in Equation 4. We tested α values of 0.25, 0.5, 0.75, and 1.0, with results detailed in Table 5. An optimal balance was achieved at $\alpha = 0.5$, emphasizing the significance of both SC and MC in our evaluation framework. Consequently, we set $\alpha = 0.5$ for all experiments.

Weight α	Prompt	Accuracy (%)
0.25	Let's think about this logically.	77.2
0.5	After careful analysis, the conclusion is evident.	77.7
0.75	Let's approach this problem with utmost creativity, innovation, and strategic thinking.	76.4
1.0	Let's break it down step by step.	75.1

Table 5: Optimization results on the GSM8K dataset using different values of balance weight α as specified in Equation 4.

The second parameter is the training dataset size. We evaluated various sizes: 10, 20, 50, 100, and 200, as shown in Table 6. Based on these results, we selected a dataset size of 100 to balance accuracy and computational efficiency.

Dataset Size	Prompt	Accuracy (%)
10	Let's break it down step by step.	75.1
20	Let's carefully analyze each aspect of the problem thoroughly and devise the most optimal plan.	75.5
50	Let's approach this problem with utmost creativity, innovation, and strategic thinking.	76.4
100	After careful analysis, the conclusion is evident.	77.7
200	Let's break it down step by step.	77.9

Table 6: Optimization results on the GSM8K dataset using different training dataset sizes.

Dataset	Method	Prompt	Accuracy (%)
GSM8K	Baseline (Wang et al., 2022)	Let's think step by step.	74.8
	APE (Zhou et al., 2022)	Let's work this out in a step by step way to be sure we have the right answer.	76.3
	APO (Pryzant et al., 2023)	Given the scenario, perform necessary calculations and provide a step-by-step explanation to arrive at the correct numerical answer. Consider all information provided.	76.5
	PE2 (Ye et al., 2023)	Let's solve the problem step-by-step and calculate the required total value correctly.	77.7
	GLaPE-based (Ours)	After careful analysis, the conclusion is evident.	77.7
MultiArith	Baseline (Wang et al., 2022)	Let's think step by step.	98.0
	APE (Zhou et al., 2022)	Let's work this out in a step by step way to be sure we have the right answer.	97.8
	APO (Pryzant et al., 2023)	Given the scenario, perform the necessary calculations step by step to find the final result. Consider all parts of the input and the sequence of events.	99.0
	PE2 (Ye et al., 2023)	Let's solve this problem by considering all the details. Pay attention to each piece of information, remember to add or subtract as needed, and perform the calculations step by step	99.6
	GLaPE-based (Ours)	Let's approach this problem strategically, methodically, and innovatively, exploring groundbreaking solutions.	99.3

Table 7: Optimization results (optimal prompt and corresponding accuracy) of our GLaPE-based prompt optimization method and other popular methods.

A.2 Computation Detail of Figure 2

First, we calculate the self-consistency c_i for each prompt according to the definition in Section 3.2, which are:

$$c_1 = 100.0, \quad c_2 = 70.0, \quad c_3 = 70.0,$$
$$c_4 = 40.0, \quad c_5 = 30.0.$$

Thus, the loss function L_{self} is:

$$L_{\text{self}} = \sum_{i=1}^{5} (f_i - c_i)^2 = (f_1 - 100)^2 + (f_2 - 70)^2 + (f_3 - 70)^2 + (f_4 - 40)^2 + (f_5 - 30)^2.$$

Next, we calculate the loss function of mutual-consistency refinement L_{refine}, which is:

$$L_{\text{refine}} = \sum_{1 \le i < j \le 5} \mathbb{1}_{a_i = a_j} (f_i - f_j)^2,$$

since prompts 1 and 2 share the same answer 31, while prompts 3, 4, and 5 share the same answer 36.

Clearly, f_1 and f_2 are unrelated to f_3, f_4, and f_5 since their answers are different.

The evaluation scores are then computed as follows (ignoring the coefficient 0.5 for both L_{self} and L_{refine}):

$$f_1, f_2 = \arg\min_{f_1, f_2} \left[(f_1 - 100)^2 + (f_2 - 70)^2 + (f_2 - 70)^2 + (f_1 - f_2)^2 \right]$$

and

$$f_3, f_4, f_5 = \arg\min_{f_3, f_4, f_5} \left[(f_3 - 70)^2 + (f_4 - 40)^2 + (f_5 - 30)^2 + (f_3 - f_4)^2 + (f_3 - f_5)^2 + (f_4 - f_5)^2 \right].$$

Ultimately, the solution is:

$$f_1 = 87.9, \quad f_2 = 81.8, \quad f_3 = 50.0,$$
$$f_4 = 45.7, \quad f_5 = 44.2.$$

A.3 Further Comparison of Prompt Optimization Methods

To emphasize the efficacy of our method, we conducted additional comparisons between our GLaPE method and other recent prompt optimization approaches for private LLMs, including APE (Zhou et al., 2022), APO (Pryzant et al., 2023), and PE2 (Ye et al., 2023). The results are presented in Table 7. These comparisons demonstrate that GLaPE is not only competitive but also exceeds the performance of other existing supervised methods in various cases.

	AddSub	AQuA	Big-Bench Date	GSM8K	MultiArith	SVAMP	StrategyQA	MATH
Cleaned Dataset	0.61(+0.17)	0.40(+0.36)	0.94(+0.06)	0.69(+0.20)	0.93(+0.05)	0.81(+0.12)	0.41(+0.13)	0.61(+0.14)
Control Group	0.42(-0.07)	-0.01(-0.05)	0.86(-0.02)	0.40(-0.09)	0.84(-0.04)	0.61(-0.08)	0.16(-0.02)	0.46(-0.01)
Original Dataset	0.44	0.04	0.88	0.49	0.88	0.69	0.18	0.47

Table 8: Comparison of Spearman correlation coefficients (↑) before and after excluding challenging questions that surpass the intrinsic capabilities of LLM. Evaluation of the control group is conducted by randomly selecting 10 subsets of the original dataset, and the average Spearman correlation coefficient is computed.

A.4 Spearman Correlation Coefficients on Cleaned Datasets

It is imperative to recognize that our methodology evaluates prompts on individual questions, and the evaluation score of a prompt across the entire dataset is derived from the sum of its evaluation scores on each question. Consequently, inaccuracies in evaluations for questions stated in Section 7 can significantly compromise the effectiveness of the overall dataset evaluation, particularly on challenging datasets. To gauge the impact of challenging questions on our GLaPE, we exclude questions for which no prompt results in a correct answer with a self-consistency level greater than 50% from the dataset. The cleaned dataset was then compared to a control group, consisting of an equally large subset of the original dataset, to mitigate the influence of dataset size bias. On the initial dataset, the control group, and the cleaned dataset, we calculate the Spearman correlation coefficient.

In Table 8, the Spearman correlation coefficient on the cleaned dataset demonstrates a considerable improvement compared to that on the original dataset or control group. This improvement underscores the pronounced adverse influence of intricate questions on our evaluation process.