

# Exploring Generative Process Reward Modeling for Semi-Structured Data: A Case Study of Table Question Answering

Lei Tang<sup>2</sup> Wei Zhou<sup>1,3</sup> Mohsen Mesgar<sup>3</sup>

<sup>1</sup>University of Augsburg, Germany

<sup>2</sup>Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany

<sup>3</sup>Bosch Center for Artificial Intelligence, Renningen, Germany

letang9-c@my.cityu.edu.hk {wei.zhou3|mohsen.mesgar}@de.bosch.com

## Abstract

Process reward models (PRMs) enhance complex reasoning in large language models (LLMs) by evaluating candidate solutions step-by-step and selecting answers based on aggregated step scores. While effective in domains such as mathematics, their applicability to tasks involving semi-structured data, like table question answering (TQA), remains unexplored. TQA poses unique challenges for PRMs, including abundant irrelevant information, loosely connected reasoning steps, and domain-specific reasoning. This work presents the first systematic study of PRMs for TQA. We evaluate state-of-the-art generative PRMs on TQA from both answer and step perspectives. Results show that PRMs that combine textual and code verification can aid solution selection but struggle to generalize to out-of-domain data. Analysis reveals a weak correlation between performance in step-level verification and answer accuracy, possibly stemming from weak step dependencies and loose causal links. Our findings highlight limitations of current PRMs on TQA and offer valuable insights for building more robust, process-aware verifiers.

## 1 Introduction

Recent advances in process reward models (PRMs) have demonstrated remarkable improvements in the reasoning capabilities of large language models (LLMs) (Wang et al., 2024b; Yin et al., 2025; Ong et al., 2025). PRMs enable sampling-based test time scaling (TTS) by grading multiple solutions step-wise. The best solution is then selected based on aggregated step scores, e.g., average step scores. This verification-driven paradigm has proven highly effective in domains such as math (Zhao et al., 2025; Zhang et al., 2025b,a), visual (Ong et al., 2025; Wang et al., 2025), and textual QA (Yin et al., 2025). However, to the best of our knowledge, no previous work explores the effectiveness of PRMs on tasks using semi-structured

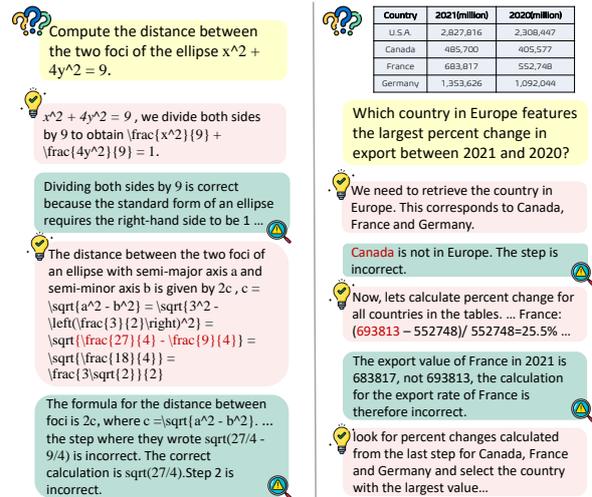


Figure 1: Comparing the verification process for a math (left) and a TQA problem (right). Examples draw from Zhao et al. (2025) and Zhou et al. (2025). Compared to math, not all information in TQA is relevant.

data, e.g., tables. This raises a natural and important question: *Is the success of PRMs in domains like math transferable to other complex reasoning tasks, where semi-structured data are involved?* To answer the question, we focus on table question answering (TQA), where a system needs to solve a question based on information in a table.

TQA differs fundamentally from math problems. In math, most given information is relevant, and reasoning steps are tightly linked in deductive chains. In contrast, tables in TQA can be large with abundant irrelevant information for solving a question. Moreover, the reasoning steps are often loosely connected, i.e., later steps may not strictly follow from earlier ones and can start a new topic. Consider “Step 1” and “Step 2” on the right side of Figure 1. Nevertheless, both tasks require multi-step and complex reasoning to be solved.

We hypothesize that TQA can benefit from PRM-based verification. Unlike existing TQA approaches, which predominantly optimize policy

models through prompt engineering (Zhou et al., 2025; Wang et al., 2024c), or fine-tuning (Zhang et al., 2024; Deng and Mihalcea, 2025), PRMs are model agnostic: they can be applied to outputs from any policy models and rewards given by PRM can be used to further improve performance of policy model via reinforcement learning with verifiable rewards (RLVR) (Su et al., 2025). Moreover, solutions selected by PRMs reflect both answer and step quality, compared to majority voting that considers only answer frequencies (Liu et al., 2024; Zhou et al., 2025). Note that the effectiveness of majority voting depends strongly on the base model’s capability: it has to produce the correct answer in most sampled outputs. By contrast, PRMs are specifically trained for step-level evaluation. Last but not least, PRMs can enhance system trustworthiness, providing human-readable rationales for verification decisions on step correctness.

Yet, applying PRMs to TQA is non-trivial. Retrieval noise, loosely connected reasoning steps, and domain-specific reasoning patterns may limit their effectiveness. As an initial step, this work assesses the potential of applying PRMs for step-level verification in TQA. We explore to what extent such a step-based evaluation paradigm assists in solution selection, with a focus on PRMs that generate human-understandable rationales. Specifically, we evaluate multiple PRM variants from both answer-level and step-level perspectives against strong baselines, including self-consistency (Wang et al., 2022) and LLM-as-a-judge (Gu et al., 2024). Our results show that PRMs can aid in selecting high-quality solutions for TQA, particularly when combining textual and code verification. However, their effectiveness remains limited, with notable weaknesses such as poor generalization to out-of-domain datasets.

Through qualitative and quantitative analysis of verification chains, we uncover a weak correlation between step verification and answer correctness: better average step scores do not necessarily yield higher answer accuracy in TQA. This stems from characteristics of the problem profile, such as weak step dependence and loose causal relationships between reasoning steps. By bridging PRMs with TQA, our work clarifies when and why PRMs succeed or fail, and provides insights for designing more robust, process-aware systems for complex multi-step reasoning tasks. Code is available.<sup>1</sup>

<sup>1</sup><https://github.com/boschresearch/TQA-PRM>

## 2 Examining PRMs on TQA.

This section describes the PRMs we examine, the training setup, baselines, TQA datasets for training and evaluation, as well as evaluation protocols.

**Problem Formulation.** Let  $T$  denote a table,  $Q$  a natural language question, and  $S$  a set of candidate reasoning paths generated by a policy model  $M_p$ , where each path produces a predicted answer  $A'$ . A PRM evaluates each reasoning step  $s_i \in S$  by generating a natural language rationale  $v_i \in V$  and assigning a scalar reward  $r_i \in R$ . The goal is to identify the reasoning path  $S^* \in S$  that maximizes the likelihood of producing a correct answer, i.e.,  $A' = A$ , where  $A$  is the ground-truth answer.

**Examined PRMs.** We focus on state-of-the-art generative PRMs that output both step-level rewards and rationales explaining these rewards. GENPRM (Zhao et al., 2025) verifies each step through a two-stage process: (1) a textual verification stage, in which the PRM examines a step using chain-of-thought (CoT) reasoning (Wei et al., 2023), and (2) a code verification stage, where Python code is generated and executed for the current step. Step correctness is determined based on the CoT reasoning, generated code, and its execution results. TEXTUAL PRM (Zhao et al., 2025) is an ablation of GENPRM that removes the code verification stage, relying solely on CoT analysis. MULTI-DIMENSIONAL ANALYSIS (MDA) (She et al., 2025) is a textual verifier that evaluates each step from five distinct perspectives: restatement, data check, logical consistency, numeric accuracy, and format. Finally, REPHRASE-REACT-ANALYSIS (RRA) (Deng et al., 2023) rephrases the current step as a question, solves the question using the ReAct method (Yao et al., 2022), and then provides a step-by-step analysis. Implementation details are provided in Appendix A.1.

**Training PRMs for TQA.** We use the training sets of WTQ (Pasupat and Liang, 2015) and TabFact (Chen et al., 2020) as our source problem sets for obtaining reasoning paths. For each question table pair, we sample  $N = 8$  reasoning paths from Qwen2.5-7B-Instruct (Qwen et al., 2025) with a temperature of 0.6. For verification generation with GENPRM, we adopt the default hyperparameters from Zhao et al. (2025) and perform relative progress estimation (RPE), defined as the ratio between the scores of  $s_i$  and  $s_{i-1}$ . These

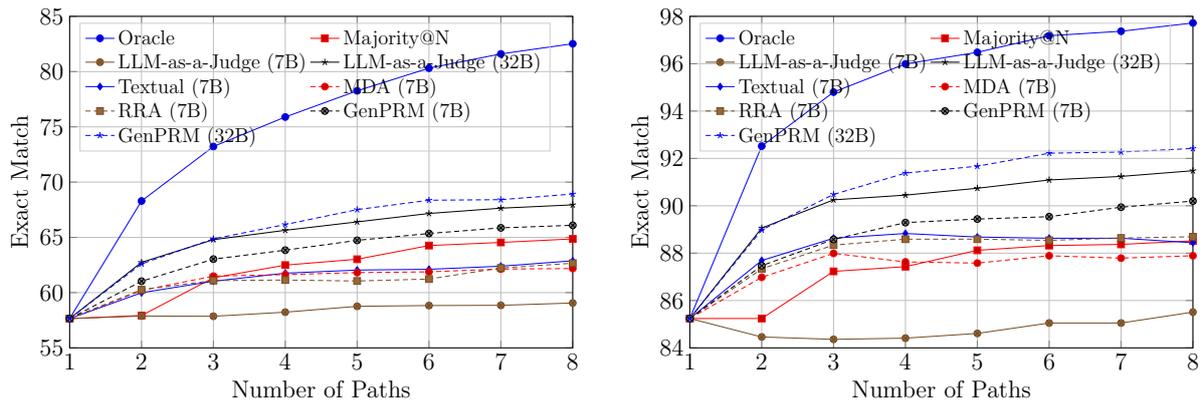


Figure 2: EM measured by  $Best@N$  across varying numbers of paths. Left: WTQ. Right: Tabfact. ORACLE always selects correct answers if they exist. MAJORITY selects the most frequent predictions as answers.

scores are approximated using Monte Carlo sampling (Wang et al., 2024a). We employ QwQ-32B (Team, 2025) as the rationale generation model to produce both textual analyses and Python code. The normalized probability of the Yes token is taken as the final step reward. If the rewards generated by RPE and the rationale model disagree, the instance is discarded. We also filter out incomplete outputs, instances with improper formatting, and cases where the predicted label is neither Yes nor No. To generate training data for textual verification, QwQ-32B is used. We provide analysis of generated data quality in Appendix A.2. For PRM training, we fully fine-tune DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025), chosen for its large context length, on the filtered dataset. The training data are formatted as standard multi-turn conversations. Training parameters can be found in Appendix A.3.

**Baselines and Evaluation.** We compare against the following baselines: **ORACLE** which selects the correct answer from the candidate set, serving as an upper bound on verification performance. **MAJORITY@N** chooses the most frequent answer among  $N$  candidates; in the case of a tie, the first prediction is selected. **LLM-AS-JUDGE** uses the QwQ-32B model as an LLM judge, with the prompt provided in Appendix A.4. We adopt two evaluation paradigms: **answer-focused** and **process-focused**. In the *answer-focused* setting, we apply a *best-of-N* strategy: each sampled reasoning path is scored by verifiers, and the final answer is taken from the path with the highest average step score. Performance is measured using exact

match (EM). We report results in both **in-domain** and **out-of-domain** scenarios, evaluating on the test sets of WTQ (Pasupat and Liang, 2015), TabFact (Chen et al., 2020), CRT (Zhang et al., 2023), and SciTab (Lu et al., 2023). Dataset statistics are shown in Appendix A.5. In the *process-focused* settings, we manually annotate 300 instances which are randomly sampled from the four datasets, resulting in 1916 steps (1366 correct and 550 incorrect).

### 3 Results and Analysis

Figure 2 presents EM of various PRMs and baseline approaches across different path sampling budgets on WTQ and TabFact. Results for step-level verification are shown in Table 1.

**GENPRMS achieve promising performance compared to other methods, while falling short of ORACLE for in-domain settings.** From Figure 2, we observe that GENPRMS consistently delivers the strongest results **compared with examined generative PRMs**. We attribute it to its integrated code verification component. A detailed analysis is provided in the following paragraph. **In comparison to baseline methods**, both the 7B and 32B variants of GENPRMS substantially outperform majority voting, underscoring the benefit of process-level verification for answer selection. Nevertheless, the fine-tuned 7B GENPRMS lags behind the 32B LLM-as-a-judge. Though scaling GENPRMS to 32B closes this gap, performance advantages over LLM-as-a-judge remain limited. Across all methods, a substantial performance gap remains compared to ORACLE, indicating considerable room for improvement in leveraging multiple

Method	WTQ	TabFact	CRT	SciTab
LLM-as-a-Judge (32B)	90.29	88.56	81.60	84.82
Textual PRM (7B)	81.06	80.35	71.58	74.70
MDA (7B)	80.79	81.25	72.10	72.61
RRA (7B)	71.52	84.82	70.37	77.08
GenPRM w/o code (7B)	80.99	80.29	71.58	75.00
GenPRM (7B)	91.38	90.13	77.58	84.93
GenPRM w/o code (32B)	84.79	87.50	79.80	80.92
GenPRM (32B)	<b>93.71</b>	<b>94.72</b>	<b>83.22</b>	<b>85.71</b>

Table 1: Process verification accuracy of PRMs and LLM-as-a-Judge on four TQA datasets. RRA stands for Rephrase, React, Analysis (7B). MDA stands for multi-dimensional analysis.

candidate solutions. We discuss the efficiency of different methods in Appendix A.6.

### Incorporating code verifier improves PRM performance in both answer and step correctness.

From the answer-level perspective, a comparison between CoT PRM and GENPRM in Figure 2 shows that GENPRM substantially outperforms CoT PRM. Since CoT PRM is trained on the same data as GENPRM but without the code verification component, this performance gain can be attributed to the incorporation of code verification. From the process-level perspective, we examine step-level correctness by comparing rewards derived from pure textual analysis with those obtained after applying code verification in GENPRM. The results, presented in the last four rows of Table 1, demonstrate that incorporating code verification markedly improves step-level correctness verification across datasets and model sizes.

**When Code Helps and Hurts.** We conduct a manual analysis to characterize when code-based verification is beneficial and when it is not. Specifically, we review 60 randomly sampled steps on which GenPRM with code and without code disagree (e.g., a step labeled 1 by GenPRM with code but 0 by the model without code). Among these 60 instances, GenPRM with code assigns the correct label in 43 cases and an incorrect label in 17 cases. Within the correctly classified instances, we observe two recurring patterns in which code execution provides substantial benefits: (1) **Numerical operations.** Executing code supports precise and quantitative reasoning (e.g., counting, aggregation, and threshold checks), thereby reducing ambiguity and limiting error propagation. (2) **Targeted retrieval.** Programmatic access to specific rows or columns provides accurate relevant evidence and

enables verification of individual cell values. We also identify two common failure modes of code-based verification: (1) **Exhaustive retrieval.** The program retrieves large portions of a table (or even a full table) even when the claim depends on a small subset, implicitly shifting the burden to long-context recall rather than selective filtering. (2) **Index misunderstandings.** The model sometimes confuses row indexing conventions, inconsistently switching between absolute row indices (including the header) and row indices that exclude the header, which leads to incorrect cell retrieval.

### Fine-tuned PRMs with code verification are not robust on out-of-domain datasets.

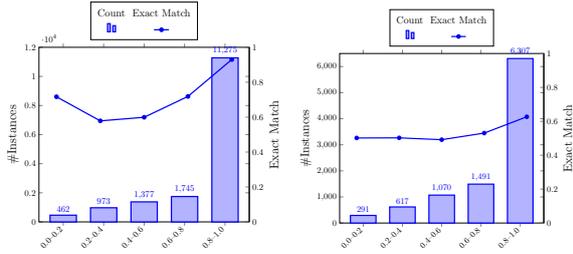
As shown in Figure 2, GENPRM consistently outperforms majority voting on in-domain datasets. However, this advantage diminishes in out-of-domain settings, with performance dropping to 53 vs.56 on CRT and 60 vs.62 on SciTab, as reported in Table 3. We attribute this gap to differences in problem complexity: CRT contains highly complex questions, while SciTab requires specialized scientific knowledge. These challenging scenarios often lead the policy model to generate reasoning paths containing spurious or noisy steps, which hinder effective verification. An analysis of variance in reasoning complexity across datasets is provided in Appendix A.7. In summary, supervised fine-tuned PRMs are prone to generalization issues. Mitigating this limitation may require more balanced fine-tuning data that reflects varying verification difficulties, or additional post-training techniques to enhance the generalizability of PRMs.

### Process-level evaluation highlights the competitiveness of GENPRM.

While the fine-tuned 7B GENPRM under-performs the 32B LLM-as-a-judge in answer evaluation, process-level verification reveals a different trend: 7B GENPRM attains comparable or higher accuracy across most datasets, with the exception of CRT (Table 1). This suggests that strong step-level verification does not necessarily translate into improved answer accuracy. We elaborate on this point in the next paragraph. Moreover, purely text-based PRMs consistently lag behind GENPRM, underscoring the benefit of integrating code verification.

### Weak correlation between PRM verification and answer correctness in TQA.

Our manual analysis reveals that average step scores are not strongly correlated with answer correctness. Figure 3 con-



(a) GenPRM (7B) on Tabfact (b) GenPRM (7B) on SciTab

Figure 3: Counts per confidence bin (bars, left axis) and Exact Match (line, right axis). Bins are PRM score intervals:  $[0, 0.2)$ ,  $[0.2, 0.4)$ ,  $[0.4, 0.6)$ ,  $[0.6, 0.8)$ ,  $[0.8, 1.0]$ . It shows that average step scores are not strongly correlated with answer correctness.

firmly this trend by plotting EM accuracy across bins of average step scores. We observe that solutions with higher average step scores do not necessarily feature higher EM accuracy. A likely reason is that averaging step-level rewards treats all steps as equally informative, regardless of their causal contribution to the final answer. Consequently, (i) chains containing many irrelevant but locally correct steps receive inflated scores, whereas (ii) chains with a local error that does not affect the final answer are over-penalized. Examples of both cases are provided in Appendix A.8. Unlike in math, reasoning chains in TQA are more loosely linked, so an error at step  $i$  does not necessarily propagate to subsequent steps or the final prediction. This reduces the effectiveness of step verification for answer selection. Nevertheless, PRMs remain valuable; the core bottleneck lies in the current answer-focused evaluation paradigm and the absence of process-annotated evaluation datasets for TQA. We advocate future work on process-level evaluation and modeling approaches that strengthen step dependencies and assess each step by both its local correctness and its causal contribution to the final answer, for example, by filtering irrelevant steps.

**Potential Alternative Aggregation Strategies.** We observe that simple uniform averaging is sub-optimal, as it implicitly assumes that all reasoning steps contribute equally to the final answer; this assumption allows irrelevant or non-causal steps to influence the estimated solution quality and obscure the impact of steps that are truly determinative. Beyond uniform averaging, we explore alternative aggregation strategies that more accurately reflect answer quality in TQA. A simple and

intuitive approach is to assign greater weight to the final reasoning steps. When evaluated using a 32B coding verifier, this strategy yields consistent improvements of +3.56% EM on WTQ, +1.14% on TabFact, +4.55% EM on CRT, and +2.69% on SciTab. In addition, relevance- and causality-aware aggregation strategies represent promising directions for future study. A relevance-aware strategy assigns a relevance score to each step, indicating the step’s relevance to solving the problem. Similarly, a causality-aware strategy assigns a causality score to each step, indicating whether the step influences the final outcome. Incorporating these weights when computing a solution score can down-weight steps that do not contribute to the final answer, thereby leading to a score that more faithfully reflects answer quality.

## 4 Conclusions

In this paper, we examined four generative PRMs on TQA and compared them to strong baselines such as LLM-as-a-judge and majority vote from both answer and process evaluation perspectives. We found that PRMs that combine textual and code verification can aid solution selection but struggle to generalize to out-of-domain data. Our analysis suggests that the primary bottleneck is not step-level verification, but rather the aggregation of step rewards. Uniform averaging is inadequate for TQA as it assigns equal weight to all steps, regardless of their relevance to the problem or their causal influence on the final answer. Future work can therefore investigate relevance-aware and causality-aware aggregation strategies to generate solution scores that more faithfully reflect answer quality.

## Limitations

To examine the effectiveness of generative PRMs over semi-structured data, we choose table question answering as a proxy task for initial exploration. Nevertheless, there exist other tasks involving semi-structured data, such as table generation or table summarization. Future studies can explore these tasks to validate the effectiveness of PRMs. Moreover, we mainly train PRM using the Qwen models, following the practice of PRM training in previous work (Zhao et al., 2025). Though we experiment with varying model sizes, results from a different model family might be insightful.

## References

- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. [Tabfact: A large-scale dataset for table-based fact verification](#). *Preprint*, arXiv:1909.02164.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *ArXiv*, abs/2501.12948.
- Naihao Deng and Rada Mihalcea. 2025. [Rethinking table instruction tuning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21757–21780, Vienna, Austria. Association for Computational Linguistics.
- Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. [Rephrase and respond: Let large language models ask better questions for themselves](#). *Preprint*, arXiv:2311.04205.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. [A survey on llm-as-a-judge](#). *ArXiv*, abs/2411.15594.
- Tianyang Liu, Fei Wang, and Muhao Chen. 2024. [Rethinking tabular data understanding with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.
- Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. [SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables](#).
- Brandon Ong, Tej Deep Pala, Vernon Toh, William-Chandra Tjhi, and Soujanya Poria. 2025. [Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned](#).
- Panupong Pasupat and Percy Liang. 2015. [Compositional semantic parsing on semi-structured tables](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. [Rprm: Reasoning-driven process reward modeling](#). *Preprint*, arXiv:2503.21295.
- Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. [Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains](#). *ArXiv*, abs/2503.23829.
- Qwen Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#).
- Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024a. [Math-shepherd: Verify and reinforce llms step-by-step without human annotations](#). *Preprint*, arXiv:2312.08935.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. [Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. [Visualprm: An effective process reward model for multimodal reasoning](#). *ArXiv*, abs/2503.10291.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. [Self-consistency improves chain of thought reasoning in language models](#). *ArXiv*, abs/2203.11171.
- Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024c. [Chain-of-table: Evolving tables in the reasoning chain for table understanding](#). *Preprint*, arXiv:2401.04398.
- 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](#). *Preprint*, arXiv:2201.11903.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. [React: Synergizing reasoning and acting in language models](#). *ArXiv*, abs/2210.03629.

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, and Xuanjing Huang. 2025. [Dynamic and generalizable process reward modeling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4203–4233, Vienna, Austria. Association for Computational Linguistics.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025a. [Process-based self-rewarding language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18097–18110, Vienna, Austria. Association for Computational Linguistics.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024. [TableLlama: Towards open large generalist models for tables](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044, Mexico City, Mexico. Association for Computational Linguistics.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023. [CRT-QA: A dataset of complex reasoning question answering over tabular data](#).

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. [The lessons of developing process reward models in mathematical reasoning](#). *Preprint*, arXiv:2501.07301.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. 2025. [Genprm: Scaling test-time compute of process reward models via generative reasoning](#). *Preprint*, arXiv:2504.00891.

Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2025. [Efficient multi-agent collaboration with tool use for online planning in complex table question answering](#). *Preprint*, arXiv:2412.20145.

## A Appendix

### A.1 Implementation and Adaptation Details

We adapt five perspectives from [She et al., 2025](#)—previous steps analysis, now step analysis, data source analysis, consistency analysis, and calculation analysis—into ours: restatement, data check, logical consistency, numeric accuracy, and format, as shown in Figure 10. For RRA, we follow [Deng et al., 2023](#): the PRM first rephrases the current step, then solves the rephrased question using the ReAct paradigm ([Yao et al., 2022](#)), and finally provides a step-by-step analysis conditioned on the current step, the rephrased problem, and the ReAct trace, as shown in Figure 11.

Label	After Filtering		Before Filtering	
	#Steps	%	#Steps	%
Correct	254	94.42	240	84.21
Incorrect	15	5.58	45	15.79

Table 2: Manual evaluation of reasoning-step correctness before and after the GenPRM filtering process. Filtering substantially improves data quality, increasing step-level accuracy from 84.21% to 94.42%.

### A.2 Data Quality Analysis

Training data quality may partially account for the limited performance of GenPRM on TQA. To examine this, we randomly sample 50 reasoning paths before and after the GenPRM filtering procedure and manually assess their correctness. As shown in Table 2, the filtering process effectively removes incorrect steps. After filtering, which produces the data used for PRM training, the step-level accuracy reaches 94%, indicating overall high data quality. This suggests that data noise is unlikely to be the primary bottleneck for GenPRM performance on TQA.

### A.3 Training Parameters

Models were fine-tuned with full-parameter. Training proceeded for a single epoch with a batch size of 64 and a maximum sequence length of 2,048 tokens. We initialized the learning rate at  $2 \times 10^{-5}$  and applied a cosine decay schedule. Optimization used AdamW with bfloat16 (bf16) precision.

### A.4 Prompts

Prompts for generating reasoning paths for WTQ, TabFact, CRT and SciTab are shown in Figure 6, 7, 6 and 8, respectively. Prompts for generating rationales for a textual PRM are presented in Figure 9. Prompts for generating rationales for MULTI-DIMENSIONAL ANALYSIS and REPHRASE, REACT, ANALYSIS are shown in Figure 10 and Figure 11, respectively. Refer to Figure 12 for the prompt for data generation for GENPRM. Lastly, Figure 13 shows the prompt for collecting data for LLM-as-a-judge.

### A.5 Dataset Statistics and Licenses

The statistics of datasets we used are shown in the Table 4. WTQ ([Pasupat and Liang, 2015](#)) is under the license of CC-BY-SA-4.0<sup>2</sup>. TabFact ([Chen](#)

<sup>2</sup><https://creativecommons.org/licenses/by-sa/4.0/>

Method	WTQ	TabFact	CRT	SciTab
Upper Bound	82.52	97.72	71.01	88.39
Pass@1	57.65	85.24	51.09	57.10
Majority@8	64.86	88.52	56.45	62.58
LLM-as-a-Judge (7B)	59.06	85.51	50.13	58.00
LLM-as-a-Judge (32B)	67.94	91.48	55.49	64.62
CoT-style PRM (7B)	62.88	88.44	51.51	60.07
CoT-style PRM (32B)	67.48	91.86	55.08	63.36
Multi-dimensional (7B)	62.18	87.89	51.94	59.96
Rephrase-React-Analysis (7B)	62.66	88.69	52.74	59.93
GenPRM (7B)	66.09	90.20	53.31	60.64
GenPRM (32B)	68.92	92.42	56.00	63.78

Table 3: Main results on in-domain (WTQ, TabFact) and out-of-domain (CRT, SciTab) datasets, measured with the  $Best@N$  metric ( $N = 8$ ). PRMs with code execution achieve the best overall performance.

Datasets	#instances	Domain
WTQ	4344	Wikipedia
TabFact	2024	Wikipedia
SCITAB	1224	scientific paper
CRT	728	Wikipedia

Table 4: Test data statistics.

	WTQ	TabFact	CRT	SciTab
Textual PRM	61.56	80.60	56.82	90.32
Multi-Dimensional	82.75	88.86	83.32	112.64
Rephrase-React-Analysis	150.78	133.81	208.07	163.73
GenPRM w/o code (7B)	107.22	122.41	154.84	188.31
GenPRM (7B)	222.93	228.38	290.30	469.45
GenPRM w/o code (32B)	185.96	159.95	255.65	191.29
GenPRM (32B)	508.66	433.21	608.91	529.66

Table 5: Time efficiency (seconds; lower is better) of PRM variants across four tasks. GenPRMs with code have longer inference times than Textual PRMs, MDA, RRA, and GenPRMs w/o code, indicating that code generation substantially increases latency.

et al., 2020), CRT (Zhang et al., 2023) and SCITAB (Lu et al., 2023) are under the MIT<sup>3</sup> license.

### A.6 Time Efficiency.

Despite dataset-specific variation, PRMs without code (Textual, MDA, RRA, and GenPRM without execution) are consistently cheaper than GenPRM. Code execution introduces significant runtime overhead, especially with larger models. This suggests that integrating code improves performance at the cost of higher latency and longer inference time.

### A.7 Variance in Reasoning Complexity Across Datasets

Table 6 reveals a clear in-/out-of-domain split in path structure. WTQ and TabFact predominantly exhibit *consistent* trajectories (86% and 81%), i.e.,

Reasoning Paths	WTQ	TabFact	CRT	SciTab
Total	100	100	100	100
Consistent	86	81	63	50
Inconsistent	14	19	37	50

Table 6: Distribution of human-annotated reasoning paths by consistency. A path is **consistent** if its labels are **monotonically decreasing** (e.g.,  $[1, 1, 1, 0, 0]$ ); paths that are always correct or always wrong also count as consistent. A path is **inconsistent** if it exhibits any  $0 \rightarrow 1$  jump or multiple flips (e.g.,  $[0, 1, 1, 0, 1]$ ).

monotone label sequences with at most one change-point. By contrast, CRT and SciTab show far more *inconsistent* paths (37% and 50%), with multiple flips between correct and incorrect steps. This suggests fewer spurious/noisy steps in the in-domain sets (WTQ/TabFact) and a higher noise ratio in CRT/SciTab.

### A.8 Examples of Inflated Scores and Over-Penalization

Examples with inflated and penalized scores are shown in Figure 4 and Figure 5, respectively.

<sup>3</sup><https://opensource.org/license/mit>

### Chain with Locally Correct Steps but an Incorrect Final Answer

Table is: [['Team', 'No', 'Driver', 'Class', 'Rounds'],  
['Josef Kaufmann Racing', '4', 'Robin Frijns', '', 'All'],  
['Josef Kaufmann Racing', '5', 'Hannes van Asseldonk', 'R', 'All'],  
['Josef Kaufmann Racing', '6', 'Petri Suvanto', 'R', 'All'],  
['Möfcke Motorsport', '7', 'Maciej Bernacik', 'R', 'All'],  
['Möfcke Motorsport', '8', 'Timmy Hansen', '', 'All'],  
['EuroInternational', '11', 'Daniil Kvyat', 'R', 'All'],  
['EuroInternational', '12', 'Carlos Sainz, Jr.', 'R', 'All'],  
['EuroInternational', '14', 'Michael Lewis', '', 'All'],  
['DAMS', '15', 'Javier Tarancón', '', 'All'],  
['DAMS', '16', 'Dustin Sofyan', '', '5'],  
['DAMS', '16', 'Luciano Bacheta', '', '2013'],  
['DAMS', '17', 'Fahmi Ilyas', '', '120136'],  
['DAMS', '17', 'Dustin Sofyan', '', '8'],  
['Eifelland Racing', '18', 'Facundo Regalia', '', 'All'],  
['Eifelland Racing', '19', 'Côme Ledogar', '', 'All'],  
['Eifelland Racing', '20', 'Marc Coleselli', 'R', 'All'],  
['Fortec Motorsport', '24', 'Jack Harvey', '', 'All'],  
['Fortec Motorsport', '25', 'George Katsinis', '', 'All'],  
['Fortec Motorsport', '26', 'Christof von Grünigen', '', 'All']]

Question is: how many teams had at least three drivers each for the 2010 formula bmw europe season?

Step 1: First, we need to identify which teams have at least three drivers each. We'll count the drivers for each team. (✓)

Step 2: Josef Kaufmann Racing: 3 drivers (4, 5, 6), (✓)

Step 3: Möfcke Motorsport: 2 drivers (7, 8), (✓)

Step 4: EuroInternational: 3 drivers (11, 12, 14), (✓)

Step 5: DAMS: 4 drivers (15, 16, 17, 16), (✓)

Step 6: Eifelland Racing: 3 drivers (18, 19, 20), (✓)

Step 7: Fortec Motorsport: 3 drivers (24, 25, 26), (✓)

Step 8: Count the teams with at least 3 drivers: 4 (Josef Kaufmann Racing, EuroInternational, DAMS, Eifelland Racing, Fortec Motorsport). Final answer: \boxed{4}. (✗)

Figure 4: A chain with many locally correct steps but an incorrect final answer, inflating the average step score. Labels here are human-annotated.

### Chain with Local Errors but Correct Final Answer

Table is: [['Polling Firm', 'Source', 'Date Published', 'N.Anastasiades', 'G.Lillikas', 'S.Malas', 'Others'],  
['RAI Consultants', '[1][dead link]', '16 September 2012', '37.2%', '14.2%', '21.9%', '1.5%'],  
['Evresis', '[2]', '18 September 2012', '35.2%', '17.5%', '19.7%', '1.7%'],  
['Noverna', '[3]', '23 September 2012', '35.02%', '15.81%', '17.78%', ''],  
['Prime Consulting Ltd', '[4]', '7 October 2012', '34.7%', '17.4%', '18.5%', ''],  
['CMR Cypronetwork / Cybc', '[5][dead link]', '18 October 2012', '36.9%', '17%', '23.8%', '1.2%'],  
['Evresis', '[6]', '2 November 2012', '36.9%', '17.7%', '20.6%', '1.4%'],  
['RAI Consultants', '[7]', '4 November 2012', '38.8%', '19.8%', '21.1%', '2.3%'],  
['CMR Cypronetwork / Cybc', '[8]', '15 November 2012', '36.8%', '18.9%', '22.8%', '1.6%'],  
['Prime Consulting Ltd', '[9]', '18 November 2012', '35.9%', '18.7%', '19.6%', '0.6%'],  
['Evresis', '[10]', '27 November 2012', '37.1%', '19.6%', '20.8%', '0.6%'],  
['Noverna', '[11]', '2 December 2012', '35.6%', '17.2%', '18.1%', '4.1%'],  
['Prime Consulting Ltd', '[12]', '3 December 2012', '35%', '19.1%', '18.6%', '1.4%'],  
['CMR Cypronetwork / Cybc', '[13][dead link]', '17 December 2012', '37.1%', '20.4%', '23.1%', '3.1%'],  
['Evresis', '[14]', '22 December 2012', '37.4%', '19.8%', '21.8%', '0.5%'],  
['RAI Consultants Ltd', '[15][dead link]', '13 January 2013', '40.3%', '17.9%', '20.5%', '6.1%'],  
['CMR Cypronetwork / Cybc', '[16]', '17 January 2013', '38%', '19.7%', '23.7%', '2.7%'],  
['Prime Consulting Ltd', '[17]', '27 January 2013', '39.2%', '18.8%', '19.8%', '4%'],  
['Evresis', '[18]', '1 February 2013', '40.8%', '19.9%', '22.2%', '2.5%'],  
['Prime Consulting Ltd', '[19]', '4 February 2013', '39.8%', '19.3%', '20%', '3%'],  
['Prime Consulting Ltd', '[20]', '9 February 2013', '40.6%', '19.6%', '20.4%', '2.9%'],  
['RAI Consultants Ltd', '[21]', '9 February 2013', '42.1%', '19.4%', '21.1%', '4.4%'],  
['CMR Cypronetwork / Cybc', '[22]', '9 February 2013', '39.9%', '20.2%', '24.2%', '3%'],  
['Average (only valid votes)', '2013', '2013', '48.4%', '22.52%', '25.29%', '3.79%']]

Question is: which poll did anastasiades win by a larger margin, feb 1 2013 or dec 3 2012?

Step 1: Identify the relevant polls for comparison - February 1, 2013 poll by Prime Consulting Ltd - December 3, 2012 poll by Prime Consulting Ltd, (X)

Step 2: Extract the percentages for N. Anastasiades from these two polls - February 1, 2013: 40.6% - December 3, 2012: 35.6%, (X)

Step 3: Calculate the difference between N. Anastasiades and the second-place candidate (G. Lillikas) for each poll- For February 1, 2013:  $40.6\% - 19.6\% = 21.0\%$  - For December 3, 2012:  $35.6\% - 17.2\% = 18.4\%$ , (X)

Step 4: Compare the margins - February 1, 2013: 21.0% - December 3, 2012: 18.4%, (X)

Step 5: Conclude which poll had the larger margin - The February 1, 2013 poll shows a larger margin for N. Anastasiades. Final answer:  $\boxed{\text{February 1, 2013}}$ , (✓)

Figure 5: An example of a chain with local errors that does not affect the final answer; in this case, the chain's score is over-penalized. Labels here are human-annotated.

### Reasoning Path Generation prompt on WTQ and CRT

Please inspect the table(s) and then provide an Answer to the question. Please reason step by step. Attention: You MUST put your final answer using the following format: Final answer: `\boxed{final answer}`. Attention: You MUST put your final answer using the following format: Final answer: `\boxed{final answer}`.

Figure 6: Prompt for Generating Reasoning Paths on WTQ and CRT Dataset.

### Reasoning Path Generation prompt on TabFact

Please inspect the table(s) and then provide a True or False answer to the question. Please reason step by step. Attention: You MUST put your final answer using the following format: Final answer: `\boxed{True/False}`, Attention: You MUST put your final answer using the following format: Final answer: `\boxed{True/False}`.

Figure 7: Prompt for Generating Reasoning Paths on TabFact Dataset.

### Reasoning Path Generation prompt on SciTab

Please inspect the table(s) and then provide an Answer to the question. Please reason step by step. Attention: You MUST put your final answer using the following format: Final answer: `\boxed{True/False/not enough info}`, Attention: You MUST put your final answer using the following format: Final answer: `\boxed{True/False/not enough info}`.

Figure 8: Prompt for Generating Reasoning Paths on SciTab Dataset.

### Prompt for generating CoT analysis

The following is the table question answering problem and a solution (split into paragraphs, enclosed with tags and indexed from 1):

[Table Question Answering Problem]

{problem}

[Solution]

{solution section}

Your task is to verify the correctness of the paragraph in the solution. Split your verification by '### Paragraph {ID}'.

Your verification for each paragraph should be constructed by 2 parts, wrapped by '<analyze>' and '<output>' separately.

1. In '<analyze>' part, you need to analyze the reasoning process and explain why the paragraph is correct or incorrect in detail.

2. In '<output>' part, judge if this paragraph is correct or incorrect and put the answer into '<output>' part, e.g., '<output> \*\*Judgement\*\*: `\boxed{Yes/No}` </output>'. Every Paragraph must have an '<output>' part.

Figure 9: Prompt for generating CoT-style rationales.

### Prompt for generating Multi-Dimensional Rationales

The following is the table question answering problem and a solution (split into paragraphs, enclosed with tags and indexed from 1):

[Table Question Answering Problem]

{problem}

[Solution]

{solution Section}

Your task is to verify the logical and factual correctness of each solution paragraph.

Split your verification by '###Paragraph ID'.

Your verification for each paragraph should be constructed by 2 parts, wrapped by '<analyze></analyze>', and '<output></output>' separately.

1. In '<analyze>' part, address these five aspects:

Part 1: **Restatement**: Verbalize the claim, statement, or answer this paragraph makes.

Part 2: **Data Check**: Traverse the table and quote row and column to show where the evidence comes from.

Part 3: **Logical Consistency**: Check if the claim is logically valid.

Part 4: **Numeric Accuracy**: Check if calculations such as sums, percentages, and table lookups, are computed correctly.

Part 5: **Format**: Verify that the final paragraph puts the final answer using  $\boxed{\text{Final Answer}}$ .

If any aspect is wrong, then the paragraph is wrong.

2. In '<output>' part, judge if this paragraph is correct or incorrect and put the answer into '<output>' part, e.g., "<output>**Judgement**:  $\boxed{\text{Yes/No}}$ </output>". Every Paragraph must have an '<output>' part.

Figure 10: Prompt used for generating analyses from five distinct perspective.

### Prompt for generating Rephrase, React, and Analysis

The following is the table question answering problem and a solution (split into paragraphs, enclosed with tags and indexed from 1):

[Table Question Answering Problem]

{problem}

[Solution]

{solution section}

Your task is to verify the correctness of paragraphs in the solution. Split your verification by ‘### Paragraph {ID}’.

Your verification for each paragraph should be constructed by 4 parts, wrapped by ‘<rephrase></rephrase>’, ‘<react></react>’, ‘<analyze></analyze>’ and ‘<output></output>’ separately.

1. In ‘<rephrase>’ part, rephrase the paragraph as a clear, self-contained question, preserving all information from the original paragraph.

2. In ‘<react>’ part, use alternating steps of **Thought**, **Action**, **Observation** to systematically solve the problem from <rephrase> part.

**Thought**: Based on the information currently available, reason through the problem and determine the goal of the next action.

**Action**: Each action must be one of the following five types: 1. LOOKUP\_ROW[row]: Find a row in the table whose label matches the given text. 2. LOOKUP\_COLUMN[column]: Find a column in the table whose header matches the given text. 3. READ\_CELL[row, column]: Retrieve the value in the specified row and column. 4. COMPUTATION[values]: Perform arithmetic or other clearly defined mathematical operations on the provided values. 5. Finish[answer]: Once a clear answer has been determined, use this action to return the answer and terminate the task.

**Observation**: Record the factual result of the action.

3. In ‘<analyze>’ part, based on ‘<rephrase>’ and ‘<react>’ results, analyze in detail if the current paragraph is correct or incorrect, citing the relevant evidence.

4. In ‘<output>’ part, make a final judgement and put the judgement into ‘<output>’ part, e.g., ‘<output> **Judgement**:  $\boxed{\text{Yes/No}}$  </output>’.

Figure 11: Prompt for generating Rephrase, React, and Analysis rationales.

### Prompt for generating analysis with code

The following is the table question answering problem and a solution (split into paragraphs, enclosed with tags and indexed from 1):

[Table Question Answering Problem]

{problem}

[Solution]

{solution section}

Your task is to verify the correctness of paragraph in the solution. Split your verification by '### Paragraph {ID}'.

Your verification for each paragraph should be constructed by 3 parts, wrapped by '<analyze></analyze>', '<verify></verify>' and '<output></output>' separately.

1. In '<analyze>' part, you need to analyze the reasoning process and explain why the paragraph is correct or incorrect in detail.
2. In '<verify>' part, you must write **Python code** in the form of `python {CODE}` to verify every details in the current paragraph that can be verified by code. You must use python code to verify every details in the paragraph. Make sure to print the critic results in the code. Every code will be executed automatically by system. You need to analyze the '[Code Output]' after code executing. Pay attention that you must follow the format of `python{CODE}` when you write the code, otherwise the code will not be executed.
3. In '<output>' part, judge if this paragraph is correct or incorrect and put the answer into '<output>' part, e.g., `<output> Judgement: \boxed{Yes/No} </output>`. Every Paragraph must have an '<output>' part.

Figure 12: Prompt for generating analysis with code execution.

### Prompt for LLM-as-a-Judge

The following is the table question answering problem and a solution (split into paragraphs, enclosed with tags and indexed from 1):

[Table Question Answering Problem]

{problem}

[Solution]

{solution section}

Your task is to verify the correctness of paragraph in the solution. Split your verification by '### Paragraph {ID}'.

Your verification for each paragraph should be constructed by 2 parts, wrapped by '<analyze></analyze>' and '<output></output>' separately.

1. In '<analyze>' part, you need to analyze the reasoning process and explain why the paragraph is correct or incorrect in detail.
2. In '<output>' part, judge if this paragraph is correct or incorrect and put the answer into '<output>' part, e.g., `<output> Judgement: $\boxed{Yes/No} </output>`. Every Paragraph must have an '<output>' part.

Figure 13: Prompt for step-wise LLM-as-a-Judge.