

IF-CRITIC: Towards a Fine-Grained LLM Critic for Instruction-Following Evaluation

Bosi Wen^{1,†,*} Yilin Niu^{2,*} Cunxiang Wang² Pei Ke³ Xiaoying Ling²
Ying Zhang² Aohan Zeng⁴ Hongning Wang¹ Minlie Huang^{1,‡}

¹The Conversational Artificial Intelligence (CoAI) Group, Tsinghua University

²Zhipu AI ³University of Electronic Science and Technology of China

⁴The Knowledge Engineering Group (KEG), Tsinghua University

wbs23@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Instruction-following is a fundamental ability of Large Language Models (LLMs), requiring their generated outputs to follow multiple constraints imposed in input instructions. Numerous studies have attempted to enhance this ability through preference optimization or reinforcement learning based on reward signals from LLM-as-a-Judge. However, existing evaluation models for instruction-following still possess many deficiencies, such as substantial costs and unreliable assessments. To this end, we propose IF-CRITIC, an LLM critic for fine-grained, efficient, and reliable instruction-following evaluation. We first develop a checklist generator to decompose instructions and generate constraint checklists. With the assistance of the checklists, we collect high-quality critique training data through a multi-stage critique filtering mechanism and employ a constraint-level preference optimization method to train IF-CRITIC. Extensive experiments show that the evaluation performance of IF-CRITIC can beat strong LLM-as-a-Judge baselines, including o4-mini and Gemini-3-Pro. With the reward signals provided by IF-CRITIC, LLMs can achieve substantial performance gains in instruction-following optimization under lower computational overhead compared to strong LLM critic baselines. Our code and model are available at <https://github.com/thu-coai/IF-CRITIC>.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in various NLP tasks (Zhao et al., 2023). Among these, instruction-following stands out as one of the most crucial requirements for LLM applications (Ouyang et al., 2022). In

[†]Work done when this author interned at Zhipu AI.

^{*}Equal contribution

[‡]Corresponding author

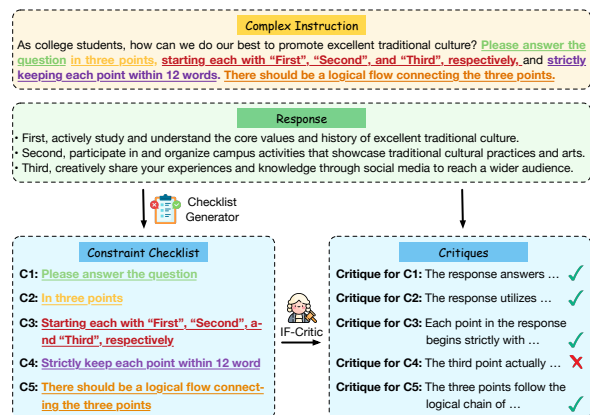


Figure 1: A usage example of IF-CRITIC: Given an instruction and a response, a *checklist generator* first decomposes the instruction to generate a constraint checklist. Then, IF-CRITIC can provide fine-grained evaluations for the response with respect to its compliance with all included constraints in one inference pass.

the real-world use of LLMs, nearly all tasks are formulated as instruction-following, where human instructions specify task requirements and impose corresponding constraints on the model output (Jiang et al., 2024). Accurately following instructions serves as a key factor in ensuring the helpfulness and reliability of LLMs (Huang et al., 2024).

However, LLMs still fall short in instruction-following, especially when dealing with complex instructions that involve numerous constraints (Jiang et al., 2024). Many research attempts have been made to improve the instruction-following ability of LLMs. They often leverage LLMs to synthesize complex instructions and employ LLM-as-a-Judge to evaluate compliance with each constraint, thereby deriving reward signals to perform preference optimization or reinforcement learning (Zhang et al., 2025; He et al., 2024a; Ren et al., 2025; Liu et al., 2025c; Peng et al., 2025a).

Despite these efforts, the usage of LLM-as-a-Judge for instruction-following still faces two major yet underexplored challenges: (1) **Substantial**

Costs: Existing methods typically rely on proprietary LLMs (e.g., GPT-4o) or large reasoning models (e.g., QwQ-32B) to judge each constraint in an instruction separately and use the resulting judgments as reward signals. Since complex instructions often involve multiple constraints, these methods require repeated per-constraint inference, which is costly and difficult to scale up in model training. (2) **Unreliable Assessments:** Current LLMs still struggle to accurately assess instruction-following (Zhang et al., 2024c), exhibiting low recall in error detection (Kamoi et al., 2024) and limited counting ability (Zhang et al., 2024b; Fu et al., 2024). These limitations introduce noise to the evaluation and restrict the gains in model optimization. To tackle these challenges, some works attempt to introduce handcrafted code-verifiable constraints in data construction, aiming to enhance evaluation efficiency and reliability (Dong et al., 2025; Kim et al., 2025; Li et al., 2025; Peng et al., 2025a). However, these constraints are often homogeneous and atomized, failing to cover the diversity and complexity of human instructions, such as constraint composition (Wen et al., 2024).

In this work, we propose IF-CRITIC, an LLM critic tailored for instruction-following evaluation. To enhance inference efficiency and obtain a more holistic and granular perception of the instructions, IF-CRITIC adopts a novel **checklist-guided critique generation paradigm**. As shown in Figure 1, we first utilize a *checklist generator* to decompose the instructions and generate constraint checklists. Then, IF-CRITIC can evaluate compliance with all included constraints in one inference pass, eliminating the need to generate separate judgments for different constraints multiple times. To train IF-CRITIC, we elaborately prompt Deepseek-R1 (Guo et al., 2025a) to generate expert critiques guided by the checklists, then employ a **multi-stage critique filtering mechanism** to select the highest-quality critique for each constraint, which significantly mitigates noise and enhances the reliability of critiques, including those involving counting. After fine-tuning on this data, we leverage the property that the critiques contain multi-constraint evaluation results, further improving IF-CRITIC via a novel **constraint-level preference optimization method**. This method performs fine-grained, constraint-level comparisons between positive and negative critiques using Direct Preference Optimization (DPO) (Rafailov et al., 2023), reinforcing the perception of crucial preference in-

formation and leading to better alignment between IF-CRITIC and the expert critiques. Empirically, IF-CRITIC consistently outperforms various LLM-as-a-Judge baselines on four instruction-following meta-evaluation benchmarks, including the latest state-of-the-art LLMs o4-mini (Jaech et al., 2024) and Gemini-3-Pro (Team et al., 2023).

Leveraging the fine-grained evaluation results from IF-CRITIC as reward signals, we adapt DPO and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to improve the instruction-following ability of LLMs. Experiments demonstrate that IF-CRITIC brings substantially greater improvements over strong baselines with significantly less computational overhead, highlighting its effectiveness and practicability. Our contributions can be summarized as follows:

- We propose a novel framework for the development of instruction-following critics, including the checklist-guided critique generation paradigm, the multi-stage critique filtering mechanism for high-quality training data collection, and the constraint-level preference optimization method for model training.
- We introduce IF-CRITIC, the first instruction-following critic developed through our framework. Experiments show that IF-CRITIC achieves superior evaluation performance compared to strong LLM-as-a-Judge baselines, including o4-mini and Gemini-3-Pro.
- We validate that IF-CRITIC can provide scalable reward signals to improve the instruction-following abilities of LLMs using DPO and GRPO. IF-CRITIC can bring substantially greater improvements under lower computational overhead compared to strong baselines.

2 Related Work

LLM-based Evaluation. Utilizing LLMs as text evaluators has gradually become prevalent (Wang et al., 2023a; Chen et al., 2023; Ke et al., 2023) due to their advantages in flexibility, interpretability, and generalization. Although proprietary LLMs have demonstrated excellent evaluation capabilities (Zheng et al., 2023; Wang et al., 2023a), their practical application is limited by issues such as high cost and potential data leakage (Ke et al., 2024). To reduce the reliance on proprietary LLMs, a series of works attempt to train smaller LLMs specialized in evaluation using the data generated from

humans or proprietary LLMs (Kim et al., 2024a; Li et al., 2024; Ke et al., 2024; Hu et al., 2024; Kim et al., 2024b). Distinct from these models that focus on evaluating overall text quality, we develop an evaluation model that provides constraint-level, fine-grained assessments of instruction-following and leverages them as reward signals to enhance the instruction-following ability of LLMs.

Checklist-based Evaluation. Recently, using checklists for evaluation has emerged as a promising paradigm for reward modeling and has been widely applied across diverse domains, including instruction following (Peng et al., 2025b; Liu et al., 2025c; Qin et al., 2025b; Viswanathan et al., 2025; Huang et al., 2025b), reasoning (Zhou et al., 2025; Sanders et al., 2026), and long-form generation (Gunjal et al., 2025; Shao et al., 2025). This paradigm associates each instruction with a list of verifiable checklist items and derives the reward for a response based on the proportion of satisfied items. Through fine-grained verification of each item to calculate the final reward, this approach alleviates the problem of overly coarse supervision (Liu et al., 2025b) and improves evaluation robustness (Shen et al., 2026). However, current work typically relies on proprietary LLMs or large reasoning models to separately verify each checklist item, which is costly and difficult to scale. To address this scaling bottleneck, we develop a fine-grained LLM critic to efficiently provide reliable verification results for checklist items.

Instruction-following Optimization. As LLMs are increasingly applied to real-world complex tasks, the ability to follow complex instructions with multiple constraints emerges as a key determinant of their practical utility (Liu et al., 2023; Lou et al., 2024), driving extensive studies to enhance this ability of LLMs. They typically leverage LLMs to synthesize complex instructions and perform preference optimization or reinforcement learning based on the reward signals from LLM-as-a-Judge (Sun et al., 2024; Zhang et al., 2025; Cheng et al., 2025; Ren et al., 2025; Liu et al., 2025c; Peng et al., 2025a). Specifically, SPaR (Cheng et al., 2025) employs self-play tree search refinement through a fine-tuned refiner to construct DPO training data. RECAST (Liu et al., 2025c) categorizes soft and hard constraints, leveraging GPT-4o and code execution, respectively, to obtain evaluation results for GRPO training. However, the effectiveness of existing methods is restricted

by the high costs and limited instruction-following evaluation ability of LLM-as-a-Judge, and code-verifiable constraints fail to cover the diversity and complexity of human instructions.

3 Methodology

3.1 Task Definition and Method Overview

Distinct from previous LLM critics that typically generate evaluation results for a single criterion or question (Vu et al., 2024; Wang et al., 2025; Anugraha et al., 2025), IF-CRITIC leverages a constraint checklist to guide critique generation and evaluates compliance with all included constraints in one inference pass. Specifically, given an instruction x , corresponding response y , and a constraint checklist $\{c_k\}_{k=1}^n$ which contains all constraints within x , IF-CRITIC can provide a critique $C = \bigcup_{k=1}^n (e_k, j_k)$ for y , where each segment (e_k, j_k) consists of an explanation e_k and a binary judgment j_k (0 or 1) for compliance with the respective constraint. This paradigm not only enhances inference efficiency but also equips IF-CRITIC with a more holistic and granular perception of the instructions, thus improving the reliability of instruction-following assessments.

The development of IF-CRITIC consists of the following steps: After collecting complex user instructions and LLM-generated responses, we first develop a checklist generator to decompose the instructions and generate constraint checklists. With the assistance of the checklists, we elaborately prompt Deepseek-R1 to generate expert critiques and employ a multi-stage critique filtering mechanism to further enhance their quality. Finally, we utilize the collected expert critiques and perform a constraint-level performance optimization method to train IF-CRITIC. The overall pipeline is illustrated in Figure 2. IF-CRITIC can be leveraged to provide reliable reward signals, improving the instruction-following ability of LLMs via DPO or GRPO, as validated in our experiments.

3.2 Instruction and Response Collection

To obtain diverse instructions for critic training, we collect instructions from real-world application scenarios and utilize LLMs to categorize them according to the task taxonomy of CritiqueLLM (Ke et al., 2024), which encompasses 10 categories and covers diverse NLP applications in real-world scenarios. To improve data quality, we prompt LLMs to score instruction quality and develop a classi-

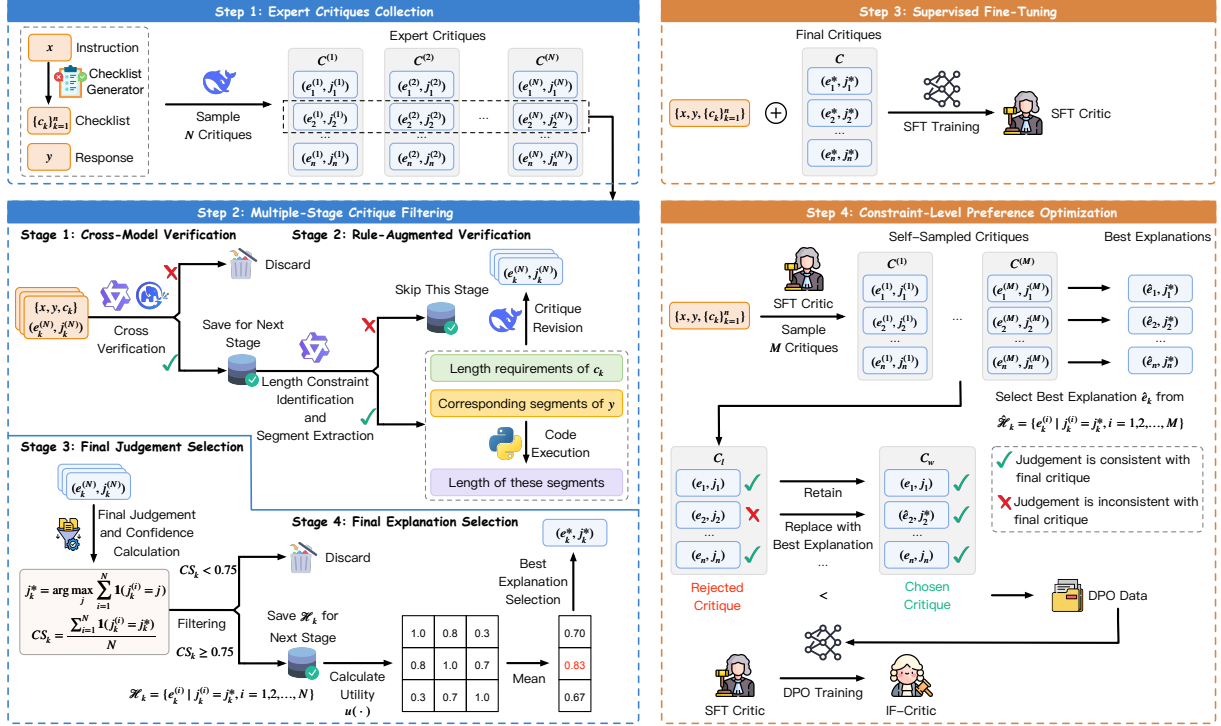


Figure 2: The pipeline of IF-CRITIC development. The left section illustrates the process of critique training data construction, while the right section presents the process of training IF-CRITIC.

fier to assess their constraint complexity, which is trained with a small amount of human annotation data. Finally, we balance the amount of instruction across task categories and acquire 55K high-quality instructions with relatively high complexity.

Then, we collect LLM responses from 15 representative models, which possess varying abilities in instruction-following. The model list is provided in Appendix A. For each instruction, we randomly select two models to generate responses, resulting in a total of 110k evaluation inputs, each comprising an instruction and its corresponding LLM response.

3.3 Critique Training Data Construction

To assist IF-CRITIC in fine-grained evaluation, we first develop a checklist generator to decompose the constraints within instruction x and generate a constraint checklist $\{c_k\}_{k=1}^n$. To achieve this, we leverage a powerful LLM, Deepseek-R1¹ for automatic checklist annotation in our collected instructions, and then utilize these checklists to fine-tune a base model as our checklist generator, which supports efficient checklist generation after deployment. Manual inspection of 1,000 test samples for the generator shows that 99.29% of generated constraints and 97.50% of entire checklists are correct,

¹<https://huggingface.co/deepseek-ai/DeepSeek-R1>

indicating its reliability.

With the assistance of the checklists provided by our checklist generator, we proceed to employ Deepseek-R1 for expert critique annotation via a carefully designed prompt, which elicits a concise and specific explanation before the judgment for each constraint. Detailed prompts are provided in the Appendix B. For each response, we collect N expert critiques and apply a multi-stage critique filtering mechanism to select the highest-quality critique for each constraint in the checklist, which is then used for training IF-CRITIC.

The mechanism comprises four stages: cross-model verification, rule-augmented verification, final judgement selection, and final explanation selection. The first two stages aim to filter out noise in critiques caused by two major limitations of LLM-as-a-Judge: potential evaluation bias and limited counting ability, respectively. The remaining two stages then select the highest-quality judgements and explanations from the filtered critiques based on consistency, further enhancing their reliability.

Cross-Model Verification. Previous works point out that LLM-as-a-Judge may suffer from potential biases (Zheng et al., 2023; Ye et al., 2025a), undermining the reliability of its generated critiques. To mitigate this, we utilize two additional LLMs,

GLM-4-Plus (GLM et al., 2024) and Qwen2.5-72B-Instruct (Yang et al., 2024), to independently verify two aspects of the expert critiques for each constraint: (1) The correctness of the explanation, and (2) The consistency between the explanation and judgment. Detailed prompts can be found in Appendix B. Any critiques that fail verification by any model in any aspect are strictly filtered, accounting for 11.3% of our data.

Rule-Augmented Verification. Considering that LLMs often struggle with accurate counting (Zhang et al., 2024b; Fu et al., 2024), we introduce rule-augmented verification to enhance the reliability of critiques involving length constraints (e.g., *each title must be 8 characters long*). Methodologically, we first prompt Qwen2.5-72B-Instruct to identify length constraints and extract relevant segments from the responses. Subsequently, we employ Python scripts to identify length information within these segments automatically. This information then guides Deepseek-R1 to revise its critiques. The prompts for these steps are in the Appendix B. This approach integrates the flexibility of LLM-based evaluation with the precision of rule-based evaluation, thereby effectively handling the diverse length constraints in human instructions.

Final Judgement Selection. Given that critique generation is analogous to a Chain-of-Thought (CoT) (Wei et al., 2022) reasoning process, we incorporate self-consistency (Wang et al., 2023b), a widely used decoding strategy in CoT reasoning, to select final judgments and improve their correctness. Specifically, for each constraint c_k , we select the final judgment j_k^* via majority voting among multiple expert critiques. Furthermore, previous works show that *confidence*, which is calculated as the proportion of the majority result to the total samples, is an important indicator of answer reliability (Xiong et al., 2024; Hu et al., 2024; Prasad et al., 2024), with higher confidence generally implying greater reliability. Inspired by this, we discard the final critiques of constraints with confidence lower than 0.75 in our experiment, which empirically balances data retention and reliability.

Final Explanation Selection. After obtaining the correct final judgments, the last step is to select the highest-quality explanations that support these judgments. To achieve this, we devise a selection strategy inspired by Minimum Bayes Risk Decoding (Bickel and Doksum, 2015): For each

constraint c_k , we first construct a hypothesis set \mathcal{H}_k containing all explanations whose corresponding judgments align with the final judgments j_k^* . Then, we utilize a text similarity measure $u(\cdot)$ to calculate the consistency between each pair of explanations in \mathcal{H}_k . The explanation with the highest average consistency among others is selected as the final explanation e_k^* . The formal definition is as follows:

$$\mathcal{H}_k = \{e_k^{(i)} \mid j_k^{(i)} = j_k^*, i = 1, 2, \dots, N\} \quad (1)$$

$$e_k^* = \arg \max_{e \in \mathcal{H}_k} \frac{1}{|\mathcal{H}_k|} \sum_{\tilde{e} \in \mathcal{H}_k} u(\tilde{e}, e) \quad (2)$$

Through these rigorous filtering processes, we manually inspect the quality of 70 final critiques with 353 constraints and find that 96.03% judgments and 92.35% explanations are completely correct, demonstrating their high quality.

3.4 IF-CRITIC Training

After constructing the final critique training dataset $D = \{C_i = \bigcup_{k=1}^{n_i} (e_{ik}^*, j_{ik}^*)\}_{i=1}^{|D|}$, we partition D into two subsets $D_{\text{sft}} \cup D_{\text{ref}}$ and employ a two-stage paradigm to train IF-CRITIC, comprising supervised fine-tuning (SFT) and constraint-level preference optimization. In the first stage, we utilize a critique generation prompt p to concatenate the instruction x , response y , and checklist $\{c_k\}_{k=1}^n$, applying a standard SFT loss on vanilla LLM θ :

$$\mathcal{L}_{\text{SFT}} = - \sum_{i=1}^{|D_{\text{sft}}|} \log P_{\theta}(C_i | p_i) \quad (3)$$

After obtaining the SFT critic, we further enhance its ability via preference optimization. Existing methods usually sample multiple independent responses from the policy model and utilize response-level rewards to select chosen and rejected examples (Yuan et al., 2024; Ye et al., 2025b; Yu et al., 2025). However, since the critiques in our task involve multiple segments for different constraints, only the segments where chosen and rejected critiques differ in judgments encapsulate crucial preference information. Other segments may introduce unnecessary interference and dilute important supervision signals, leading to insufficient optimization (Chen et al., 2024; Cheng et al., 2025).

To mitigate this, we introduce fine-grained, constraint-level comparisons in preference pairs construction, thus reinforcing the perception of crucial preference information. Specifically, for each evaluation input p in D_{ref} , we first sample

M critiques from the SFT critic. Then, we identify the best self-sampled explanation \hat{e}_k for each constraint c_k whose judgements align with the expert final judgement j_k^* in a manner analogous to Equation 1 and 2 (if any). Finally, self-sampled critiques with at least one judgment that misaligns with the expert final critique are selected as rejected critiques $C_l = \bigcup_{k=1}^n (e_k, j_k)$. Retaining the segments that align with the expert final critique while replacing other misaligned segments with the best self-sampled explanations \hat{e}_k and expert final judgement j_k^* to construct chosen critiques C_w :

$$C_w = \bigcup_{k=1}^n \begin{cases} (e_k, j_k), & j_k = j_k^* \\ (\hat{e}_k, j_k^*), & j_k \neq j_k^* \end{cases} \quad (4)$$

Rejected critiques are discarded when some misaligned segments lack substitute best self-sampled explanations. Consequently, we construct a preference dataset D_{dpo} where the preference pairs only differ in the segments with inconsistent judgments, thereby reinforcing crucial preference information. Moreover, self-sampled critiques are closer to the SFT critic decoding space than the expert critique, which is generally beneficial for optimization (Guo et al., 2024). Finally, we conduct preference optimization via the following DPO loss:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = & - \mathbb{E}_{(p, C_l, C_w) \sim D_{\text{dpo}}} [\log \sigma(\beta \log \frac{\pi_\theta(C_w|p)}{\pi_{\text{ref}}(C_w|p)} \\ & - \beta \log \frac{\pi_\theta(C_l|p)}{\pi_{\text{ref}}(C_l|p)})] \end{aligned} \quad (5)$$

where both π_θ (i.e., the trained policy model) and π_{ref} (i.e., the fixed reference model) are initialized from the SFT critic.

3.5 Evaluation Results Utilization

Leveraging the evaluation results of IF-CRITIC as reward signals, we can apply DPO or GRPO algorithms to improve the instruction-following abilities of LLMs. For every input instruction x , the checklist generator produces the constraint checklist $\{c_k\}_{k=1}^n$, and the policy model π_θ samples K candidate responses $\{y_i\}_{i=1}^K$. Then, the reward r_i for each response y_i can be determined by the generated critique $C_i = \bigcup_{k=1}^n (e_{ik}, j_{ik})$ for y_i :

$$r_i = \frac{1}{n} \sum_{k=1}^n j_{ik} \quad (6)$$

Then, we can use this reward to construct DPO training data or to conduct GRPO training.

4 Experiments

4.1 Instruction-Following Evaluation

Evaluation Benchmarks. Given the scarcity of existing instruction-following meta-evaluation datasets, we sample instructions from four benchmarks: **EvalCritic**, **CFBench** (Zhang et al., 2024a), **TRACE** (Zhang et al., 2025), and **Multi-IF** (He et al., 2024b) to construct meta-evaluation datasets. EvalCritic is a random split of the dataset we used for instruction-following optimization (§4.2), whose distribution of instructions differs from IF-CRITIC training data. For the first three benchmarks, we randomly sample 150 instructions. For Multi-IF, we select the first round from the Chinese version, comprising 454 instructions. Responses for each instruction are generated by 2 of 6 LLMs with varying capabilities. We use our checklist generator to generate constraint checklists for EvalCritic and directly adapt the original checklists for other benchmarks. For the first three benchmarks, two annotators independently judge compliance with each constraint, with a third inspector conducting cross-validation. Any discrepancies will be discussed to reach a consensus. For Multi-IF, we use the provided verification code for evaluation. Across these benchmarks, there are 1411 / 1070 / 1634 / 1220 constraints, respectively, of which 406 / 252 / 297 / 320 are labeled as "not followed".

Baselines. We choose state-of-the-art general and evaluation-specific LLMs as our baselines, which are listed in Appendix C. General LLMs can serve as evaluators for instruction-following via elaborately prompting. We explore two prompting strategies: (1) **Checklist-Level Prompt**: evaluate all constraints in the given checklist within a single inference pass, which is the same as IF-CRITIC. (2) **Constraint-Level Prompt**: evaluate each constraint individually. Evaluation-specific LLMs are mainly designed for pairwise comparison. To accommodate this evaluation setting, we convert the above datasets into pairwise versions. Specifically, for two responses generated from the same instruction, we select pairs in which one response follows all constraints while the other does not to construct pairwise samples. The rewards of IF-CRITIC according to Equation 6 for both responses are converted into pairwise comparisons to calculate the agreement rate. To ensure a fair comparison, we remove samples that result in a tie for IF-CRITIC. Greedy search decoding is used for reproducibility,

| Prompt | Model | EvalCritic | | CFBench | | TRACE | | Multi-IF | | Avg. |
|----------------------|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Average F1 |
| Constraint-Level | Gemini-3-Pro | 0.887 | 0.744 | 0.934 | <u>0.794</u> | <u>0.932</u> | <u>0.704</u> | <u>0.962</u> | <u>0.898</u> | <u>0.857</u> |
| | o4-mini | 0.887 | 0.717 | 0.927 | 0.755 | 0.923 | 0.599 | 0.960 | 0.890 | 0.832 |
| | GPT-4.1 | 0.883 | 0.608 | 0.919 | 0.689 | 0.924 | 0.597 | 0.932 | 0.801 | 0.794 |
| | Deepseek-R1 | 0.863 | 0.686 | 0.905 | 0.718 | 0.883 | 0.555 | 0.933 | 0.824 | 0.796 |
| | QwQ-32B | 0.856 | 0.675 | 0.900 | 0.684 | 0.901 | 0.595 | 0.931 | 0.795 | 0.792 |
| | Qwen-2.5-72B-Instruct | 0.854 | 0.362 | 0.891 | 0.459 | 0.915 | 0.374 | 0.896 | 0.590 | 0.668 |
| Checklist-Level | Llama-3.3-70B-Instruct | 0.869 | 0.486 | 0.904 | 0.612 | 0.916 | 0.426 | 0.921 | 0.756 | 0.736 |
| | GLM-4-32B-0414 | 0.867 | 0.451 | 0.900 | 0.531 | 0.913 | 0.384 | 0.898 | 0.610 | 0.694 |
| | Gemini-3-Pro | 0.902 | 0.742 | 0.945 | 0.808 | <u>0.932</u> | 0.655 | 0.960 | 0.892 | 0.855 |
| | o4-mini | <u>0.911</u> | <u>0.753</u> | 0.935 | 0.761 | 0.925 | 0.639 | 0.964 | 0.900 | 0.849 |
| | GPT-4.1 | 0.883 | 0.560 | 0.918 | 0.637 | 0.923 | 0.517 | 0.933 | 0.799 | 0.771 |
| | Deepseek-R1 | 0.895 | 0.716 | 0.928 | 0.725 | 0.910 | 0.579 | 0.938 | 0.828 | 0.815 |
| IF-CRITIC-14B (Ours) | QwQ-32B | 0.892 | 0.663 | 0.922 | 0.716 | 0.922 | 0.570 | 0.933 | 0.793 | 0.801 |
| | Qwen-2.5-72B-Instruct | 0.849 | 0.278 | 0.894 | 0.424 | 0.914 | 0.330 | 0.889 | 0.548 | 0.641 |
| | Llama-3.3-70B-Instruct | 0.862 | 0.391 | 0.905 | 0.585 | 0.915 | 0.400 | 0.916 | 0.718 | 0.712 |
| | GLM-4-32B-0414 | 0.855 | 0.347 | 0.897 | 0.512 | 0.909 | 0.325 | 0.909 | 0.664 | 0.677 |
| | IF-CRITIC-14B (Ours) | 0.927 | 0.806 | <u>0.937</u> | 0.785 | 0.942 | 0.739 | 0.945 | 0.844 | 0.866 |

Table 1: Positive F1 and Negative F1 scores of constraint-following verification on the instruction-following evaluation task. The best performance among these models is **bold**, while the second-best performance is underlined.

| Model | EvalCritic | CFBench | TRACE | Multi-IF |
|------------------------------------|--------------|--------------|--------------|--------------|
| Skywork-Reward-V2-Llama-3.1-8B-40M | 0.571 | 0.689 | 0.671 | 0.705 |
| Prometheus-BGB-8x7B | 0.464 | 0.556 | 0.529 | 0.489 |
| RM-R1-Deepseek-Distilled-Qwen-32B | 0.607 | 0.711 | 0.765 | 0.625 |
| RRM-32B | 0.534 | 0.689 | 0.765 | 0.636 |
| IF-CRITIC-14B (Ours) | 0.964 | 0.956 | 0.882 | 0.977 |

Table 2: Pairwise agreement rates on the instruction-following evaluation task.

which has a negligible impact on the evaluation performance of reasoning models, as verified in Appendix E.

Implementation Details. We choose Qwen-2.5-14B-Instruct (Yang et al., 2024) as the base model for both the checklist generator and IF-CRITIC. We also experiment with other base models, and the results are provided in Appendix F. Deepseek-R1 generates $N = 5$ expert critiques per response. During final explanation selection, we use `diff1ab`² to implement the text similarity measure $u(\cdot)$. During IF-CRITIC training, the ratio of D_{sft} to D_{ref} is set to 6:4. The number of self-sampled critiques, M , is set to 10, and up to 1 preference pair is constructed for each evaluation input. More implementation details are in Appendix D.

Main Results. Table 1 presents the F1-scores for both positive and negative classes in the instruction-following evaluation task of general LLMs, while Table 2 shows the pairwise agreement for evaluation-specific LLMs. **Firstly**, IF-CRITIC demonstrates strong performance on various benchmarks, outperforming all baselines on av-

erage F1, including the latest state-of-the-art LLM o4-mini and Gemini-3-Pro. In contrast, the performance of existing general LLMs remains limited, with relatively low negative F1. This aligns with our motivation for developing IF-CRITIC to enable more reliable evaluation. **Secondly**, existing evaluation-specific LLMs still perform poorly in evaluating instruction-following and lag far behind IF-CRITIC, highlighting the necessity of a critic specifically designed for this task. **Finally**, reasoning models (e.g. o4-mini, DeepseekR1, and QwQ-32B) typically perform better with **Checklist-Level Prompt**, while other LLMs benefit more from **Constraint-Level Prompt**. We attribute this to two factors: (1) Evaluating multiple constraints at one inference pass increases task complexity. (2) Long-chain reasoning enables a more holistic perception of the relationships between constraints with the help of checklists. This observation validates the rationale of our checklist-guided critique generation paradigm.

Ablation Study. To investigate the impact of each design component of our framework, we conduct additional ablation studies. For the training data format, we decompose expert critiques to train an alternative model that generates critiques for individual constraints. Table 3 shows that the performance of IF-CRITIC declines across all benchmarks, demonstrating that checklist-guided training can facilitate instruction perception and contribute to the final performance.

For the critique filtering mechanism, we remove

²<https://docs.python.org/3/library/diff1ab.html>

| Method | EvalCritic | CFBench | TRACE | Multi-IF |
|---------------------------------|--------------|--------------|--------------|--------------|
| IF-CRITIC-14B | 0.861 | 0.863 | 0.840 | 0.895 |
| <i>Data Format</i> | | | | |
| w/ Constraint-Level Critique | 0.844 | 0.830 | 0.816 | 0.859 |
| <i>Critique Filtering</i> | | | | |
| w/ Raw Data | 0.814 | 0.792 | 0.774 | 0.780 |
| w/o Cross-Model Verification | 0.851 | 0.858 | 0.832 | 0.874 |
| w/o Rule-Augmented Verification | 0.827 | 0.823 | 0.789 | 0.825 |
| w/o Final Judgement Selection | 0.840 | 0.804 | 0.821 | 0.849 |
| w/o Final Explanation Selection | 0.840 | 0.846 | 0.807 | 0.858 |
| <i>Preference Learning</i> | | | | |
| w/ Vanilla DPO | 0.797 | 0.797 | 0.785 | 0.841 |
| w/ Expert Critique | 0.828 | 0.836 | 0.801 | 0.840 |
| w/o Preference Learning | 0.815 | 0.810 | 0.810 | 0.841 |

Table 3: The average of Positive F1 and Negative F1 scores under different ablation settings for IF-CRITIC.

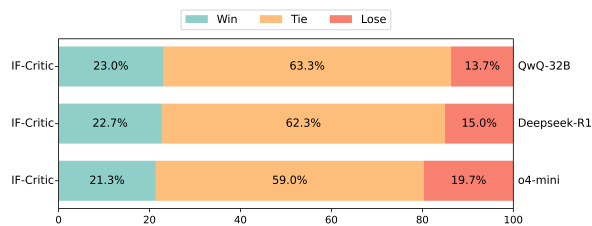


Figure 3: Explanation quality evaluation results. The percentages indicate the preference between IF-CRITIC and other evaluation models via human annotation.

each step individually to explore its impact on evaluation performance. Table 3 shows that all steps contribute to the final performance, but removing rule-augmented verification yields the largest performance drop, underscoring its key role in improving discriminability for length constraints. Notably, directly using all raw training data without critique filtering leads to the most severe performance decline, highlighting its effectiveness.

For the preference optimization method, we explored three variants of our method: (1) Constructing preference pairs directly based on the agreement rates between expert critiques, denoted as Vanilla DPO, (2) Replacing misaligned segments of rejected critiques with expert critiques rather than the best self-sampled ones, and (3) Performing SFT on the entire training dataset without preference optimization. Table 3 shows that our constraint-level preference optimization method achieves the best performance, evidencing its effectiveness in enabling better alignment with expert critiques.

Analysis on Explanation Quality. To assess the quality of the explanations generated by IF-CRITIC, we randomly sample 300 constraints from the first three evaluation benchmarks and collect corresponding generated explanations from IF-CRITIC, QwQ-32B, Deepseek-R1, and o4-mini.

For each pair of explanations (one from IF-CRITIC and the other from a baseline evaluation model, given the same evaluation input), we ask human annotators to judge which explanation is better (i.e., win, lose, or tie) in terms of correctness, informativeness, and helpfulness. The priority of these aspects follows the above order. As shown in Figure 3, IF-CRITIC surpasses QwQ-32B and Deepseek-R1 in win rates by 9.3% and 7.7%, respectively, and performs on par with o4-mini. Many results are ties because all four evaluators are strong and correct on many constraints. This demonstrates that IF-CRITIC can generate high-quality explanations for instruction-following evaluation.

4.2 Instruction-Following Optimization

Training Datasets & Settings. To collect diverse and high-quality complex instructions for model training, we first curate seed instructions from real-world application scenarios. These scenarios are distinct from the training data of IF-CRITIC and generally exhibit lower instruction complexity. Stratified sampling based on task categories is adopted to enhance diversity. Subsequently, we employ Deepseek-R1 to introduce constraints to these instructions based on a comprehensive constraint taxonomy that encompasses 6 primary constraint categories and 4 constraint composition types. Deepseek-R1 is also used to validate the generated instructions, ensuring they are reasonable, unambiguous, and logically consistent. In total, we obtain 18.5k high-quality complex instructions for training. Further details are in Appendix G.

We choose Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct as the initial policy models for DPO and GRPO training. Three evaluation models are used to provide reward signals: IF-CRITIC, the state-of-the-art reward model Skywork-Reward-V2-Llama-3.1-8B-40M, and the powerful open-source general LLM QwQ-32B. For DPO training, we sample 10 responses per instruction, with a temperature of 1.0 and a top-p value of 0.9. The responses with the highest and lowest rewards are used to construct preference pairs. For GRPO training, we perform 32 rollouts per instruction. More training details are in Appendix G.

Evaluation Benchmarks. We evaluate our models on three popular instruction-following benchmarks: **Multi-IF** (He et al., 2024b), which assesses multi-turn and multilingual hard constraints following; **CFBench** (Zhang et al., 2024a), which covers

| Policy Model | Method | Reward Model / Critic | Relative Time | Multi-IF | | | CFBench | | | SysBench | | |
|-----------------------|--------|------------------------------------|---------------|--------------|--------------|--------------|-------------|-------------|-------------|--------------|--------------|--------------|
| | | | | Turn 1 | Turn 2 | Turn 3 | CSR | ISR | PSR | CSR | ISR | SSR |
| Qwen2.5-7B-Instruct | - | - | - | 76.14 | 61.65 | 51.41 | 0.79 | 0.46 | 0.56 | 60.33 | 35.72 | 19.10 |
| Qwen2.5-32B-Instruct | - | - | - | 83.02 | 72.43 | 63.12 | 0.84 | 0.55 | 0.65 | 81.76 | 66.44 | 44.83 |
| Qwen2.5-7B-Instruct | DPO | Skywork-Reward-V2-Llama-3.1-8B-40M | 0.79 | 77.86 | 63.12 | 53.15 | 0.82 | 0.53 | 0.63 | 61.59 | 38.60 | 23.60 |
| | | QwQ-32B | 13.4 | 80.44 | 66.05 | 56.82 | 0.82 | 0.50 | 0.61 | 64.85 | 41.32 | 24.23 |
| | | IF-CRITIC-14B (Ours) | 1.00 | 81.25 | 67.51 | 57.73 | 0.83 | 0.52 | 0.63 | 67.40 | 44.84 | 28.71 |
| | GRPO | Skywork-Reward-V2-Llama-3.1-8B-40M | 1.04 | 75.28 | 61.12 | 51.90 | 0.81 | 0.50 | 0.60 | 65.82 | 44.48 | 25.79 |
| | | QwQ-32B | 3.08 | 78.59 | 63.51 | 52.60 | 0.83 | 0.54 | 0.64 | 76.25 | 59.28 | 37.58 |
| | | IF-CRITIC-14B (Ours) | 1.00 | 81.87 | 67.36 | 58.76 | 0.85 | 0.59 | 0.69 | 81.19 | 66.80 | 44.44 |
| LLama3.1-8B-Instruct | - | - | - | 73.13 | 64.27 | 56.84 | 0.70 | 0.32 | 0.43 | 60.14 | 37.32 | 19.38 |
| LLama3.1-70B-Instruct | - | - | - | 83.06 | 74.25 | 66.78 | 0.82 | 0.52 | 0.62 | 78.64 | 62.52 | 39.49 |
| LLama3.1-8B-Instruct | DPO | Skywork-Reward-V2-Llama-3.1-8B-40M | 0.79 | 76.71 | 69.04 | 60.59 | 0.75 | 0.38 | 0.50 | 67.88 | 47.53 | 28.22 |
| | | QwQ-32B | 12.8 | 79.72 | 71.94 | 62.97 | 0.76 | 0.41 | 0.52 | 68.36 | 46.97 | 27.53 |
| | | IF-CRITIC-14B (Ours) | 1.00 | 80.75 | 72.52 | 64.13 | 0.76 | 0.41 | 0.53 | 69.67 | 49.04 | 28.15 |
| | GRPO | Skywork-Reward-V2-Llama-3.1-8B-40M | 1.33 | 75.57 | 67.88 | 58.59 | 0.77 | 0.42 | 0.53 | 76.99 | 59.96 | 37.42 |
| | | QwQ-32B | 3.10 | 80.67 | 72.49 | 64.40 | 0.77 | 0.44 | 0.53 | 78.22 | 62.32 | 40.79 |
| | | IF-CRITIC-14B (Ours) | 1.00 | 83.83 | 73.32 | 65.05 | 0.81 | 0.50 | 0.59 | 79.90 | 66.40 | 44.27 |

Table 4: Experimental results (%) on Multi-IF, CFBench, and SysBench. The best performance among different critics is **bold**. **Relative Time** refers to the relative time to IF-CRITIC of reward calculation in DPO and average per-step training in GRPO. Skywork-Reward-V2 is a reward model that generates scalar rewards, whereas QwQ-32B and IF-CRITIC generate natural language critiques.

a comprehensive range of constraint types; and **Sysbench** (Qin et al., 2025a), which focuses on the following of system prompts. GPT-4o-2024-11-20 is used as the evaluation model for CFBench and SysBench. For **Multi-IF**, we follow RPO (Huang et al., 2025a) and report the average performance across the English and Chinese subsets.

Main Results. The main results are presented in Table 4. **Firstly**, optimization with IF-CRITIC brings substantially superior performance gains across various benchmarks compared to all strong baselines, substantiating its advanced capacity of instruction-following evaluation. Notably, LLMs after GRPO training with IF-CRITIC even achieve comparable performance to significantly larger LLMs (32B or 70B) of the same model family. In contrast, the general reward model, Skywork-Reward-V2, is ill-suited for instruction-following optimization, yielding only marginal improvements. **Secondly**, compared to the powerful LLM critic, QwQ-32B, IF-CRITIC only incurs less than a third of computational overhead, highlighting its practicality. **Finally**, GRPO proves more effective for instruction-following tasks compared to DPO, delivering greater performance gains. The reward curves during GRPO training are provided in the Appendix H. We also validate that optimization with IF-CRITIC does not compromise the general performance of LLMs in the Appendix I.

5 Conclusion

Our work proposes IF-CRITIC, an LLM critic for fine-grained instruction-following evaluation. IF-

CRITIC adopts a checklist-guided critique generation paradigm that evaluates all constraints in a checklist within one inference pass, improving efficiency and reliability. We collect high-quality critique training data through a multi-stage filtering mechanism and employ constraint-level preference optimization to train IF-CRITIC. Extensive experiments show that IF-CRITIC has stronger evaluation ability and can bring superior performance gains to LLMs in instruction-following optimization, compared to powerful LLM critic baselines.

Limitations

The limitations of our work are summarized as follows:

Scope of Rule-Augmented Verification. In the critique training data construction stage of IF-CRITIC, we only introduce rule-augmented verification for length-related constraints, while other code-verifiable constraints (e.g., keywords or structural format) are not comprehensively covered. We prioritize length-related constraints because they are prevalent in real-world complex instructions, and the previous version of IF-CRITIC performs well on other verifiable constraints but struggles with length constraints, as observed in our primary experiments. Although this approach has yielded substantial performance gains, extending this approach to a broader spectrum of constraint types remains an important direction for future work.

Potential Evaluation Bias. Similar to other model-based evaluation methods, IF-CRITIC may also suffer from potential evaluation biases such as

self-enhancement or verbosity bias (Zheng et al., 2023). These biases could harm the evaluation correctness of certain models. Although our multi-stage critique filtering mechanism has effectively mitigated noise and bias in training critique collection, incorporating inference-time strategies (e.g., multi-agent debate) may further attenuate bias and enhance evaluation performance. We reserve further investigation of evaluation bias mitigation as important future work.

Ethical Considerations

In this work, we recruit some human annotators to validate the quality of training data and the performance of IF-CRITIC. The annotator pool primarily consists of college students. Throughout the data annotation process, we adhere to several key principles. Firstly, all annotators are fully informed about the purpose of our study and the involved tasks. Secondly, all annotators received fair compensation for their contributions based on the market price. Finally, any harmful instructions are filtered before annotation to avoid ethical issues.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and the Natural Science Foundation of China (No. 62536008). We would also like to thank Zhipu AI for sponsoring the computational resources and annotation costs in this work.

References

- Anthropic. 2023. [Introducing claude](#).
- David Anugraha, Zilu Tang, Lester James V Miranda, Hanyang Zhao, Mohammad Rifqi Farhansyah, Garry Kuwanto, Derry Wijaya, and Genta Indra Winata. 2025. R3: Robust rubric-agnostic reward models. *arXiv preprint arXiv:2505.13388*.
- Peter J Bickel and Kjell A Doksum. 2015. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Xiuxi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.
- Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.
- Zhipeng Chen, Kun Zhou, Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. 2024. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5694–5711, Bangkok, Thailand. Association for Computational Linguistics.
- Jiale Cheng, Xiao Liu, Cunxiang Wang, Xiaotao Gu, Yida Lu, Dan Zhang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. 2025. SPAR: Self-play with tree-search refinement to improve instruction-following in large language models. In *The Thirteenth International Conference on Learning Representations*.
- Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2025. Self-play with execution feedback: Improving instruction-following capabilities of large language models. In *The Thirteenth International Conference on Learning Representations*.
- Doubao Team. 2025. [Doubao-1.5-pro: Model release](#). Accessed: 2025-01-22.
- Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. 2024. Why do large language models (llms) struggle to count letters? *arXiv preprint arXiv:2412.18626*.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025b. Reward reasoning model. *arXiv preprint arXiv:2505.14674*.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, and 1 others. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.
- Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024a. [From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10864–10882, Miami, Florida, USA. Association for Computational Linguistics.
- Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, and 1 others. 2024b. [Multiif: Benchmarking llms on multi-turn and multilingual instructions following](#). *arXiv preprint arXiv:2410.15553*.
- Zhenyu Hou, Yilin Niu, Zhengxiao Du, Xiaohan Zhang, Xiao Liu, Aohan Zeng, Qinkai Zheng, Minlie Huang, Hongning Wang, Jie Tang, and 1 others. 2024. Chatglm-rlhf: Practices of aligning large language models with human feedback. *arXiv preprint arXiv:2404.00934*.
- Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. [Themis: A reference-free NLG evaluation language model with flexibility and interpretability](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.
- Xiang Huang, Ting-En Lin, Feiteng Fang, Yuchuan Wu, Hangyu Li, Yuzhong Qu, Fei Huang, and Yongbin Li. 2025a. [Reverse preference optimization for complex instruction following](#). *arXiv preprint arXiv:2505.22172*.
- Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, and 1 others. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175.
- Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, and 1 others. 2025b. [Reinforcement learning with rubric anchors](#). *arXiv preprint arXiv:2508.12790*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2023. [Follow-bench: A multi-level fine-grained constraints following benchmark for large language models](#). *arXiv preprint arXiv:2310.20410*.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. [Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.
- Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Haoran Ranran Zhang, Sujeeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. 2024. [Evaluating LLMs at detecting errors in LLM responses](#). In *First Conference on Language Modeling*.
- Pei Ke, Fei Huang, Fei Mi, Yasheng Wang, Qun Liu, Xiaoyan Zhu, and Minlie Huang. 2023. [DecompEval: Evaluating generated texts as unsupervised decomposed question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9676–9691, Toronto, Canada. Association for Computational Linguistics.
- Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. [CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054, Bangkok, Thailand. Association for Computational Linguistics.
- Joongwon Kim, Anirudh Goyal, Aston Zhang, Bo Xiong, Rui Hou, Melanie Kambadur, Dhruv Mahajan, Hannaneh Hajishirzi, and Liang Tan. 2025. [A systematic examination of preference learning through the lens of instruction-following](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*

- (*Volume 1: Long Papers*), pages 11062–11082, Albuquerque, New Mexico. Association for Computational Linguistics.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. [Prometheus 2: An open source language model specialized in evaluating other language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. [Generative judge for evaluating alignment](#). In *The Twelfth International Conference on Learning Representations*.
- Ming Li, Han Chen, Chenguang Wang, Dang Nguyen, Dianqi Li, and Tianyi Zhou. 2025. [RuleR: Improving LLM controllability by rule-based data recycling](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 926–943, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025b. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*.
- Wenhao Liu, Zhengkang Guo, Mingchen Xie, Jingwen Xu, Zisu Huang, Muzhao Tian, Jianhan Xu, Muling Wu, Xiaohua Wang, Changze Lv, and 1 others. 2025c. Recast: Strengthening llms’ complex instruction following with constraint-verifiable data. *arXiv preprint arXiv:2505.19030*.
- Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. [AlignBench: Benchmarking Chinese alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640, Bangkok, Thailand. Association for Computational Linguistics.
- Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*.
- Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.
- OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025a. [VerIF: Verification engineering for reinforcement learning in instruction following](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30324–30339, Suzhou, China. Association for Computational Linguistics.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025b. Verif: Verification engineering for reinforcement learning in instruction following. *arXiv preprint arXiv:2506.09942*.
- Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. 2024. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*.
- Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, sunhaoze, Yan Zhang, Yujing Qiao, weipeng chen, Zenan Zhou, Wentao Zhang, and Bin CUI. 2025a. [Sysbench: Can LLMs follow system message?](#) In *The Thirteenth International Conference on Learning Representations*.

- Yulei Qin, Gang Li, Zongyi Li, Zihan Xu, Yuchen Shi, Zhekai Lin, Xiao Cui, Ke Li, and Xing Sun. 2025b. Incentivizing reasoning for advanced instruction-following of large language models. *arXiv preprint arXiv:2506.01413*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.
- Qingyu Ren, Jie Zeng, Qianyu He, Jiaqing Liang, Yanghua Xiao, Weikang Zhou, Zeye Sun, and Fei Yu. 2025. [Step-by-step mastery: Enhancing soft constraint following ability of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19581–19596, Vienna, Austria. Association for Computational Linguistics.
- Kate Sanders, Nathaniel Weir, Sapana Chaudhary, Kaj Bostrom, and Huzefa Rangwala. 2026. Generating data-driven reasoning rubrics for domain-adaptive reward modeling. *arXiv preprint arXiv:2602.06795*.
- Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, and 1 others. 2025. Dr tulur: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- William F Shen, Xinchu Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. 2026. Rethinking rubric generation for improving llm judge and reward modeling for open-ended tasks. *arXiv preprint arXiv:2602.05125*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Bao-hua Dong, Ran Lin, and Ruohui Huang. 2024. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Qwen Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#).
- Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*.
- Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. [Foundational autoraters: Taming large language models for better automatic evaluation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17086–17105, Miami, Florida, USA. Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- PeiFeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2025. [Direct judgement preference optimization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1979–2009, Suzhou, China. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. [MMLU-pro: A more robust and challenging multi-task language understanding benchmark](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le,

- and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaying Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. [Benchmarking complex instruction-following with multiple constraints composition](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025a. [Justice or prejudice? quantifying biases in LLM-as-a-judge](#). In *The Thirteenth International Conference on Learning Representations*.
- Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun LIU. 2025b. [Learning LLM-as-a-judge for preference alignment](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiayu Yan, Kaidong Yu, and Xuelong Li. 2025. [Improve llm-as-a-judge ability as a general ability](#). *arXiv preprint arXiv:2502.11689*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. [Self-rewarding language models](#). In *Forty-first International Conference on Machine Learning*.
- Tao Zhang, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, Bin Cui, and 1 others. 2024a. [Cfbench: A comprehensive constraints-following benchmark for llms](#). *arXiv preprint arXiv:2408.01122*.
- Xiang Zhang, Juntao Cao, and Chenyu You. 2024b. [Counting ability of large language models and impact of tokenization](#). *arXiv preprint arXiv:2410.19730*.
- Xianren Zhang, Xianfeng Tang, Hui Liu, Zongyu Wu, Qi He, Dongwon Lee, and Suhang Wang. 2024c. [Divide-verify-refine: Aligning llm responses with complex instructions](#). *arXiv preprint arXiv:2410.12207*.
- Xinghua Zhang, Haiyang Yu, Cheng Fu, Fei Huang, and Yongbin Li. 2025. [IOPO: Empowering LLMs with complex instruction following via input-output preference optimization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22185–22200, Vienna, Austria. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*, 1(2).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Sidhartha Brahma, Sujay Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *arXiv preprint arXiv:2311.07911*.
- Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, and 1 others. 2025. [Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning](#). *arXiv preprint arXiv:2508.16949*.

A Model List for Response Generation

We use 15 representative LLMs for response generation, including GPT-4o (OpenAI, 2023), GPT-4o-mini (OpenAI, 2023), seven versions of GLM-4 (GLM et al., 2024), three versions of Qwen2.5 (Yang et al., 2024), Deepseek-R1 (Guo et al., 2025a), Doubao1.5 (Doubao Team, 2025), and Claude3.7-sonnet (Anthropic, 2023).

B List for Prompt Templates

This section lists all the prompt templates applied throughout this work, including the prompt for **scoring the quality of user instruction** in Table 9, the prompt for **constraint checklist generation**

| Model | Decoding Strategy | EvalCritic | | CFBench | | TRACE | | Multi-IF | | Avg. |
|-------------|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Positive F1 | Negative F1 | Average F1 |
| Deepseek-R1 | Greedy Search | 0.895 | 0.716 | 0.928 | 0.725 | 0.910 | 0.579 | 0.938 | 0.828 | 0.815 |
| | Default | 0.901 | 0.702 | 0.933 | 0.730 | 0.915 | 0.565 | 0.936 | 0.829 | 0.814 |
| QwQ-32B | Greedy Search | 0.892 | 0.663 | 0.922 | 0.716 | 0.922 | 0.570 | 0.933 | 0.793 | 0.801 |
| | Default | 0.882 | 0.643 | 0.928 | 0.713 | 0.919 | 0.555 | 0.937 | 0.805 | 0.798 |

Table 5: Positive F1 and Negative F1 scores of constraint-following verification on the instruction-following evaluation task under different decoding strategies.

in Table 10, the prompt for **checklist-guided critique generation** in Table 11, the prompt for **cross-model verification** in Table 12, and the prompt for **rule-augmented verification** in Table 13 and 14.

C List for Baseline Evaluation Models

For general baselines, we adapt 3 advanced proprietary LLMs, GPT-4.1 (OpenAI, 2023), o4-mini (Jaeck et al., 2024), and Gemini-3-Pro (Team et al., 2023), as well as 5 strong open-source LLMs, GLM-4-32B-0414 (GLM et al., 2024), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen2.5-72B-Instruct (Yang et al., 2024), QwQ-32B (Team, 2025), and Deepseek-R1 (Guo et al., 2025a). For evaluation-specific baselines, we select 4 powerful discriminative or generative reward models, Skywork-Reward-V2-Llama-3.1-8B-40M (Liu et al., 2025a), Prometheus-BGB-8x7B (Kim et al., 2024b), RM-R1-Deepseek-Distilled-Qwen-32B (Chen et al., 2025), and RRM-32B (Guo et al., 2025b).

D Implementation Details of IF-CRITIC

We use the Llama-Factory (Zheng et al., 2024) framework and the Zero Redundancy Optimizer (ZeRO) (Rajbhandari et al., 2020) stage 3 from the Deepspeed library (Rasley et al., 2020) to train IF-CRITIC. All experiments are conducted on 8 H800 GPUs. We utilize the AdamW (Kingma and Ba, 2014) optimizer with a weight decay of 0.1, and set the maximum sequence length to 8192 tokens during all the experiments.

During the supervised fine-tuning, the peak learning rate is set to $5e-6$ with a 10% warmup ratio and a linear scheduler. The batch size is 16, and training is conducted for 2 epochs. During the constraint-level preference-optimization, the peak learning rate is set to $1e-6$ with a 10% warmup ratio and a cosine scheduler. The batch size is 128. Training utilizes a sigmoid loss function with a beta value of 0.1 and spans 1 epoch. Following previous

| Base Model | EvalCritic | CFBench | TRACE | Multi-IF |
|-----------------------|--------------|--------------|--------------|--------------|
| Qwen2.5-14B-Instruct | 0.861 | 0.863 | 0.840 | 0.895 |
| Qwen-2.5-7B-Instruct | 0.822 | 0.798 | 0.806 | 0.813 |
| Qwen-3-8B | 0.855 | 0.823 | 0.820 | 0.848 |
| GLM-4-9B-0414 | 0.797 | 0.810 | 0.801 | 0.810 |
| Qwen-3-14B | 0.868 | 0.863 | 0.828 | 0.860 |
| Qwen-2.5-32B-Instruct | 0.879 | 0.854 | 0.833 | 0.867 |

Table 6: The average of Positive F1 and Negative F1 scores of IF-CRITIC under different base models. Qwen-2.5-14B-Instruct is the base model we used for the main experiments.

works (Hou et al., 2024), an additional SFT loss is added to the chosen critique with a weight of 0.1. The overall API cost for critique training data collection is approximately 4000\$.

E Experimental Results on Other Decoding Strategies

In our preliminary experiment, we have compared the performance of greedy search decoding against sampling-based decoding strategies. Specifically, we conduct additional experiments on DeepSeek-R1 and QwQ-32B using their default decoding strategies (a temperature of 1.0 and a top-p value of 1.0 for DeepSeek-R1; a temperature of 0.6, a top-p value of 0.95, and a top-k value of 20 for QwQ-32B). The F1 scores of constraint-following verification using the checklist-level prompting strategy are presented in Table 5. We find that greedy decoding performs almost identically to their default decoding strategies and does not hurt model performance. Consequently, we use greedy decoding to ensure reproducibility.

F Experimental Results on Other Base Models

In our primary experiment, we experiment with various base models for IF-CRITIC, and the results are presented in Table 6. We observe that LLMs with approximately 8B parameters are still insufficient to fully acquire critique generation capabilities, ex-

hibiting a notable performance gap to larger LLMs. In contrast, LLMs with approximately 14B parameters are generally adequate for learning critique generation, and further increasing model scale yields negligible performance improvement. Considering the trade-off between performance and efficiency, we ultimately choose Qwen-2.5-14B-Instruct as the base model for our main experiments.

G Details of Instruction-Following Optimization

Training Instructions Collection. To ensure the diversity and complexity of constraints, we first construct a comprehensive and hierarchical constraint taxonomy, which consists of 6 primary categories (i.e., format constraints, content constraints, behavioral constraints, style constraints, role constraints, and environmental constraints) and 54 secondary categories. During constraint synthesis, we randomly sample 2 to 5 secondary constraint categories and require Deepseek-R1 to introduce several corresponding constraints into the seed instructions, thereby increasing their complexity. In contrast to previous work that only considers independent and atomic constraints (Jiang et al., 2023; Ren et al., 2025; Peng et al., 2025a), we further ask Deepseek-R1 to compose multiple constraints according to the composition types proposed in ComlexBench (Wen et al., 2024), including: (1) *And*: Coordination between different constraints, (2) *Chain*: Sequential completion of constraints, (3) *Selection*: Conditional selection of constraints, and (4) The nested structure of the above composition types. In this way, our constraint synthesis can more comprehensively encompass the complex instruction types in real-world scenarios.

In this scenario, it is notable that most of the hard constraints cannot be reliably verified using automatic verification codes alone, such as constraints targeting for specific segments of the response (e.g., *Generate 10 titles, each no more than 8 words*) and the composition of soft and hard constraints (e.g., *Bold all adjectives that express feelings, such as happy*), as it is challenging for verification code to accurately extract the corresponding segments of the response that should follow these constraints. We provide some examples of the seed instructions and the corresponding final constructed complex instructions in Table 8.

Training Configuration. For DPO training, we use the Llama-Factory (Zheng et al., 2024) frame-

work, and all other settings are the same as the preference optimization for IF-CRITIC. During the reward calculation, IF-CRITIC and QwQ-32B utilize the vllm (Kwon et al., 2023) inference framework, while Skywork-Reward-V2-Llama-3.1-8B-40M is deployed by sglang³ (since this reward model is not compatible with the vllm framework currently). The reward calculation and DPO training are both conducted on 8 H800 GPUs.

For GRPO training, we use the VerL (Sheng et al., 2025) framework. The learning rate is set to 2e-6, with the maximum sequence length for both prompts and responses set to 4096 tokens. The train-batch size and PPO mini-batch size are each set to 32. For rollout responses generation, we employ the vllm (Kwon et al., 2023) framework, utilizing 60% of the available GPU memory, with a temperature and top-p value of 1.0. To encourage stable training, we also incorporate KL divergence regularization with a coefficient of 0.001, using the low-variance KL implementation. We save checkpoints every 30 steps during training. Following TULU 3 (Lambert et al., 2024) and VerIF (Peng et al., 2025a), we use IFEval (Zhou et al., 2023) as the validation set to select the best checkpoint. All experiments are conducted on 24 H800 GPUs, with 8 GPUs dedicated to model training and the remaining 16 GPUs allocated for API deployment of the instruction-following critic or reward model. The deployment framework is identical to the inference framework described in the above DPO training.

H Reward Curves During GRPO Training

We provide the reward curves during GRPO training when IF-CRITIC and QwQ-32B are employed as the LLM critic in Figure 4, and Skywork-Reward-V2-Llama-3.1-8B-40M as the reward model in Figure 5. IF-Critic has a stronger error-detection capability in instruction following, which leads to it generally assigning lower rewards to policy models than QwQ-32B. We also observe that the rewards provided by Skywork-Reward-V2-Llama-3.1-8B-40M may have potential length bias. Using it as a reward model for GRPO training significantly increases the average response length of the policy model, resulting in a longer average per-step training time compared to IF-CRITIC.

³<https://github.com/sgl-project/sglang>

| Policy Model | Method | Reward Model / Critic | AlignBench | MMLU-Pro | Omni-MATH | HumanEval | Average |
|----------------------|--------|-----------------------|------------|----------|-----------|--------------|--------------|
| Qwen2.5-7B-Instruct | - | - | 58.61 | 57.32 | 32.20 | 79.27 | 56.85 |
| | DPO | Skywork-Reward-V2 | 61.36 | 57.41 | 31.95 | 75.00 | 56.43(-0.42) |
| | | QwQ-32B | 60.81 | 57.14 | 32.05 | 77.44 | 56.86(+0.01) |
| | | IF-CRITIC-14B (Ours) | 61.16 | 57.20 | 32.10 | 77.44 | 56.98(+0.13) |
| | GRPO | Skywork-Reward-V2 | 61.49 | 57.43 | 31.18 | 74.39 | 56.12(-0.73) |
| | | QwQ-32B | 62.64 | 57.51 | 32.69 | 73.78 | 56.66(-0.19) |
| IF-CRITIC-14B (Ours) | | 64.83 | 56.57 | 32.11 | 76.83 | 57.59(+0.74) | |
| LLama3.1-8B-Instruct | - | - | 40.66 | 44.17 | 16.05 | 64.02 | 41.23 |
| | DPO | Skywork-Reward-V2 | 44.93 | 44.75 | 15.44 | 63.41 | 42.13(+0.90) |
| | | QwQ-32B | 41.23 | 44.61 | 16.54 | 65.85 | 42.06(+0.83) |
| | | IF-CRITIC-14B (Ours) | 45.76 | 44.11 | 16.02 | 63.41 | 42.33(+1.10) |
| | GRPO | Skywork-Reward-V2 | 46.97 | 45.65 | 17.38 | 64.02 | 43.51(+2.28) |
| | | QwQ-32B | 45.95 | 44.76 | 17.25 | 60.37 | 42.08(+0.85) |
| IF-CRITIC-14B (Ours) | | 46.26 | 45.52 | 17.75 | 60.98 | 41.98(+1.40) | |

Table 7: Experimental results (%) on various general benchmarks. The original results of AlignBench range from 1 to 10; we convert the scale to 10-100 for ease of calculating average performance. Skywork-Reward-V2 refers to Skywork-Reward-V2-Llama-3.1-8B-40M.

I Experimental Results on General Benchmarks

We provide the performance of LLMs after instruction-following optimization in four general benchmarks: AlignBench (Liu et al., 2024), MMLU-Pro (Wang et al., 2024), Omni-MATH (Gao et al., 2025), and HumanEval (Chen et al., 2021) in Table 7. GPT-4o-2024-11-20 is used as the evaluation model for AlignBench, and we use OmniJudge⁴ for the evaluation of Omni-MATH. The results reveal that conducting instruction-following optimization with IF-CRITIC does not harm the general performance of LLMs.

J Human Annotation Guideline for Constraint Verification

The human annotation guideline for constraint verification in the section **Instruction-Following Evaluation** is shown in Table 15. Given an instruction and a corresponding model response, as well as the constraint checklist for the instruction, human annotators are instructed to judge whether the model response follows each constraint in the checklist.

⁴<https://huggingface.co/KbsdJames/Omni-Judge>

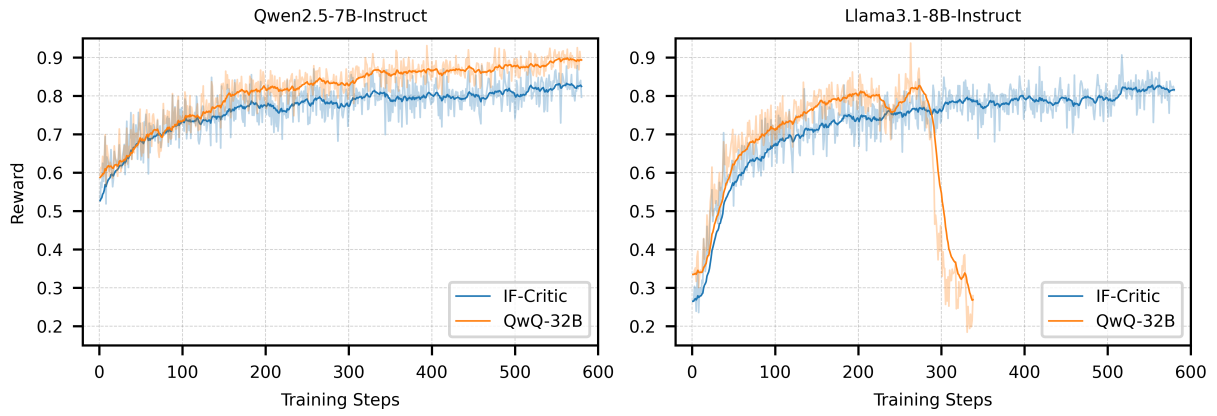


Figure 4: Reward curves during GRPO training when IF-CRITIC and QwQ-32B are employed as the LLM critics. For LLama-3.1-8B-Instruct, training with QwQ-32B results in a model collapse after 300 steps, with the model tending to generate extensive repetitive and meaningless content. Due to efficiency considerations, we terminate further training and calculate the average per-step training time for all critics and reward models based on the first 300 steps.

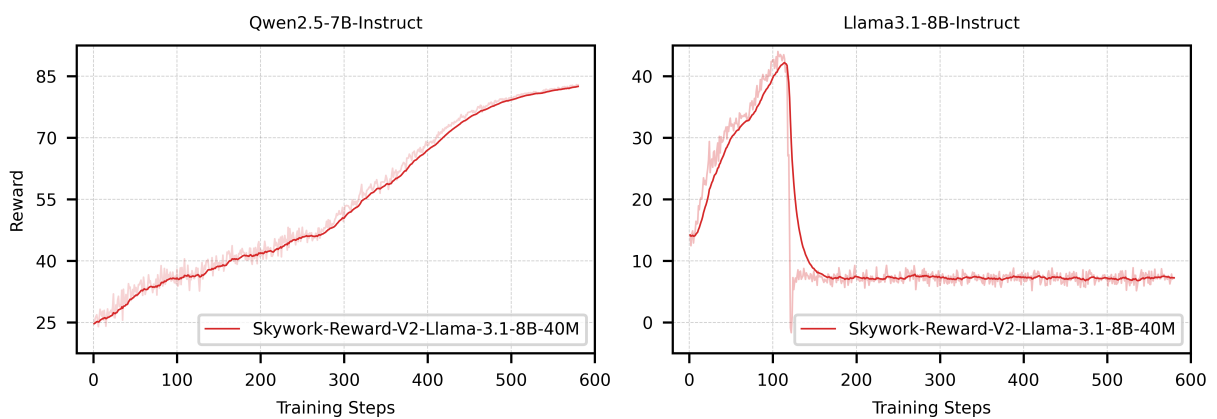


Figure 5: Reward curves during GRPO training when Skywork-Reward-V2-Llama-3.1-8B-40M is employed as the reward model.

| Type | Prompt |
|-------------------|---|
| Seed Instruction | You're an expert in artificial intelligence training. How is a machine learning model built? Please explain in a clear and easy-to-understand tone. |
| Final Instruction | You're an expert in artificial intelligence training. Please explain the process of building a machine learning model in five stages, following a "general-to-specific-to-general" structure. Ensure the following requirements are satisfied: 1. Use the symbols ❶❷❸ to notate each sub-step within each stage. 2. Do not include any mathematical formulas or code examples. 3. Integrate real-life kitchen or cooking analogies within each stage. 4. Finally, use ▷ to highlight three common mistakes people make when building a machine learning model. 5. Write in the second-person perspective, ensure no technical terms exceed middle school math knowledge, and make sure all analogies involve kitchen or cooking elements. |
| Seed Instruction | Translate the following English words into Chinese: attachment, surpasses, dissent, retention, unobtrusive, glazed, entails, confederates, skeptical, perceive, enormity, fanatical |
| Final Instruction | Please translate the following English words into Chinese, strictly following these requirements: 1. Precede each translated Chinese word with a numbered bullet point (e.g., ❶, ❷, ❸). 2. Indicate the part of speech of the original English word (e.g., noun/verb/adjective) in parentheses after each translation. 3. Provide a Chinese synonym after each Chinese translation word, separated by a dash. Words to be translated: attachment, surpasses, dissent, retention, unobtrusive, glazed, entails, confederates, skeptical, perceive, enormity, fanatical |
| Seed Instruction | Write a conversation between a Chinese international student and a Japanese student who decide to try a new restaurant together. While they're having their meal, one of them suggests that they go to see a movie afterwards. |
| Final Instruction | Please write a conversation between Xiaolin, a Chinese international student, and Yamada, a Japanese student. In the conversation, they decide to try a new restaurant together. While they're having their meal, one of them suggests that they go to see a movie afterwards. Ensure the following requirements are satisfied: 1. The conversation should contain three stages: (1) Stage 1: Arranging to dine at the restaurant (at least 2 turns), (2) Stage 2: Discussing the food during the meal (at least 3 turns), (3) Stage 3: Proposing to watch a movie (at least 3 turns). Separate each stage of the conversation with two blank lines. 2. If the restaurant is a Chinese restaurant, the suggested movie genre should be kung fu films. If the restaurant is a sushi restaurant, the suggested movie genre should be animated films. 3. Xiaolin should naturally mention Mid-Autumn Festival customs in the conversation while Yamada should incorporate elements of the Obon Festival. Please use specific festival details (e.g., types of mooncakes, or forms of Bon Odori). 4. Please write the conversation in Chinese. Indicate the speaker (Xiaolin: / Yamada:) at the start of each turn. The entire conversation should be no more than 10 turns. Please first determine the restaurant type (Chinese restaurant or sushi restaurant), and then write the conversation according to the corresponding scene. |

Table 8: Some examples of the seed instructions and the corresponding final constructed complex instruction in the instruction-following optimization.

You are an expert specializing in evaluating the quality of user instructions. Please evaluate the quality of the user instruction based on the following criteria.

Evaluation Criteria:

1. **Low Quality:** The user instruction contains significant issues such as unclear, incomplete, or ambiguous information, making it impossible to determine the intent behind the instruction.
2. **Medium Quality:** The user instruction may include some unclear, incomplete, or ambiguous elements; however, the overall intent can still be inferred.
3. **High Quality:** The user instruction is clear, complete, and unambiguous, allowing the intent to be easily and definitively understood.

Note:

1. You do not need to answer the user instruction, but only need to output the evaluation result.
2. If the user instruction requires additional retrieval or the use of tools to obtain an answer, it should be evaluated as **Medium Quality**.

Here are some examples and the user instruction to be evaluated:

[The Start of Examples]

{in_context_examples}

[The End of Examples]

[The Start of User Instruction]

{instruction}

[The End of User Instruction]

Output Format

""

Prompt Quality: **Low Quality** / **Medium Quality** / **High Quality**

""

Please directly output the evaluation result:

Table 9: The prompt template for scoring the quality of user instructions.

You are an information extraction expert specializing in identifying all constraints within instructions.

We will provide you with a user instruction. Your task is to extract all constraints present in that instruction. **Constraints** refer to all instructional content, excluding auxiliary information such as **background knowledge**, **text materials**, and **in-context examples**. This includes, but is not limited to, descriptions of basic tasks to be completed, and specific requirements on output content, style, and format. Output all constraints using this format exactly:

""

[The Start of Constraint 1]

Constraint: ... (Please generate a specific constraint of the instruction. It must be complete and detailed, with no information omitted or altered in any way)

[The End of Constraint 1]

[The Start of Constraint 2]

Constraint: ... (Please generate a specific constraint of the instruction. It must be complete and detailed, with no information omitted or altered in any way)

[The End of Constraint 2]

...

""

Remember the following points:

- (1) Each constraint you output must be a verbatim excerpt from the given instruction. You must not fabricate, paraphrase, or add any content.
- (2) You must extract and output all constraints in the order in which they appear in the instruction, without omitting any constraints or inventing constraints not present in the instruction. To reiterate: **Constraints** refer to all instructional content, excluding auxiliary information such as **background knowledge**, **text materials**, and **in-context examples**.
- (3) Do not output duplicate constraints. Each portion of the given instruction should appear in at most one constraint, and must not be repeated across multiple constraints.
- (4) Each constraint should be atomic and independent, without inclusion or dependency relationships. At the same time, ensure appropriate granularity of constraints: neither too fine nor too coarse. Each constraint should have complete semantics and be understandable on its own, without reference to other constraints. At the same time, relatively independent constraints must not be merged into a single constraint.

Here are the user instruction:

[The Start of User Instruction]

{instruction}

[The End of User Instruction]

Table 10: The prompt template for constraint checklist generation.

You are a fair judge, specializing in evaluating the quality of an AI assistant's responses to user instructions.

We will provide you with a user instruction, a constraint checklist for the user instruction, and an AI assistant's response. Please think carefully and provide a detailed analysis of whether the AI assistant's response follows each item on the constraint checklist. You need to analyze and assess every constraint in the checklist, and do not need to analyze anything that is not listed in the checklist. You must output your explanation and judgment for each constraint, strictly following the format below:

""

[The Start of Constraint 1]

Constraint: ... (Directly output the first constraint from the checklist here with no modifications)

Explanation: ... (Provide a thorough, step-by-step analysis of whether the AI assistant's response follows this constraint, referencing specific details of the response)

Judgment: [[The AI assistant's response follows this constraint]] or [[The AI assistant's response does not follow this constraint]]

[The End of Constraint 1]

[The Start of Constraint 2]

Constraint: ... (Directly output the second constraint from the checklist here with no modifications)

Explanation: ... (Provide a thorough, step-by-step analysis of whether the AI assistant's response follows this constraint, referencing specific details of the response)

Judgment: [[The AI assistant's response follows this constraint]] or [[The AI assistant's response does not follow this constraint]]

[The End of Constraint 2]

...

Remember the following points:

- (1) You should only focus on whether the response follows the items in the checklist, not on the overall quality of the response. Other shortcomings in the response (such as not fulfilling a constraint to a high standard) should not affect your judgment unless the constraint is not followed.
- (2) Your judgment should be as strict as possible. You should output "[[The AI assistant's response follows this constraint]]" only if the response completely follows every part of the constraint as stated, without any omission or mistake.
- (3) When analyzing a given constraint, you should not take into account whether other constraints are followed.

Here are the user instruction, constraint checklist, and AI assistant's response:

[The Start of User Instruction]

{instruction}

[The End of User Instruction]

[The Start of Constraint Checklist]

{checklist}

[The End of Constraint Checklist]

[The Start of AI Assistant's Response]

{response}

[The End of AI Assistant's Response]

Table 11: The prompt template for checklist-guided critique generation.

| Template | Prompt |
|--------------------------|---|
| Correctness Verification | <p>You are a fair judge, specializing in evaluating the quality of an AI assistant's responses to user instructions.</p> <p>We will provide you with a user instruction, an AI assistant's response, a specific constraint from the user instruction, and a critique of the AI's response regarding whether the response follows this constraint from the user instruction. The critique comprises two parts: "Explanation" and "Judgment". Your task is to carefully analyze this critique to judge whether it is reasonable and correct. After your analysis, output either "[[The given critique is correct]]" or "[[The given critique is not correct]]" as your final conclusion.</p> <p>A correct critique must simultaneously satisfy all of the following three principles:</p> <ol style="list-style-type: none"> (1) Only focus on whether the response follows the constraint, not on the overall quality of the response. Other shortcomings in the response (such as not fulfilling the constraint to a high standard) should not affect the judgment unless the constraint is not followed. (2) Be as strict as possible. Output "[[The AI assistant's response follows this requirement]]" only if the response completely follows every part of the requirement as stated, without any omission or mistake. (3) Independently analyzes the given constraint, and does not take into account whether other constraints are followed. <p>Here are the user instruction, the constraint, the AI assistant's response, and the critique to be judged:</p> <p>[The Start of User Instruction] {instruction} [The End of User Instruction]</p> <p>[The Start of Constraint] {constraint} [The End of Constraint]</p> <p>[The Start of AI Assistant's Response] {response} [The End of AI Assistant's Response]</p> <p>[The Start of Critique] {critique} [The End of Critique]</p> |
| Consistency Verification | <p>You are a fair judge, specializing in evaluating the quality of an AI assistant's responses to user instructions.</p> <p>We will provide you with a constraint, a critique that comprises an explanation and a judgment generated by an evaluator, whose main purpose is to analyze whether an AI assistant's response follows the constraint. You do not need to consider the content of the constraint or whether the critique is correct. You only need to judge whether the explanation and judgment within the critique are logically consistent, specifically:</p> <ol style="list-style-type: none"> (1) If the explanation section determines that the response follows the constraint, the judgment should be: "[[The AI assistant's response follows the constraint]]". (2) If the explanation section identifies any mistake or shortcoming in the response, indicating that the constraint is not fully followed, the judgment should be: "[[The AI assistant's response does not follow the constraint]]". <p>Remember the following points:</p> <ol style="list-style-type: none"> (1) There should not be situations such as "partially follow the constraint". As long as the explanation identifies any mistakes or shortcomings in the response that result in the constraint not being fully followed, the judgment should be: "[[The AI assistant's response does not follow the constraint]]." (2) If the explanation states that the constraints and the response are unrelated and there is no need for evaluation, the judgment should be: "[[The AI assistant's response follows the constraint]]." <p>Please analyze the logical consistency between the explanation and judgment following the above principles. Output your analysis first, and then output either "[[Explanation and Judgment are consistent]]" or "[[Explanation and Judgment are not consistent]]" to represent your final verdict.</p> <p>Here are the constraint and the critique to be judged:</p> <p>[The Start of Constraint] {constraint} [The End of Constraint]</p> <p>[The Start of Critique] {critique} [The End of Critique]</p> |

Table 12: The prompt templates for cross-model verification.

You are a fair judge, specializing in evaluating the quality of an AI assistant's responses to user instructions.

We will provide you with a user instruction, an AI assistant's response, and a specific constraint from the user instruction. Please first determine whether the given constraint is related to length (e.g., "The article should be no less than 800 words" or "Each title must be 8 characters long"). If so, please extract the segments from the AI assistant's response that should follow this length requirement (extract the original text from the response without any modifications).

Reference Example
{in_context_examples}

Here are the user instruction, the constraint, and the AI assistant's response:

[The Start of User Instruction]
{instruction}
[The End of User Instruction]

[The Start of Constraint]
{constraint}
[The End of Constraint]

[The Start of AI Assistant's Response]
{response}
[The End of AI Assistant's Response]

Remember the following points:

- (1) The given constraints may contain length requirements for multiple segments of the response. Therefore, numerous segments of content may need to be extracted.
- (2) You only need to determine whether the given constraint is related to length, without considering other constraints of the user instruction.
- (3) Output your analysis and conclusion strictly following the format below:

Analysis: ...(Provide a detailed analysis to determine whether the given constraint is related to length)

Conclusion: ...

1. (If the constraint is NOT related to length)

```
"""json
{
  "Length Constraint" : False
}
"""
```

2. (If the constraint is related to length)

```
"""json
{
  "Length Constraint" : True,
  "Extracted Segments" : [
    {
      "Length Requirement within the Constraint": ...(Provide the length requirement within the constraint. Be
sure to extract the original text from the constraint without making any modification),
      "Corresponding Segment in Response": ...(Provide the segment of the response that should follow this
requirement without making any modification. If no corresponding segment exists in the response, output "No corresponding
segment exists.")
    },
    {
      "Length Requirement within the Constraint": ...,
      "Corresponding Segment in Response": ...
    },
    ... (Analogously, extracting all segments in the response that should follow the length requirements of the given
constraint)
  ]
}
"""
```

Table 13: The prompt template for length constraint identification and segment extraction.

You are a fair judge, specializing in evaluating the quality of an AI assistant's responses to user instructions.

We will provide you with a user instruction, a specific constraint from the user instruction, and the AI assistant's response. Please think carefully and provide a detailed analysis of whether the AI assistant's response follows the given constraint.

The given constraint is related to length. We will also provide a JSON data, which contains multiple items. Each item consists of three elements: (1) a length requirement within the constraint, (2) a segment of the responses that should follow the length requirement, and (3) the actual length of the segment. It is important that you consider the "actual length provided in the JSON data" as the golden-truth length, which is absolutely accurate. Do not make your own calculation about "the length of the segment". You must rely solely on the "actual length" as the basis for length validation. Output your explanation and judgment for the given constraint based on the actual length provided in the JSON data, strictly following the format below:

""

[The Start of Constraint]

Constraint: ... (Directly output the first constraint from the checklist here with no modifications)

Explanation: ... (Provide a thorough, step-by-step analysis of whether the AI assistant's response follows this constraint, referencing specific details of the response)

Judgment: [[The AI assistant's response follows this constraint]] or [[The AI assistant's response does not follow this constraint]]

[The End of Constraint]

""

Remember the following points:

- (1) You should only focus on whether the response follows the constraint, not on the overall quality of the response. Other shortcomings in the response (such as not fulfilling the constraint to a high standard) should not affect your judgment unless the constraint is not followed.
- (2) Your judgment should be as strict as possible. You should output "[[The AI assistant's response follows this constraint]]" only if the response completely follows every part of the constraint as stated, without any omission or mistake.
- (3) Independently analyzing the given constraint, and does not take into account whether other constraints are followed.
- (4) The given constraint may contain both length requirements and other requirements. For length requirements, you must rely solely on the "actual length provided in the JSON data" as the basis for length validation. For other requirements, please refer to the above principles and carefully evaluate them by yourself.
- (5) When generating your explanation, it is forbidden to mention phrases such as "according to the given JSON data" or "based on the provided actual length". Your explanation should convey that the length is calculated independently by you, without referencing the given JSON data.

Here are the user instruction, the constraint, the AI assistant's response, and the JSON data about length:

[The Start of User Instruction]

{instruction}

[The End of User Instruction]

[The Start of the Constraint]

{constraint}

[The End of the Constraint]

[The Start of AI Assistant's Response]

{response}

[The End of AI Assistant's Response]

[The Start of The Json Data]

{json_data}

[The End of The Json Data]

Table 14: The prompt template for rule-augmented critique revision.

Below, an instruction, a corresponding model response, and a constraint checklist for the instruction will be provided. The checklist contains some of the constraints within the instruction. Your task is to verify whether the model response follows each constraint in the checklist, choose "Followed" or "Not Followed", and provide a brief justification.

Task Details:

1. Please analyze whether the response follows each constraint listed in the given checklist, providing a judgment for each constraint respectively.
2. Your judgments must be strict. Only responses that fully satisfy a constraint can be judged as "Followed". If there is any omission or error regarding a constraint, it must be judged as "Not Followed".
3. Please focus exclusively on the constraints within the given checklist. It is unnecessary to consider whether the response follows any other constraints beyond the checklist.
4. When judging the following of each constraint, your judgement should consider the complete context of the instructions, rather than interpreting the constraint in isolation.

{in-context examples}

[Instruction]

{instruction}

[Model Response]

{model_response}

[Constraint Checklist]

{checklist}

Your choice for the first constraint in the checklist: {option} A. Followed B. Not Followed

Your justification: {justification}

Your choice for the second constraint in the checklist: {option} A. Followed B. Not Followed

Your justification: {justification}

.....

Table 15: Human annotation guideline for constraint verification. The red part is the information provided to the annotators, and the blue part is content that requires the annotators to make annotations.