The Critique of Critique

Shichao Sun[1,5], **Junlong Li**[2,5], **Weizhe Yuan**[4], **Ruifeng Yuan**[1], **Wenjie Li**[1]*, **Pengfei Liu**[2,3,5]*
[1] The Hong Kong Polytechnic University, [2] Shanghai Jiao Tong University
[3] Shanghai Artificial Intelligence Laboratory, [4] New York University
[5] Generative AI Research Lab (GAIR)
cswjli@comp.polyu.edu.hk, pengfei@sjtu.edu.cn

Abstract

Critique, as a natural language description for assessing the quality of model-generated content, has played a vital role in the training, evaluation, and refinement of LLMs. However, a systematic method to evaluate the quality of critique is lacking. In this paper, we pioneer the critique of critique, termed METACRITIQUE, which builds specific quantification criteria. To achieve a reliable evaluation outcome, we propose Atomic Information Units (AIUs), which describe the critique in a more fine-grained manner. METACRITIQUE aggregates each AIU's judgment for the overall score. Moreover, METACRITIQUE delivers a natural language rationale for the intricate reasoning within each judgment. Lastly, we construct a meta-evaluation dataset covering 4 tasks across 16 public datasets involving human-written and LLM-generated critiques. Experiments demonstrate that METACRITIQUE can achieve near-human performance. Our study can facilitate future research in LLM critiques based on our following observations and released resources: (1) superior critiques judged by METACRITIQUE can lead to better refinements, indicating that it can potentially enhance the alignment of existing LLMs; (2) the leaderboard of critique models reveals that open-source critique models commonly suffer from factuality issues; (3) relevant code and data are publicly available at https://github.com/GAIR-NLP/MetaCritique to support deeper exploration; (4) an **API** at PyPI with the usage documentation in Appendix C allows users to assess the critique conveniently.

1 Introduction

Natural language critique has assumed a crucial role in advancing the development of Large Language Models (LLMs), ranging from the training of a more helpful and harmless model (Bai et al., 2022; OpenAI, 2023; Scheurer et al., 2023; Wu et al., 2023), alignment evaluation of model generations (Wang et al., 2023b; Zheng et al., 2023; Chan et al., 2023; Li et al., 2023) to the refinement of defective model outputs (Madaan et al., 2023; Gou et al., 2023; Ye et al., 2023; Akyurek et al., 2023). While a bunch of recent works are being done using generated critiques to assist in the development of LLMs (Cui et al., 2023; Kim et al., 2023; Wang et al., 2023a; Li et al., 2023; Ke et al., 2023), there has not been enough emphasis on how to automatically and efficiently evaluate the quality of these critiques due to the following challenges: (i) **quantification**: establishing specific criteria to qualify the critique rating, (ii) **reliability**: ensuring transparency to calculate the comparable score, and (iii) **intricacy**: grasping the complex relations among multiple concepts in the critique evaluation.

In this paper, we pioneer the critique of critique, termed METACRITIQUE, to get over the above hurdles. An example of METACRITIQUE is shown in Figure 1. Firstly, METACRITIQUE tackles the **quantification** issue by establishing specific criteria, i.e., a meaningful critique should provide factual statements and comprehensive assessments, which will be quantified by two metrics: *precision* and *recall*. Precision serves to gauge the accuracy of the critique's content, ensuring each point is factual, while recall measures the extent to which the critique fully covers the necessary breadth of information, reflecting its comprehensiveness. Secondly, METACRITIQUE addresses the **reliability** concern by introducing Atomic Information Units (AIUs). AIUs symbolize the fundamental segments of informative critique that cannot be divided further. METACRITIQUE converts the critique-level evaluation into AIU-level evaluation to minimize ambiguity in the evaluation process. Subsequently, METACRITIQUE aggregates these AIU-level outcomes to produce the overall score, thereby guaranteeing the transparency of the scoring process.

* Corresponding Authors.
Work done while visiting GAIR Lab.

Findings of the Association for Computational Linguistics: ACL 2024, pages 9077–9096
August 11-16, 2024

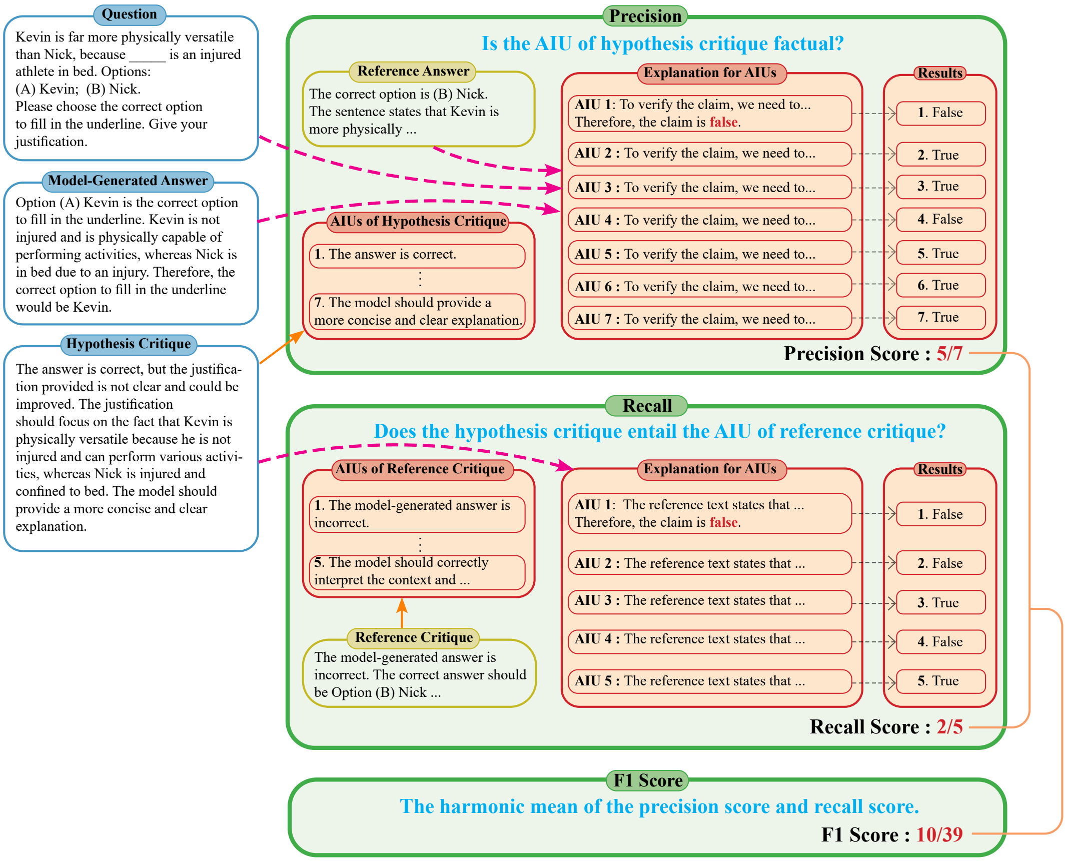


Figure 1: An example of METACRITIQUE for hypothesis critique evaluation. More details can be found in Figure 5. Atomic Information Units (AIUs) are fundamental segments of informative critique that cannot be divided further. The ratio of "True" AIUs is calculated as the corresponding score.

Lastly, inspired by the fact that complex reasoning problems can be alleviated by using LLMs with Chain-of-Thought (CoT) (Wei et al., 2022), we attempt to resolve the **intricacy** problem by generating a natural language rationale step by step for each AIU-level judgment. As a result, it can enhance the reliability of judgment and facilitate human involvement in the evaluation loop.

Given the absence of a dataset for critique evaluation, we curate a meta-evaluation dataset covering 4 tasks (question answering, reasoning, entailment, and summarization) across 16 public datasets, involving human-written and LLM-generated critiques. Our METACRITIQUE achieves near-human performance, indicating that METACRITIQUE can help understand human annotation and LLM's reflection. Besides, METACRITIQUE can identify high-quality critiques, which lead to improved results via iterative refinement. This indicates that METACRITIQUE can enhance the alignment of existing LLMs. A leaderboard of critique models also aids in identifying the pros and cons of various critique models. In conclusion, METACRITIQUE can potentially advance the progress of LLMs.

2 METACRITIQUE

METACRITIQUE evaluates the quality of a hypothesis critique by generating its critique, i.e., the critique of critique. It involves three steps: (1) reference generation, (2) AIU extraction, and (3) hypothesis critiquing as illustrated in Figure 2.

To facilitate the detailed description of METACRITIQUE, we first introduce some important concepts. **Question** denotes a user's query or instruction that prompts LLMs to produce a pertinent and insightful response. **Model-generated Answer** describes the textual content created by LLMs as a reaction to the question. **Reference Answer** is a ground-truth answer for the question. **Hypothesis Critique** is the natural language feedback to point out errors of the model-generated answer and provide actionable suggestions. It can be written by either human annotators or LLMs. **Reference Critique** is an ideal critique to comment accurately and thoroughly on the model-generated answers. **Atomic Information Unit (AIU)** is the smallest

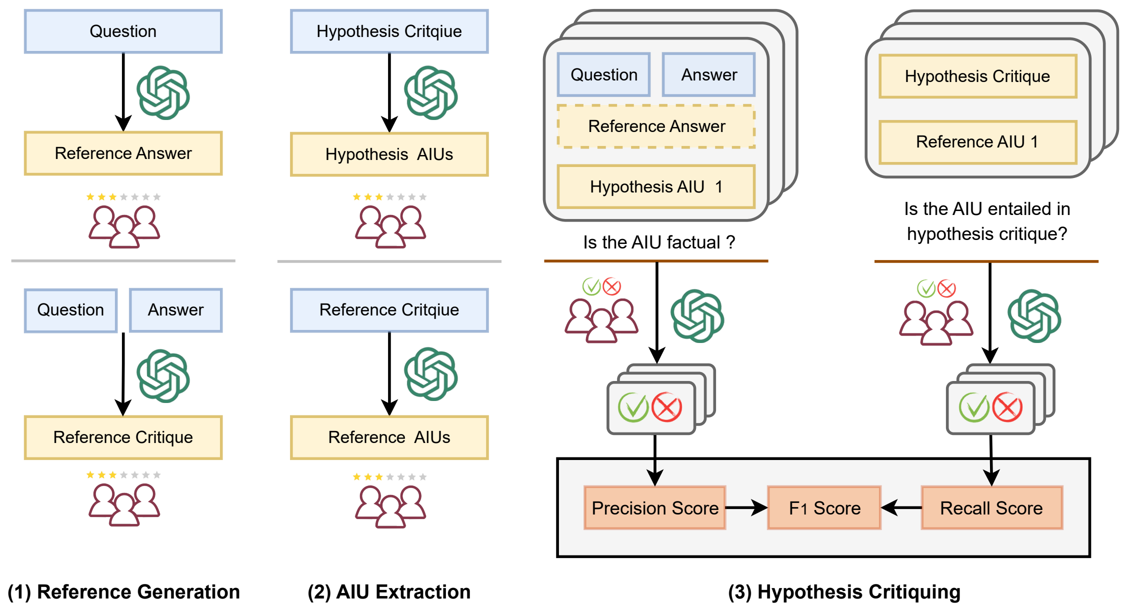


Figure 2: An overview for our MetaCritique powered by GPT-4 and human annotation for each step. The "answer" in the blue box is the model-generated answer. The star rating refers to annotators giving 1-7 likert score, and the check-and-cross mark indicates that annotators give the True or False label, just like GPT-4.

unit that can self-sufficiently convey a piece of information. It can help reduce the ambiguity of the evaluation process and improve the transparency of the evaluation outcome. Generating a numeric score for the critique depends on the information volume it encompasses. For example, a precision score considers the volume of correct information in the critique. However, information, being intangible, renders the task of determining the volume of information unfeasible (Porat, 1977; Soofi, 1994). By extracting and counting AIUs from a critique, we can approximate the volume of information as the number of AIUs because AIUs are the most fundamental elements of critiques.

2.1 Step 1: Reference Generation

Reference is essential for most text generation evaluation (Papineni et al., 2002; Lin, 2004; Zhang* et al., 2020; Yuan et al., 2021). Also, it is crucial for our MetaCritique. We need reference answers to calculate the precision score, and reference critiques to calculate the recall score. However, reference answers or critiques are always unavailable because they require significant human effort. To address this issue, we adopt the GPT-4 generated content as a proxy of the references as in previous works (OpenAI, 2023; Zheng et al., 2023; Peng et al., 2023; Cui et al., 2023; Li et al., 2023). Moreover, we conduct human evaluations to ensure the quality of the GPT-4 generated content. We provide detailed prompting instructions in Table 8.

2.2 Step 2: AIU Extraction

AIU extraction aims to split a critique into AIUs. It is similar to some prior works such as verifiable claims extraction for factuality detection (Chern et al., 2023) and atomic content units extraction for summarization evaluation (Liu et al., 2023b). It has been proven that suitably prompted GPT-4 can precisely conduct such a task as mentioned in Chern et al. (2023). Inspired by their success, we implement AIUs extraction by prompting GPT-4. Besides, we conduct human evaluations to verify the quality of the GPT-4 extraction. We provide detailed prompting instructions in Table 10.

2.3 Step 3: Hypothesis Critiquing

Precision We devise *precision* to verify the factuality of the hypothesis critique. It is motivated by the fact that good critiques should state factual information without any hallucination. Specifically, we design the **precision task**, which is a binary classification task on the AIU level to validate whether each AIU is factual or non-factual. This task receives the question, model-generated answer and the optional reference answer as the context, and outputs whether an AIU from the hypothesis critique is factual or non-factual along with a natural language rationale.

We prompt LLMs with strong instruction-following capability to implement this evaluation. We follow the idea of CoT reasoning to design the instructions. Firstly, the LLM needs to find the necessary information to verify the AIU. Then, the LLM explains and reasons whether the AIU is factual or not. Finally, the LLM states the conclusion. Detailed prompting instructions with demonstrations are shown in Table 11.

After checking each AIU in the hypothesis critique, we denote the precision score s_p as the proportion of factual AIUs relative to the total count of AIUs in the hypothesis critique.

Recall We use *recall* to assess the coverage of the hypothesis critique over the reference critique. It is motivated by the fact that good critiques should contain all key points of the reference critique without any omissions. In this evaluation, we design **recall task**, which is a binary classification task on the AIU level to classify whether the hypothesis critique entails each AIU of the reference critique. This task receives the hypothesis critique as the premise and outputs whether an AIU from the reference critique is entailed in the hypothesis critique or not along with a natural language rationale.

We prompt LLMs guided by the CoT reasoning to perform this evaluation. Firstly, the LLM analyses if the AIU from the reference critique is mentioned or logically inferred from the hypothesis critique. Subsequently, the LLM states whether the AIU is entailed or not. Detailed prompting instructions with demonstrations are in Table 12.

After checking each AIU in the reference critique, we denote the recall score s_r of the hypothesis critique as the ratio of entailed AIUs to all AIUs from the reference critique.

F1 Score We introduce the F1 Score s_f as an overall assessment score, which harmonizes the *precision* score s_p and *recall* score s_f as follows:

$$s_f = 2\frac{s_p \cdot s_r}{s_p + s_r} \tag{1}$$

3 Meta-Evaluation Dataset

In this section, we elaborate on constructing a meta-evaluation dataset and human annotation for evaluating the critique evaluation, along with presenting its statistical features. We first collect questions, model-generated answers containing flaws, and human-written or LLM-generated critiques (hypothesis critique). Subsequently, we use GPT-4 to generate critiques and answers as references and extract AIUs from critiques. Finally, human annotators use these data to complete precision and recall tasks for each AIUs as shown in Figure 2.

3.1 Collection of Question and Model-Generated Answer

To get broad coverage of NLP domains, we collect ready-made question-answer pairs from the Shepherd (Wang et al., 2023a) dataset, which consists of question-answer-critique triads for generating critique across various domains. The model-generated answer in this dataset has reasonable errors, so that the critique can be generated to improve the response. We carefully extract some data to exclude the tasks that need specific tools to find errors, like code generation. As a result, we collect data from four domains: entailment, reasoning, question answering, and summarization, across 16 datasets: Entailment Bank (Dalvi et al., 2021), e-SNLI (Camburu et al., 2018), Adversarial NLI (Nie et al., 2020), ECQA (Aggarwal et al., 2021), CosmosQA (Huang et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020), ARC (Clark et al., 2018), RACE (Lai et al., 2017), SIQA (Sap et al., 2019), TriviaQA (Joshi et al., 2017), Natural Question (Kwiatkowski et al., 2019), BoolQ (Clark et al., 2019), GPT-3 summarization (Goyal et al., 2022), DeFacto (Liu et al., 2023a).

3.2 Collection of Hypothesis Critique

Hypothesis critique can be written by human annotators (abbreviated to ***Hypo.h*** in tables) and LLMs (abbreviated to ***Hypo.l*** in tables). It is valuable to evaluate human-written critique, as it can help to understand the common shortcomings of human annotations. Also, it is necessary to understand how well the LLM-generated critiques perform because LLM-generated critiques have been widely used for training, evaluation, and refinement.

For human-written critique, we use one ready-made critique from the Shepherd dataset. For LLM-generated critiques, we generate two critiques for each question-answer pair. These two critiques will be further used to conduct pairwise comparisons. We do not conduct the pairwise comparison between human-written critiques and LLM-generated critiques to avoid potentially misleading outcomes that could arise from their writing styles. In detail, we randomly select two LLMs from WizardLM (Xu et al., 2023) (13B and 70B), GPT-3.5, LLaMA-2 (Touvron et al., 2023) (chat-13B and chat-70B),

Type	# Pair	# Critique	Avg. # AIUs
Hypo.h	100	100	3.31
Hypo.l	100	200	8.10
Reference	100	100	7.02

Table 1: Statistics of meta-evaluation dataset.

Reference Generation		AIUs Extraction		
Answer	Critique	Hypo.h	Hypo.l	Reference
6.51	6.56	6.72	6.57	6.79

Table 2: Human evaluation for GPT-4 outcomes (Likert score on 1-7 scale). 1 is the worst, and 7 is the best.

Vicuna (Chiang et al., 2023) (13B and 33B), and SelFee (Ye et al., 2023) (13B). Then, we respectively prompt the selected LLM to generate one critique. We use the same prompt as the reference critique generation (Table 8) for GPT-3.5. As for other LLMs, the prompt is shown in Table 9.

3.3 Collection of Human Annotation

We collect human labeling results for precision and recall tasks driven by two objectives: (1) to evaluate the performance of different LLMs to execute the precision and recall tasks (Exp. II). (2) to meta-evaluate METACRITIQUE and its counterparts in the light of human judgments (Exp. III-IV).

Before annotation, we prepare the reference answer, reference critique and extracted AIUs via GPT-4 as introduced in Section 2.1 and 2.2. We engaged two postgraduate students to perform the precision and recall task. As shown in Figure 2, they replace the GPT-4 to provide solely binary labels without explanation. A third postgraduate meticulously reviews the work of the first two annotators, ensuring accuracy and resolving any discrepancies. This process is essential for maintaining the reliability of our research data. Lastly, we can calculate the METACRITIQUE scores via these annotated results. These scores (shown in Table 6) represent human judgments (named as **gold scores**).

3.4 Statistics

Table 1 shows the statistics of our meta-evaluation dataset. We collect 100 question-answer pairs. Each pair has 1 human-written critique, 2 LLM-generated critiques, and 1 reference critique. We find that the number of AIUs in human-written critiques is less than half of it in LLM-generated critiques or reference critiques. This implies that human-written critiques possibly contain less information than LLM-generated critiques.

4 Experiments

In this section, we introduce various baselines and experiments to show the feasibility (Exp. I-II) and effectiveness (Exp. III-V) of METACRITIQUE.

4.1 Baseline

We compare multiple modern LLMs for AIU-level precision and recall tasks (Exp. II). The tested models include **Zephyr** (Tunstall et al., 2023), **WizardLM**, **LLaMA-2 Chat**, **Vicuna**, **GPT 3.5**, and **GPT-4**. Moreover, we randomly choose 100 AIUs and engage a postgraduate student (not the annotator) to perform the same task. This result can approximate a ceil performance (named as **Human**). We also introduce **GPT-4 w/o ans**, where we generate the reference answer and perform the precision task in one step. It aims to confirm the importance of pre-generating a reference answer.

We compare four variants of METACRITIQUE with two GPT-4 based methods introduced by Wang et al. (2023a) for Exp. III-V: **MetaCritique$_{\mathbf{GPT4}}$-P**, **MetaCritique$_{\mathbf{GPT4}}$-R** and **MetaCritique$_{\mathbf{GPT4}}$-F1** is respectively the precision score, recall score and F1 score of METACRITIQUE powered by GPT-4. **MetaCritique$_{\mathbf{Open}}$-F1** is the F1 score of METACRITIQUE powered by open-source LLMs, whereby WizardLM 70B is used for the precision task, and WizardLM 13B is used for the recall task because they beat other open-source LLMs in the Table 3. **Pairwise$_{\mathbf{GPT4}}$** is to compare two hypothesis critiques via GPT-4 and pick up the better one. The prompting instruction is shown in Table 16. **Single$_{\mathbf{GPT4}}$** is to generate a likert score (1-7) for a hypothesis critique via GPT-4. The prompting instruction is shown in Table 17.

4.2 Exp-I: Human Evaluation for Reference Generation and AIUs Extraction

Q1: Can GPT-4 outcomes serve as references?

Setup We conduct a human evaluation to validate the quality of the reference answer, reference critique, and extracted AIUs that are generated by GPT-4. We engaged two postgraduate students as annotators. We ask each human annotator to rate the outcome on a 1–7 likert score. Detailed instructions for human annotators to evaluate reference answer generation, reference critique generation, and AIUs extraction can be found in Table 13, 14,

Model	Size	Precision Task		Recall Task	
		Hypo.h	Hypo.l	Hypo.h	Hypo.l
Human	–	90.00	86.00	86.00	85.00
Llama 2 Chat	7B	31.42	34.14	8.55	5.98
	13B	56.50	52.22	55.13	51.78
	70B	54.08	58.09	72.08	65.46
Vicuna	7B	71.00	70.37	59.69	60.83
	13B	72.81	69.14	71.94	70.66
	33B	62.84	60.12	72.93	63.82
Zephyr-β	7B	62.84	59.44	54.56	61.82
WizardLM	7B	26.89	30.19	52.28	55.56
	13B	62.54	59.81	<u>74.50</u>	<u>74.22</u>
	30B	56.50	63.89	63.53	61.68
	70B	<u>79.76</u>	<u>72.28</u>	72.51	71.44
GPT-3.5	–	81.87	77.28	80.63	81.41
GPT-4 w/o ans	–	86.40	81.48	–	–
GPT-4	–	**89.12**	**87.96**	**85.47**	**86.82**

Table 3: AIU-level accuracy. <u>Underline</u> is the best result among all open-source LLMs, and **bold** is the best outcome among all LLMs.

and 15, respectively. To obtain a reliable evaluation outcome, we present as much information as possible to annotators. We also allow annotators to carefully search online whenever they need help.

Results Table 2 shows the rating scores for the quality of GPT-4 generated answers, critiques, and AIUs. The left half shows that GPT-4 attains remarkable performance, which can confirm the feasibility of using GPT-4 generated outcomes as references. Moreover, the right half shows that GPT-4 delivers impressive results, which can justify its effective use to extract AIUs for METACRITIQUE.

4.3 Exp-II: AIU-level Accuracy

Q2: Which LLMs are capable of powering METACRITIQUE?

Setup Our METACRITIQUE centers around two binary classification tasks: precision and recall. In this experiment, we investigated the capability of various LLMs to execute two tasks. Each AIU was treated as a unique test case.

Results We present the AIU-level accuracy results in Table 3. We find that GPT-4 outperforms all LLMs by a large margin and achieves an impressive performance (**nearly 90%**), which is comparable to that of humans. This shows that it is reasonable to use METACRITIQUE powered by GPT-4 for automated evaluation of human-written critiques and LLM-generated critiques. Moreover, WizardLM-70B and WizardLM-13B stand out as the top-performing open-source models in precision tasks and recall tasks, respectively. Remarkably, they rival closely GPT-3.5. Lastly, the degradation (around 3% and 6%) without the reference answer indicates that it is necessary to pre-generate a reference answer.

4.4 Exp-III: Correlation Coefficient

Q3: Which evaluation methods can give rating scores that are close to human judgments?

Setup In this experiment, we use different methods to generate a score for the hypothesis critique. We calculate correlation coefficients to measure the correlation between different scoring baselines and human judgments. Specifically, we use Pearson correlation (Lee Rodgers and Nicewander, 1988; Mukaka, 2012), Spearman correlation (Zar, 2005), and Kendall's Tau (Kendall, 1938) as metrics. We calculate the above correlation coefficients between the outcome score and the gold F1 score. To perform a rigorous analysis, we adopt the bootstrapping method (Koehn, 2004) for significance tests.

Results In Table 4, we show the correlation between various methods with human judgments. Our MetaCritique$_{\text{GPT4}}$-F1 beats the Single$_{\text{GPT4}}$ baseline by a large margin, confirming its increased reliability. We also observe that MetaCritique$_{\text{GPT4}}$-P has a reduced correlation in human-written critiques, likely because humans make fewer factuality mistakes, resulting in clustered precision scores. In addition, MetaCritique$_{\text{Open}}$ distinctly exceeds the performance of the Single$_{\text{GPT4}}$ baseline, demonstrating that even less advanced LLMs can surpass Single$_{\text{GPT4}}$ baseline. This indicates that our METACRITIQUE framework is more effective than simple GPT-4 scoring. Lastly, precision and recall scores complement each other in assessing LLM-generated critiques. MetaCritique$_{\text{GPT4}}$-P or MetaCritique$_{\text{GPT4}}$-R is slightly inferior to MetaCritique$_{\text{GPT4}}$-F1.

4.5 Exp-IV: Pairwise Comparison

Q4: Which evaluation methods can choose the critiques that humans prefer?

Setup In this experiment, we utilize a range of scoring baselines to identify the better critique out of two critiques generated by LLMs. We calculate the agreement rate to evaluate the performance, defined as the consistency of the superior critique

Methods	Pearson		Spearman		Kendall's Tau	
	Hypo.h	Hypo.l	Hypo.h	Hypo.l	Hypo.h	Hypo.l
Single$_{\text{GPT4}}$	0.508	0.390	0.503	0.379	0.396	0.290
MetaCritique$_{\text{Open}}$-F1	0.730⋆	0.709⋆	0.745⋆	0.690⋆	0.566⋆	0.506⋆
MetaCritique$_{\text{GPT4}}$-P	0.355	0.667⋆	0.283	0.681⋆	0.227	0.504⋆
MetaCritique$_{\text{GPT4}}$-R	**0.844**†	0.831⋆	**0.837**†	0.830⋆	**0.681**†	0.649⋆
MetaCritique$_{\text{GPT4}}$-F1	0.841⋆	**0.886**†	0.836⋆	**0.899**†	0.675⋆	**0.724** †

Table 4: Correlation between different models with the gold scores. ⋆ means significantly ($p < 0.05$) outperforms the baseline method (Single$_{\text{GPT4}}$). †means significantly ($p < 0.05$) outperforms all methods.

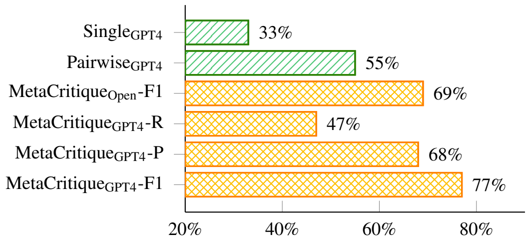


Figure 3: Agreement rate for pairwise comparison.

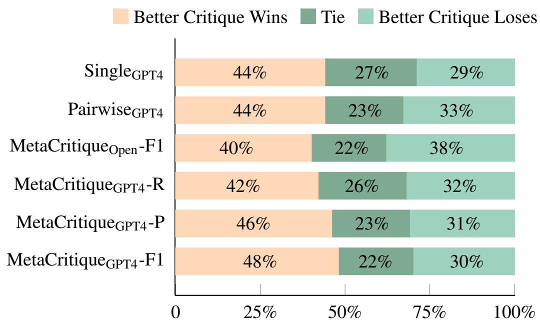
chosen by various methods with the gold standard critique determined by gold F1 score.

Results Figure 3 illustrates the agreement rate of various methods. Our MetaCritique$_{\text{GPT4}}$-F1 achieves the best performance. Especially, it exceeds the Single$_{\text{GPT4}}$ baseline by a considerable margin (**44%**). It implies that the scores of MetaCritique$_{\text{GPT4}}$-F1 are more comparable than Single$_{\text{GPT4}}$ baseline. Even MetaCritique$_{\text{Open}}$ significantly outperforms GPT-4 powered baselines Single$_{\text{GPT4}}$ and Pairwise$_{\text{GPT4}}$, confirming the effectiveness of our framework. Finally, precision and recall scores serve as complementary metrics for evaluating critiques, because MetaCritique$_{\text{GPT4}}$-P and MetaCritique$_{\text{GPT4}}$-R are somewhat less effective than MetaCritique$_{\text{GPT4}}$-F1.

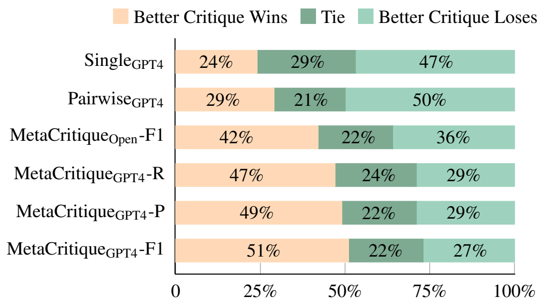
4.6 Exp V: Better Critique, Better Refinement

Q5: Can critique evaluations improve the alignment of existing LLMs?

Setup Critique is commonly applied to improve the quality of model outputs via refinement (Madaan et al., 2023). It is intuitive that superior critiques result in better refinements. To confirm this hypothesis, we conduct this experiment. Specifically, we instruct GPT-4 to refine the model outputs via the critique. Detailed instructions with demonstrations are presented in Table 18. Subsequently, we compare the refined outcomes to choose the better one. We conduct GPT-4 evaluation and human evaluation for this comparison. The prompt for GPT-4 evaluation is shown in Table 19, while the equivalent instruction for human evaluators is outlined in Table 20.



(a) GPT-4 Evaluation



(b) Human Evaluation

Figure 4: Win rates of refined results from superior critique over inferior critique. The left-hand models are used to choose the better critique. A larger yellow area means a more reliable critique.

Results We report the win rates of refined outcomes from superior critique over inferior critique, where the critique is evaluated by different methods. The outcomes of GPT-4 evaluation are depicted in Figure 4a, while the results of the human evaluation appear in Figure 4b. The superior critique chosen by MetaCritique$_{\text{GPT4}}$-F1 enhances the refinement significantly. Moreover, precision and recall scores are mutually supportive. Relying solely on one metric can lead to diminished performance.

Models	MetaCritique$_{GPT4}$		
	Precision	**Recall**	**F1 Score**
Human	**83.19**	60.65	64.02
GPT 3.5	80.79	64.27	68.72
SelFee	69.56	51.05	54.22
UltraCM	73.64	66.77	67.79
AUTO-J	<u>76.43</u>	**70.65**	**71.14**

Table 5: METACRITIQUE scores of critique models.

5 Exp VI: METACRITIQUE Leaderboard

Q6: How do various critique models perform?

We use METACRITIQUE (MetaCritique$_{GPT4}$-F1) to rank the critique models, such as **GPT 3.5**, **SelFee** (13B), **UltraCM** (13B), **AUTO-J** (13B), and **Human** annotators stemming from Shepherd dataset. As shown in Table 5, AUTO-J is the best open-source critique model, delivering more factual and comprehensive feedback than its open-source counterparts. In addition, human and GPT-3.5 achieve precision scores exceeding 80%, outshining the performance of open-source critique models. This finding highlights that the research of open-source critique models should pay more attention to factuality issues.

6 Related Work

6.1 Critique Evaluation

In light of the rapid advancements in LLMs, the significance of generating critique is increasingly acknowledged by researchers (Saunders et al., 2022; Wang et al., 2023b; Gou et al., 2023; Madaan et al., 2023; Kim et al., 2023; Ye et al., 2023; Wang et al., 2023a; Welleck et al., 2023; Li et al., 2023; Ke et al., 2023; Cui et al., 2023). However, the field is hindered by the scarcity of adequate research on critique evaluation. Critique evaluation aims to evaluate the quality of the critique. It mainly relied on the human annotators (Saunders et al., 2022; Wang et al., 2023a), which entails considerable costs and carries a substantial risk of subjectivity. Wang et al. (2023a) use GPT-4 to replace human annotators, but it still lacks transparency since the numeric score is produced directly via GPT-4 without any fine-grained calculation explanation.

6.2 Meta Evaluation

Meta evaluation is designed to assess automated metrics by determining the degree of correlation between automated scores and human evaluations (Zhang* et al., 2020; Yuan et al., 2021; Sai et al., 2022; Fu et al., 2023). This is achieved by utilizing correlation coefficients. The Spearman correlation (Zar, 2005) evaluates the monotonic relationship between two variables, focusing on their ranked values rather than raw data. Conversely, the Pearson correlation (Lee Rodgers and Nicewander, 1988; Mukaka, 2012) is used to gauge the linear relationship between two variables, employing the actual data values. Furthermore, Kendall's Tau (Kendall, 1938) is utilized to ascertain the ordinal association between two quantified variables. Lastly, the significance test (Williams., 1959; Koehn, 2004) serves as a crucial supplementary technique to measure the improved correlations.

6.3 Factuality Detection

Factuality detection aims to classify whether a textual statement, termed claim, is factual (Wang, 2017; Thorne et al., 2018; Augenstein et al., 2019; Wadden et al., 2020; Guo et al., 2022). Thorne et al. (2018) introduce the FEVER dataset to verify the given claim without related evidence, which leads to fact-checking models (Zhong et al., 2020; Krishna et al., 2022). Besides, Kamoi et al. (2023) classify whether a given claim can be entailed by the provided evidence. Similarly, in summarization tasks, FactCC (Kryscinski et al., 2020) and QAGS-based models (Wang et al., 2020) determine whether the produced summaries or summary sentences align factually with the provided document(s). Lastly, Gao et al. (2023) and Chern et al. (2023) utilize LLMs to detect factuality without giving any claim and evidence.

7 Conclusion

We are the pioneers in prioritizing critique evaluation and introducing the critique of critique, termed METACRITIQUE, involving two principles: *precision* and *recall*. Our METACRITIQUE is quantitative, reliable, and interpretable, wherein critiques are decomposed into AIUs, and two pragmatic tasks are established to calculate objective numeric scores along with plausible natural language rationale. We curate a meta-evaluation dataset containing various tasks to confirm the feasibility and effectiveness of METACRITIQUE. Experiments also show that superior critiques chosen by METACRITIQUE can result in a better refined outcome, suggesting that METACRITIQUE holds the potential to enhance generative artificial intelligence. Moreover, METACRITIQUE can aid in identifying the pros and cons of various critique models. This paper can facilitate future research in LLM critiques.

Acknowledgements

We thank Ziqiang Cao, Yiyuan Li, and Yuqing Yang for their valuable feedback and suggestions on the writing of this paper. We also thank the anonymous reviewers for their valuable feedback and helpful suggestions. This project is supported by Qingyuan Research Project and Shanghai Artificial Intelligence Laboratory, National Natural Science Foundation of China (62076212), and Research Grants Council of Hong Kong (PolyU/15203617 and PolyU/5210919).

Limitations

The creative tasks are not suitable for the recall principle, especially when there are multiple high-quality answers. It is intuitive that the suggestions for these tasks are various, so the critiques are also diverse. As a result, it is very hard to collect all effective critiques for the recall principle.

In future work, it is essential to expand upon exploring open-source large language models to achieve more robust and dependable evaluations. This also leads to reliability and numerical stability, through the transparent and verifiable process inherent in open-source developments.

Ethical Considerations

Our experimental data stems from Shepherd, which is well-established and publicly available. Dataset construction and annotation are consistent with the intellectual property and privacy rights of the original authors. The participants in our human evaluation were transparently informed of our research intent and were paid reasonable wages. All in all, this work complies with the ACL Ethics Policy[1].

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7432–7439.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

[1]https://www.aclweb.org/portal/content/acl-code-ethics

Human Language Technologies, Volume 1 (Long and Short Papers), pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. *arXiv preprint arXiv:2311.18702*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023a. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023b. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Marc Uri Porat. 1977. *The Information Economy: Definition and Measurement.* ERIC.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*.

Ehsan S Soofi. 1994. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023a. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Evan J. Williams. 1959. Regression analysis. *Wiley, New York*, 14.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

A Case Study

We present a case in Figure 5, which contains more details of Figure 1, in order to illustrate how to get our METACRITIQUE scores.

B Gold Score

According to human annotation, we calculate the gold scores of precision, recall, and F1 score at the AIU level and critique level. The results are shown in Table 6. We find that human-written critiques get higher precision scores yet lower recall scores than LLM-generated critiques. It indicates that LLMs are even more prone to creating fictional information, while human is hardly possible to make such mistakes. Also, it is worth noting that humans miss more key information than LLMs. This might be because human-written critiques contain fewer AIUs than LLM-generated critiques.

C API Usage at PyPI

We publish METACRITIQUE as a Python package at PyPI in order to allow users to assess the critique conveniently. Table 7 shows how to use METACRITIQUE in the Python code after installation.

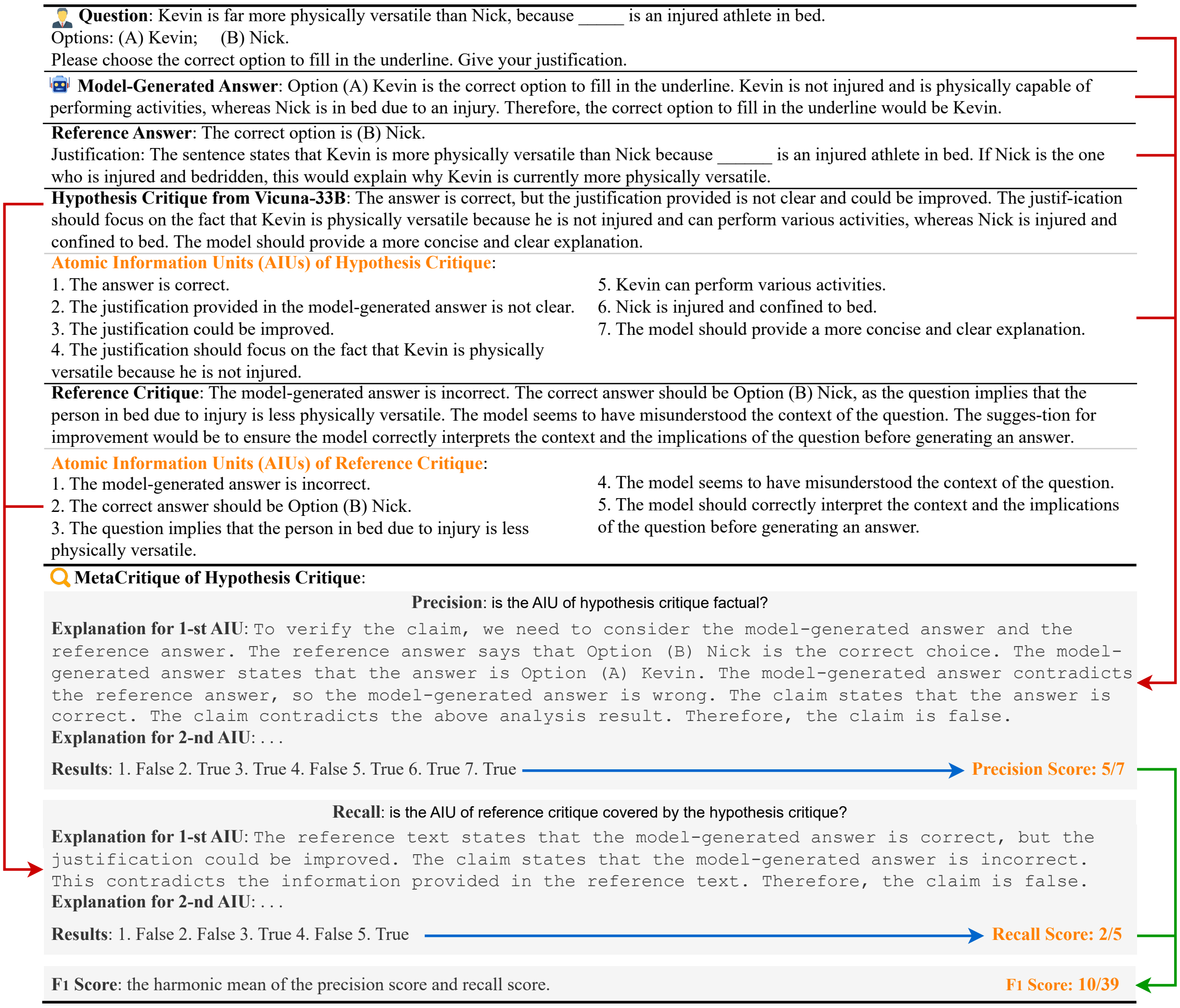


Figure 5: A case of METACRITIQUE for hypothesis critique evaluation.

	AIU Level (Micro)		
	Precision	Recall	F1 Score
Hypo.h	87.61	48.72	62.62
Hypo.l	71.85	53.28	61.19
	Critique Level (Macro)		
	Precision	Recall	F1 Score
Hypo.h	85.37	50.97	58.24
Hypo.l	71.07	54.37	58.20

Table 6: Gold scores of METACRITIQUE.

answer/critique, refined answer. For creating reference answers, the system prompt employed in our paper is "You are a brilliant AI assistant." for GPT-4. In addition, we used "claims" instead of "AIUs" in our prompts to reduce possible ambiguity and confusion for LLMs.

D Prompts and Instructions

We elaborate the prompts for GPT-4 generation, including generating hypothesis/reference critique, extracting AIUs, performing precision/recasll tasks, refining model-generated answer, implementing GPT-4 baselines (Single$_{\text{GPT4}}$ and Pairwise$_{\text{GPT4}}$), and evaluating refined answer. Additionally, we provide guidelines for human evaluation, covering generated outcomes such as AIUs, reference

```
from PyPI_package_name import MetaCritique
inputs = [
            {
              "question": "<question>",
              "response": "<response>",
              "hypothesis_critique": "<hypothesis_critique>"
            },
            {
              "question": "<question>",
              "response": "<response>",
              "hypothesis_critique": "<hypothesis_critique>"
            },
            ...
      ]

meta_critique_instance = MetaCritique(
      model_type = "gpt-4",
      batch_size = 5,
      api_key = ... ,  # here is your OpenAI key
      api_base = None,
      seed = None,
      cache_dir = "tmp_cache",
   )
precision, recall, f1_score = meta_critique_instance.score(inputs)
```

Table 7: Using METACRITIQUE in the Python code.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant. You will receive an input question and the model-generated answer. You need to generate the specific and actionable critiques for the model-generated answer, which contain the critical comments and helpful suggestions.

—————USER MESSAGE—————-

input question:
{question}
model-generated answer:
{model-generated answer}
critique:

Table 8: Prompt for generating reference critique.

You are a brilliant AI assistant. You will receive an input question and the model-generated answer. You need to generate the specific and actionable critiques for the model-generated answer, which contain the critical comments and helpful suggestions.
input question:
{question}
model-generated answer:
{model-generated answer}
critique:

Table 9: Prompt for generating hypothesis critique.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant.
You receive a critique as the input, which is the critical comment for an input question and a model-generated answer.
You need to generate all check-worthy claims of the critique.
A claim is an atomic statement that only contains a fine-grained information unit of a critique.
Each claim should be concise (a sentence) and self-contained.
Note that the 'answer' in the critique means the model-generated answer or the correct answer.
Some examples are as following:
[example 1 start]
input:
The answer violates basic principles of common sense. Flour is not sweet. Dusting it onto the bread would not sweeten the bread. Therefore, the right answer is to dust powdered sugar; sugar is, of course, sweet.
claims:
The model-generated answer violates basic principles of common sense that flour is not sweet.
Dusting Flour onto the bread would not sweeten the bread.
The correct answer is to dust powdered sugar.
Sugar is sweet.
[example 1 end]
[example 2 start]
input:
The output makes a logical error in the first bullet point of the answer, where it rejects the possibility of sunlight being the right answer. While sunlight might be produced in the sun, it doesn't stay there. Since sunlight comes to earth, it is on earth itself. Therefore sunlight, option 2, is the right answer as sunlight which is the oldest heat source on Earth.
claims:
The model-generated answer makes a logical error in rejecting the possibility of sunlight being the right answer.
Sunlight might be produced in the sun.
Sunlight doesn't stay in the sun.
Sunlight comes to earth.
Sunlight is on earth itself.
Sunlight, option 2, is the right answer as it is the oldest heat source on Earth.
[example 2 end]
[example 3 start]
input:
That the increase has "reached record levels" is not indicated in the context. Otherwise, the answer is adequate, except it should also mention the government is responding to the situation.
claims:
The content that the increase has "reached record levels" is not indicated in the context.
The correct answer should also mention the government is responding to the situation.
[example 3 end]
[example 4 start]
input:
The answer's summary was missing information about Andre Ward winning the rematch.
claims:
The model-generated answer was missing information about Andre Ward winning the rematch
[example 4 end]

—————USER MESSAGE—————-

input:
{critique}
claims:

Table 10: Prompt for AIUs extraction.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant. You receive an input question, a model-generated answer, the reference answer and a claim that is the critique for the model-generated answer. To verify the claim, you should think step by step as following:
1) you should consider the input question, the model-generated answer and the reference answer, and then you should determine what you need for verifying the claim. Please note that the input question is the most reliable, and the input question is always correct. The model-generated answer is the least reliable, and it can contain some flaws. The reference answer is more reliable than the model-generated answer, but sometimes it may contain some flaws.
2) verify whether the claim is true or false. If a claim is either a factual statement, explicitly mentioned in the needed information, or logically deducible from it, then the claim is true; otherwise, it is false.
Here are some examples:
[example 1 start]
<input question>
A shelf would fit in what part of a room? Please choose the correct option. Here are the options:
Option 1: garage
Option 2: corner
<model-generated answer>
Option 1
<reference answer>
Option 2
<claim>
The model-generated answer is wrong.
<verify claim>
To verify the claim, we need to consider the model-generated answer and the reference answer. The reference answer says that Option 2 is the correct choice. The model-generated answer states that the answer is Option 1. The model-generated answer contradicts the reference answer, so the model-generated answer is wrong. The claim states that the model-generated answer is wrong. The claim is consistent with the above analysis result. Therefore, the claim is true.
[example 1 end]
[example 2 start]
<input question>
Give a summary of the below article: Biggin Hill Airport wants to increase bank holiday and weekend opening times to increase the number of flights. Director William Curtis said he feared the airport could be asked to help the UK meet increased demand for air travel if the extension was not granted.
<model-generated answer>
Biggin Hill Airport intends to expand its hours to accommodate more flights.
<reference answer>
Biggin Hill Airport seeks to extend opening times to boost flights amidst concerns of meeting UK air travel demand.
<claim>
William Curtis wants to increase opening hours.
<verify claim>
To verify the claim, we need to consider the input question. The input question says that Biggin Hill Airport wants to increase opening times instead of William Curtis. The claim states that William Curtis wants to increase opening hours. The claim contradicts with the above reasoning result. Therefore, the claim is false.
[example 2 end]
[example 3 start]
<input question>
Context: He walked in the door and saw her pointing a gun with a laser pointer at his chest. Given the above context, How did he feel?
<model-generated answer>
He was glad.
<reference answer>
He was scared for his life.
<claim>
The correct answer is that he was afraid.
<verify claim>
To verify the claim, we need to consider the reference answer. The reference answer states that he was scared for his life, which is synonymous with being afraid. The claim states that the correct answer is that he was afraid. The claim is consistent with the above reasoning result. Therefore, the claim is true.
[example 3 end]
[example 4 start]
<input question>
Here is a premise: A black and white dog sliding down a slope on a beach.
Here is a hypothesis: A dog going uphill to the beach.
Does this premise imply the hypothesis? Please justify your answer:
<model-generated answer>
No, the dog could have been sliding down any slope, not just a sandy beach.
<reference answer>
No. The premise states that the dog is sliding down a slope on a beach, which indicates a downward movement. The hypothesis suggests that the dog is going uphill to the beach, which indicates an upward movement. These two scenarios are contradictory, so the premise does not support or imply the hypothesis.
<claim>
The model-generated answer provides a clear and concise explanation.
<verify claim>
To verify the claim, we need to consider the model-generated answer and the reference answer. The model-generated answer states that the reason is that the dog could have been sliding down any slope. However, it does not mention the direction of the dog's movement, which is the main point of contradiction. The claim states that the model-generated answer provides a clear and concise explanation. The claim contradicts with the above reasoning result. Therefore, the claim is false.
[example 4 end]

—————USER MESSAGE—————-

<input question>
{question}
<model-generated answer>
{model-generated answer}
<reference answer>
{reference answer}
<claim>
{AIU from hypothesis critique}
<verify claim>

Table 11: Prompt for performing the precision task.

——————SYSTEM MESSAGE——————-

You are a brilliant AI assistant. You receive a reference text and a claim. Note that the 'answer' in the reference text means the model-generated answer or the correct answer. You need to think step by step for verifying whether the claim is true or false. If a claim is either explicitly mentioned in the reference text or logically deducible from it, then the claim is true; otherwise, it is false.
Here are some examples:
[example 1 start]
<reference text>
The model-generated answer gives the wrong step. It is suggested to check the reasoning steps.
<claim>
The model-generated answer is wrong.
<verify claim>
The reference text points that the model-generated answer contains the wrong step, so the model-generated answer is wrong. The claim states that the model-generated answer is wrong. It is consistent with the above reasoning result. Therefore, the claim is true.
[example 1 end]
[example 2 start]
<reference text>
The model-generated answer is not right, because it provides the wrong analysis. The correct answer should be Option A.
<claim>
The correct answer should point out the invented fact.
<verify claim>
The reference text states that the correct answer should be Option A. The claim states that the correct answer should point out the invented fact. The claim is not mentioned or implied in the reference text. Therefore, the claim is false.
[example 2 end]
[example 3 start]
<reference text>
The answer lacks detail and context, like the age of the victim, the place of the incident and so on.
<claim>
The accurate summary should contain the age of the dead person.
<verify claim>
The reference text states that the model-generated answer lacks the detail, the age of victim. The claim states that the correct summary should contain the age of the dead person. It means that the model-generated answer is not mentioned the age of victim. The claim can be logically deducible from the reference text. Therefore, the claim is true.
[example 3 end]
[example 4 start]
<reference text>
The answer could be more concise and focused.
<claim>
The model-generated answer is mostly correct, but it could be improved by providing more specific details.
<verify claim>
The reference text states that the model-generated answer could be more concise. It means that the model-generated answer is elaborated. The claim states that the model-generated answer could be improved by providing more specific details. It means that the model-generated answer is brief. The claim contracts with the reference text. Therefore, the claim is false.
[example 4 end]

——————USER MESSAGE——————-

<reference text>
{hypothesis critique}
<claim>
{AIU from reference critique}
<verify claim>

Table 12: Prompt for performing the recall task.

Your task is to evaluate the generated reference answer (i.e., response for a question). Give a score 1-7 (worst-best) based on the quality of the answer.
7: Perfect. The response perfectly answer the query and provide suitable explanation.
6: Exceptional. The response perfectly answer the query and provide suitable explanation, but introduce some flaws.
5: Excellent. The response correctly answer the query and provide suitable explanation, but introduce some flaws.
4:Good. The response correctly answer the query but lacks suitable explanation.
3: Average. The response incorrectly answer the query but provides consistent explanation.
2: Poor. The response incorrectly answer the query but provides the explanation with flaws.
1: Extremely bad. The generated response is random text or simple repeats the question.
Give a score 1-3 for response with incorrect content and give a score 4-7 for response with correct content.

Table 13: Instruction for human evaluation of generated reference answer.

Your task is to evaluate the critique on a model-generated answer. Give a score 1-7 (worst-best) based on the quality of the critique.
7: When the answer is wrong, the critique clearly highlights the most important errors and provides very actionable suggestions. When the answer is correct, the critique confirms the answer is correct and provides very useful suggestions.
6: When the answer is wrong, the critique confirms that the answer is wrong and points out the most important errors. When the answer is correct, the critique confirms the answer is correct and provides useful suggestions.
5: When the answer is wrong, the critique misses the important errors but clearly confirms that the answer is wrong. When the answer is correct, the critique confirms the answer is correct and proposes some less useful suggestions.
4: The critique has a correct judgment of the answer (e.g., states correct answer is correct or states wrong answer is wrong).
3: The critique is vague about whether or not the answer is correct. Or the critique itself tries to answer the question regardless of the content in the answer.
2: The critique has a wrong judgment of the answer (e.g., states correct answer is wrong or states wrong answer is correct).
1: The critique is completely random text or simply repeats the answer.
First, please check whether the critique has correct or incorrect judgment (correct judgment means the answer is correct, critique confirms the correctness. Or if the answer is incorrect, the critique confirms the incorrectness.)
Give a score 1-3 for critique with incorrect judgment and give a score 4-7 for critique with correct judgment.

Table 14: Instruction for human evaluation of generated reference critique.

Your task is to evaluate the AIU extraction on a critique. Give a score 1-7 (worst-best) based on the quality of the extraction.
7: Perfect. Generate all salient claims without extra information. Useless or redundant information is removed.
6: Exceptional. Generate all salient claims without extra information but contains very few useless or redundant information.
5: Excellent. Generate all salient claims but introduce little extra information.
4:Good. Generate some salient claims but introduce little extra information.
3: Average. Generate some salient claims but introduce too much extra information.
2: Poor. Remove too many salient claims, or introduce too much extra information.
1: Extremely bad. The generated claims are random text or simple repeats the critique.

Table 15: Instruction for human evaluation of AIUs extraction.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant. You will receive a question, a model-generated answer, and two critiques about this answer. A good critique should point out key errors contained in the answer and provide constructive suggestions. Your task is to evaluate the quality of the critique. Your evaluation should consider factors such as the accuracy, factuality, comprehensiveness, relevance, and conciseness. Begin your evaluation by comparing the two critiques and provide a short explanation. Avoid any position biases and ensure that the order in which the answers were presented does not influence your decision. Do not allow the length of the critiques to influence your evaluation. Do not favor certain names of the answers. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if the critique A is better, "[[B]]" if the critique B is better, and "[[C]]" for a tie.

—————USER MESSAGE—————-

<input question>
{question}
<model-generated answer>
{model-generated answer}
<critique A>
{hypothesis critique from LLM 1}
<critique B>
{hypothesis critique from LLM 2}

Table 16: Prompt for pairwise comparison.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant. You will receive a question, a model-generated answer, and a critique about this answer. Your task is to evaluate the quality of the critique and give a score.
The score is based on the quality of the critique:
7: When the answer is wrong, the critique clearly highlights the most important errors and provides very actionable suggestions. When the answer is correct, the critique confirms the answer is correct and provides very useful suggestions.
6: When the answer is wrong, the critique confirms that the answer is wrong and points out the most important errors. When the answer is correct, the critique confirms the answer is correct and provides useful suggestions.
5: When the answer is wrong, the critique misses the important errors but clearly confirms that the answer is wrong. When the answer is correct, the critique confirms the answer is correct and proposes some less useful suggestions.
4: The critique has a correct judgement of the answer (e.g., states correct answer is correct or states wrong answer is wrong).
3: The critique is vague about whether or not the answer is correct. Or the critique itself tries to answer the question regardless of the content in the answer.
2: The critique has a wrong judgement of the answer (e.g., states correct answer is wrong or states wrong answer is correct).
1: The critique is completely random text or simply repeats the answer.
Begin your evaluation by considering the critique and provide a short explanation.
First, please check whether the critique has correct or incorrect judgment (correct judgment means the answer is correct, critique confirms the correctness. Or if the answer is incorrect, the critique confirms the incorrectness.) Please note that give a score 1-3 for critique with incorrect judgment and give a score 4-7 for critique with correct judgment.
Be as objective as possible. After providing your explanation, output your final score by strictly following this format: "[[score]]".

—————USER MESSAGE—————-

<input question>
{question}
<model-generated answer>
{model-generated answer}
<critique>
{hypothesis critique from LLM}

Table 17: Prompt for single critique rating.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant.
Based on the provided critique, please refine the model-generated answer to address all issues in the critique and ensure that all suggestions from the critique are incorporated to enhance the refined answer.
Here are some examples:
[example 1 start]
<input question>
A shelf would fit in what part of a room?
Here are the options:
Option 1: garage
Option 2: corner
Please choose the correct option and provide the justification.
<model-generated answer>
Option 1
<critique>
The model-generated answer is wrong and lacks the justification.
<refined answer>
Option 2.
Justification: "Corner" refers to a specific location within a room where two walls intersect, making it the appropriate choice for identifying a part of a room where a shelf could be installed.
[example 1 end]
[example 2 start]
<input question>
Context:
He walked in the door and saw her pointing a gun with a laser pointer at his chest.
Given the above context, How did he feel?
<model-generated answer>
He was pointed by a gun.
<critique>
The model-generated answer is a good starting point, but it should mention his feeling.
<refined answer>
He was scared when he saw a gun.
[example 2 end]

—————USER MESSAGE—————-

<input question>
{question}
<model-generated answer>
{model-generated answer}
<critique>
{hypothesis critique from LLM}
<refined answer>

Table 18: Prompt for refining model-generated answer using a critique.

—————SYSTEM MESSAGE—————-

You are a brilliant AI assistant. You can evaluate the quality of the model-generated answer for the input question. Your evaluation should consider factors such as the accuracy, factuality, relevance, and conciseness. Please note that details that do not clearly help to answer the input question are considered redundant. Begin your evaluation by comparing the two answers and provide a short explanation. Avoid any position biases and ensure that the order in which the answers were presented does not influence your decision. Do not allow the length of the answers to influence your evaluation. Do not favor certain names of the answers. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if the model-generated answer A is better, "[[B]]" if the model-generated answer B is better, and "[[C]]" for a tie.

—————USER MESSAGE—————-

<input question>
{question}
<model-generated answer A>
{refined answer from the LLM 1 critique}
<model-generated answer B>
{refined answer from the LLM 2 critique}

Table 19: Prompt for GPT-4 evaluation for a pair of refined answers.

Your task is to pick up the better answer from two given answers. Your evaluation should consider factors such as the accuracy, factuality, relevance, and conciseness. Please note that details that do not clearly help to answer the input question are considered redundant. Do not allow the length of the answers to influence your evaluation. Be as objective as possible. Label your final verdict: "A" if the model-generated answer A is better, "B" if the model-generated answer B is better, and "C" for a tie.

Table 20: Instruction for human evaluation for a pair of refined answers.