CAMIEval: Enhancing NLG Evaluation through Multidimensional Comparative Instruction-Following Analysis

Ziyue Fan[1,2] **Junliang He**[1,2] **Xiaoqing Li**[2] **Shaohui Kuang**[2]
Kai Song[2] **Yaqian Zhou**[1] **Xipeng Qiu**[1*]
[1]School of Computer Science, Fudan University
[2]ByteDance
`{zyfan22,jlhe22}@m.fudan.edu.cn` `xpqiu@fudan.edu.cn`

Abstract

With the rapid development of large language models (LLMs), due to their strong performance across various fields, LLM-based evaluation methods (LLM-as-a-Judge) have become widely used in natural language generation (NLG) evaluation. However, these methods encounter the following challenges: (1) distinguishing instruction-following ability, (2) being applicable across diverse NLG tasks, and (3) identifying low-quality outputs. To address these issues, we propose CAMIEVAL, a multidimensional comparative evaluation method based on instruction-following. Specifically, we define three fundamental dimensions of instruction-following: relevance, factuality, and adherence. Subsequently, we introduce a concrete Chain-of-Thoughts (ConcreteCoT) process to enhance the accuracy of evaluations. In addition, we trained a "regrettable model" REGRETLM to generate low-quality outputs, which helps the evaluator better identify the potential shortcomings of the candidate output by comparing low-quality outputs with reference outputs. Through this comparison, the evaluator can generate instruction-specific dimensions that complement the fundamental dimensions, forming a more comprehensive evaluation metric system. Experiments on two NLG evaluation benchmarks demonstrate that CAMIEVAL consistently outperforms existing methods in terms of correlation with human evaluations, providing a general and accurate framework for evaluating the outputs of LLMs.[1]

1 Introduction

Evaluating the quality of texts generated by language models has always been a challenging but essential task in natural language processing (NLP). An evaluation method more closely aligned with human judgment can provide effective feedback, including scores, explanations, and improvement suggestions. Such an approach will help better understand the actual performance of language models and promote their continuous improvement (Yuan et al., 2024).

In recent years, with the rapid advancement of large language models (LLMs) (OpenAI et al., 2024; Touvron et al., 2023; Bai et al., 2023; Jiang et al., 2023; Sun et al., 2024b), LLM-based evaluation methods, known as LLM-as-a-Judge (Zheng et al., 2023), have gradually become mainstream in natural language generation (NLG) evaluation (Gao et al., 2024; Li et al., 2024b). These methods leverage powerful LLMs for pairwise evaluation or pointwise scoring, following established evaluation processes or predefined evaluation criteria (Wang et al., 2023b; Liu et al., 2023b; Chiang and Lee, 2023a; Liu et al., 2023a; Guo et al., 2024; Lee et al., 2024). A range of studies have demonstrated that these LLM-based evaluation methods significantly outperform traditional n-gram-based (Papineni et al., 2002; Lin, 2004) or text-embedding-based methods (Kusner et al., 2015; Zhang et al., 2019). In many tasks, their performance matches or even surpasses that of human evaluation (Zheng et al., 2023; Wang et al., 2023a; Mendonça et al., 2023; Dubois et al., 2024).

However, as shown in Figure 1, the current evaluation methods exhibit significant shortcomings in evaluating the outputs of the ever-evolving LLMs: (1) Existing evaluation methods overlook the importance of evaluating instruction-following, as outputs that deviate from the given instructions are considered failures. Only those outputs that follow the instructions should proceed to further quality evaluation (Zeng et al., 2023; Qin et al., 2024). (2) Current evaluation dimensions (Hu et al., 2024; Ostheimer et al., 2023) are no longer effective in distinguishing the quality of outputs from the latest LLMs, referred to as **Dimension Bias**, as

---

[*] Corresponding author.

[1] We will release our code at `https://github.com/zyfan7/CAMIEval`

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 8708–8733
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

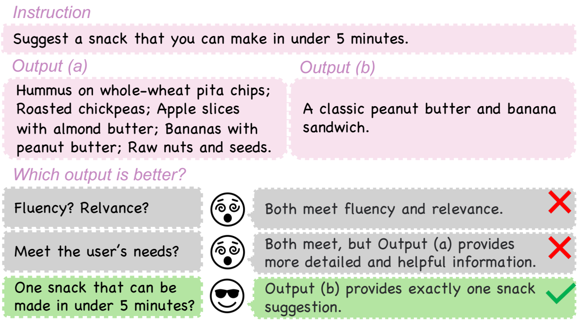
Instruction
Suggest a snack that you can make in under 5 minutes.
Output (a)
Hummus on whole-wheat pita chips; Roasted chickpeas; Apple slices with almond butter; Bananas with peanut butter; Raw nuts and seeds.
Output (b)
A classic peanut butter and banana sandwich.
Which output is better?
Fluency? Relvance?
Both meet fluency and relevance.
Meet the user's needs?
Both meet, but Output (a) provides more detailed and helpful information.
One snack that can be made in under 5 minutes?
Output (b) provides exactly one snack suggestion.

Figure 1: Comparison of current NLG evaluation methods. The first two methods encountered Dimension Bias and Description Bias, respectively.

these advanced models perform well across these dimensions. (3) There is a discrepancy between the evaluation dimension descriptions and the actual evaluation process, referred to as **Description Bias**. This makes it challenging for LLM-based evaluators to comprehend each dimension's meaning in specific instructions accurately (Zeng et al., 2023). (4) Diverse downstream tasks require meticulously defined evaluation criteria (Chiang and Lee, 2023b; Liu et al., 2023d), making various evaluation methods less generalizable and challenging to transfer to new tasks.

One of the critical challenges in this area is the inherent complexity of evaluating whether the output faithfully follows the instructions, which arises from the difficulty in defining clear criteria for what constitutes "instruction-following" for different instructions. LLM-based evaluators may become confused when asked to judge adherence without explicit standards (Sun et al., 2024a). Moreover, the limitations of current evaluation methods compound this complexity. These methods often rely on a single reference output, which makes it difficult for LLM-based evaluators to comprehensively capture the low-quality features of the candidate output, as they lack a clear example of what constitutes poor performance. Since LLMs can generate text that appears reasonable and coherent upon initial inspection, distinguishing between high-quality and low-quality outputs becomes particularly challenging, especially in pointwise scoring (Zheng et al., 2023).

To address these challenges, we must employ a more general and precise evaluation method to accurately capture the nuanced distinctions and actual performance of the model's outputs. In this paper, we introduce **CAMIEval**, a novel **C**omparative **A**utomated **M**ultidimensional **I**nstruction-following-based NLG **Eval**uation method. Firstly, from the perspective of instruction-following, we define three fundamental dimensions: *relevance*, *factuality*, and *adherence*. Meanwhile, we introduce a concrete Chain-of-Thoughts (Wei et al., 2022) process (**ConcreteCoT**) to mitigate the Description Bias. By concretizing these dimensions for each test instruction, the LLM-based evaluator can better comprehend the specific evaluation requirements, which significantly enhances the accuracy of the evaluation. Inspired by observations in pairwise evaluation, for pointwise scoring evaluation, we additionally trained a "regrettable model" **RegretLM** by employing a reversed DPO (Rafailov et al., 2024) method to capture the low-quality features of model outputs comprehensively. RegretLM generates text that follows basic instructions but, regrettably, is not of high quality, embodying common weaknesses of LLMs' outputs to serve as negative examples for reference. Starting from the test instructions, LLM-based evaluators are able to more clearly identify the potential shortcomings of the candidate output by comparing low-quality outputs with reference outputs. Through this comparison, the evaluators can generate instruction-specific dimensions that complement the fundamental dimensions, forming a more comprehensive evaluation metric system. Moreover, this approach mitigates the Dimension Bias, which in turn enhances the consistency and accuracy of the evaluation.

We conducted comprehensive experiments on two NLG evaluation benchmarks: one benchmark for pairwise evaluating instruction-following performance and another for pointwise evaluating the performance of instruction controllable summarization. The results indicate that our method typically outperforms existing NLG evaluation methods in terms of correlation with human evaluations.

To summarize, the main contributions of our work are as follows:

1. We identified significant issues in current LLM-based evaluation methods, including the oversight of evaluating instruction-following, the Dimension Bias, the Description Bias, and the requirement for manually and meticulously defining task-specific evaluation criteria.

2. We introduce CAMIEval, a novel multi-dimensional comparative evaluation method

focused on instruction-following, addressing critical shortcomings in current NLG evaluation methods.

3. We conducted comprehensive experiments on two NLG evaluation benchmarks, demonstrating that our method consistently outperforms existing NLG evaluation methods in terms of correlation with human evaluations.

2 Related Work

2.1 Automatic NLG Evaluation

Due to the inefficiency, labor-intensiveness, and lack of reproducibility inherent in manual evaluation, numerous early studies have explored various automated NLG evaluation methods, primarily including N-gram-based and embedding-based metrics.

N-gram-based Metrics. N-gram-based metrics evaluate the lexical overlap between candidate and reference texts. BLEU (Papineni et al., 2002), a widely adopted metric for evaluating machine translation quality, calculates the n-gram overlap between the candidate translations while incorporating a length penalty mechanism to generate a score—the higher the score, the better the translation quality. ROUGE (Lin, 2004), a metric for evaluating the quality of text summaries, measures the coverage and consistency of the summary by calculating the overlap in n-grams, word sequences, and longest common subsequences between the candidate and the reference.

Embedding-based Metrics. Embedding-based metrics utilize word or sentence embeddings to evaluate the semantic similarity between candidate and reference text. WMD (Kusner et al., 2015) employs word embedding techniques like Word2Vec (Mikolov, 2013) to represent words as high-dimensional vectors, subsequently calculating the minimum transportation cost between the candidate text and the reference text. WMD quantifies the similarity between texts by minimizing this cost, ultimately producing a similarity score. BERTScore (Zhang et al., 2019), on the other hand, uses the BERT model (Devlin, 2018) to generate word embeddings and calculates the cosine similarity for each word pair between the candidate and reference texts. These individual similarities are then synthesized to derive a final score.

These methods were practical and effective at the time. However, they primarily relied on superficial pattern matching, potentially failing to adequately capture the semantic and contextual subtleties of the generated text. Therefore, they have significant limitations aligning with human evaluation results (Sun et al., 2022).

2.2 LLM-as-a-Judge

With the rapid advancement of LLMs and due to their powerful capabilities in natural language understanding and reasoning, LLM-based evaluation methods have become the mainstream approach for NLG evaluation. Zheng et al. (2023) proposed MT-Bench, which utilizes LLMs to directly evaluate candidate texts, exploring both pairwise comparison and single answer grading. Liu et al. (2023b) introduced G-Eval, a CoT-based evaluation method employing a standardized evaluation process. By using GPT-4 as the evaluator, it scores the outputs of models on various NLG tasks. Based on the methods mentioned above, techniques such as probability-weighted calibration (Liu et al., 2023b), multi-agent communication (Chan et al., 2023), and batch evaluation of multiple test instruction outputs (Yuan et al., 2023) have been integrated. These LLM-as-a-Judge methods have shown considerable performance in evaluating traditional NLG tasks like text summarization and dialogue generation, rivaling or even surpassing human-level agreement in some cases.

However, a straightforward evaluation or simply incorporating CoT for assessment is inadequate for the various downstream tasks, such as evaluating instruction-following. Liu et al. (2023a) proposed AlignBench, which introduces a rule-calibration judging approach that aligns more closely with human scoring habits. They also manually set different evaluation dimensions for various types of questions, utilizing multidimensional analysis to guide LLMs in scoring. However, this method relies on meticulously crafted evaluation dimensions for each type of question and is susceptible to Description Bias, leading to failures in adversarial evaluation tasks. Additionally, some studies have proposed utilizing LLM to directly generate multiple evaluation dimensions or questions based on the test instructions, achieving commendable performance across various evaluation tasks (Gong and Mao, 2023; Li et al., 2024a). However, these generated evaluation dimensions tend to be overly general and lack specificity, failing to distinguish the quality of current LLMs' output effectively. Additionally, these evaluation methods often rely on a

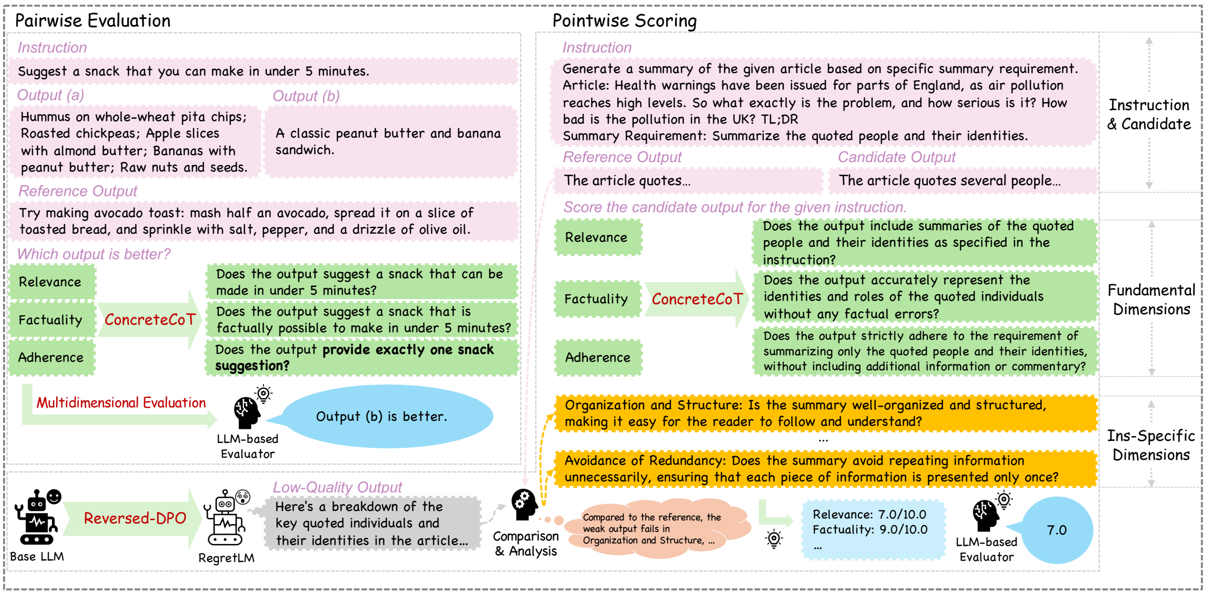
Pairwise Evaluation
Instruction
Suggest a snack that you can make in under 5 minutes.
Output (a)
Hummus on whole-wheat pita chips; Roasted chickpeas; Apple slices with almond butter; Bananas with peanut butter; Raw nuts and seeds.
Output (b)
A classic peanut butter and banana sandwich.
Reference Output
Try making avocado toast: mash half an avocado, spread it on a slice of toasted bread, and sprinkle with salt, pepper, and a drizzle of olive oil.
Which output is better?
Relevance
Factuality
Adherence
ConcreteCoT
Does the output suggest a snack that can be made in under 5 minutes?
Does the output suggest a snack that is factually possible to make in under 5 minutes?
Does the output provide exactly one snack suggestion?
Multidimensional Evaluation
LLM-based Evaluator
Output (b) is better.
Pointwise Scoring
Instruction
Generate a summary of the given article based on specific summary requirement. Article: Health warnings have been issued for parts of England, as air pollution reaches high levels. So what exactly is the problem, and how serious is it? How bad is the pollution in the UK? TL;DR Summary Requirement: Summarize the quoted people and their identities.
Reference Output
The article quotes…
Candidate Output
The article quotes several people…
Instruction & Candidate
Score the candidate output for the given instruction.
Relevance
Factuality
Adherence
ConcreteCoT
Does the output include summaries of the quoted people and their identities as specified in the instruction?
Does the output accurately represent the identities and roles of the quoted individuals without any factual errors?
Does the output strictly adhere to the requirement of summarizing only the quoted people and their identities, without including additional information or commentary?
Fundamental Dimensions
Organization and Structure: Is the summary well-organized and structured, making it easy for the reader to follow and understand?
…
Avoidance of Redundancy: Does the summary avoid repeating information unnecessarily, ensuring that each piece of information is presented only once?
Ins-Specific Dimensions
Base LLM
Reversed-DPO
RegretLM
Low-Quality Output
Here's a breakdown of the key quoted individuals and their identities in the article…
Comparison & Analysis
Compared to the reference, the weak output fails in Organization and Structure, …
Relevance: 7.0/10.0
Factuality: 9.0/10.0
…
LLM-based Evaluator
7.0

Figure 2: Overall illustration of CAMIEVAL: 1) Given an instruction, candidate outputs, and a reference output, CAMIEVAL generates instruction-based concrete descriptions of fundamental dimensions through ConcreteCoT. 2) For pointwise scoring, we trained a REGRETLM capable of producing relatively low-quality outputs that reflect common weaknesses of current LLMs. By comparing low-quality outputs with reference outputs, evaluators can generate instruction-specific (ins-specific) dimensions that complement the fundamental dimensions. 3) Conducting multidimensional evaluation across these dimensions.

single reference answer, which may mislead evaluators by the apparent high quality of the model's output, making it challenging to capture low-quality characteristics comprehensively. Therefore, we present CAMIEVAL to address these issues.

3 Methodology

CAMIEVAL is a multidimensional NLG evaluation method based on comparative analysis, beginning with instruction-following evaluation. In this section, we will provide a detailed introduction to CAMIEVAL, as shown in Figure 2.

3.1 Task Definition

In terms of the evaluation tasks currently used in NLG, there are two primary forms:

Pairwise Evaluation. Given a test instruction I, the evaluator needs to determine which of the two candidate outputs, O_1 or O_2, is better or declare the two as a tie, implying that both are equally preferred or neither is preferred:

$$p = \text{LLM}(O_1, O_2; I), \tag{1}$$

where p represents the evaluation result, indicating which output is better or declaring a tie.

Pointwise Scoring. Given a test instruction I, the evaluator needs to score the candidate output O within the given scoring range $\mathbb{S}$ (e.g., $\mathbb{S} = [1, 10]$):

$$s = \text{LLM}(O; I), \tag{2}$$

where $s \in \mathbb{S}$ represents the score.

In CAMIEVAL, from the perspective of instruction-following, we introduce three fundamental dimensions, denoted as $\mathcal{D}_{\text{funda}} = \{d_1, d_2, d_3\}$, along with N instruction-specific dimensions, represented as $\mathcal{D}_{\text{specific}} = \{d'_i\}_{i=1}^{N}$. The three fundamental dimensions are as follows:

1. *Relevance*: Does the output content directly relate to the question of the instruction?
2. *Factuality*: Does the output of the instruction not contain factual errors?
3. *Adherence*: Does the output of the instruction strictly follow the specific requirements of the instruction?

Some examples are provided in Appendix E, which demonstrate the specific forms of these dimensions in the given instructions.

3.2 Pairwise Evaluation

ConcreteCoT. For each sample (I, O_1, O_2) within the dataset, to mitigate the Description Bias, we use a LLM to generate concrete descriptions corresponding to each of the three fundamental dimensions $\mathcal{D}_{\text{funda}} = \{d_1, d_2, d_3\}$. This results in a set of concrete descriptions $\mathcal{C}_{\text{funda}} = \{c_1, c_2, c_3\}$, where

$$c_i = \text{LLM}(d_i; I), i = 1, 2, 3. \tag{3}$$

In order to ensure a more accurate evaluation, if the test set does not provide a reference output corresponding to the instructions, a strong LLM can be employed to generate a reference output:

$$O_{\text{ref}} = \text{LLM}(I). \tag{4}$$

Multidimensional Evaluation. Using these results, we can enable the LLM-based evaluator to determine which output is better for the instruction I:

$$p = \text{LLM}(O_1, O_2; I, O_{\text{ref}}, \mathcal{D}_{\text{funda}}, \mathcal{C}_{\text{funda}}). \tag{5}$$

3.3 Pointwise Scoring

ConcreteCoT. Similar to pairwise evaluation, for each sample (I, O) in the dataset, we can obtain concretized descriptions $\mathcal{C}_{\text{funda}}$ corresponding to the three predefined fundamental dimensions $\mathcal{D}_{\text{funda}}$, as well as a reference output O_{ref} generated by a strong LLM.

REGRETLM Training. Merely assessing basic instruction-following performance in pointwise evaluation tasks is insufficient. To achieve a more accurate evaluation, we need to incorporate additional dimensions that are closely related to the instructions. However, instructing the LLM to generate evaluation dimensions directly from the given instructions often results in overly general dimensions (Gong and Mao, 2023; Li et al., 2024a). Consequently, this leads to misleading evaluators with superficial quality and makes it difficult to capture the low-quality features of the model's output comprehensively. If one can obtain an output with typical low-quality features and then compare it to a reference output, an LLM-based evaluator can effortlessly identify the critical low-quality characteristics, abstracting them into additional evaluation dimensions. Therefore, it is imperative that we understand what constitutes low-quality output and how to obtain such output.

Coincidentally, we discovered that direct preference optimization (DPO, Rafailov et al., 2024) serves as the antithesis to our objective. DPO utilizes preference datasets to enhance the model's output to align with human preferences. It increases the log probability of preferred responses and decreases the log probability of non-preferred responses, ensuring higher evaluation scores for the model's output. In contrast, our goal is to capture the low-quality features of the output. To achieve this, we need to acquire a relatively weak output that receives lower scores in the evaluation.

Building upon this observation, we perform reversed-DPO on an LLM, aiming to train a "regrettable model" REGRETLM specifically designed to generate relatively weak outputs. Given the preference dataset $\mathbf{D} = \{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)\}$, where $\mathbf{y}_w$ and $\mathbf{y}_l$ are responses to $\mathbf{x}$ and marked as "win" and "lose" by human respectively, the loss function can be represented as:

$$\mathcal{L}_{\text{reversed-DPO}}(\pi_\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathbf{D}} \left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l\mid\mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w\mid\mathbf{x})}\right)\right)\right], \tag{6}$$

where we consider an LLM as a policy $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ parameterized by θ, π_{ref} is the reference model used for regularizing π_θ with Kullback-Leibler divergence.

By generating relatively low-quality outputs based on given instructions that exhibit common weaknesses of current LLM's outputs, the LLM-based evaluators will have both a positive example (reference output) and a negative example (relatively weak but essentially correct output) for comparison. This enables the LLM-based evaluators to identify where the weak output falls short compared to the reference output, allowing them to generate evaluation dimensions that are more distinct and effective in distinguishing between high-quality and low-quality outputs, while remaining precisely aligned with the given instructions.

Contrastive Analysis. Utilizing REGRETLM we have obtained, we can derive a relatively weak output O_{weak} for the given instruction I:

$$O_{\text{weak}} = \text{REGRETLM}(I). \tag{7}$$

With the relatively low-quality output O_{weak}, we can leverage the LLM to compare it with the reference output O_{ref}. By integrating the instruction I, generate N instruction-specific dimensions $\mathcal{D}_{\text{specific}} = \{d'_i\}_{i=1}^{N}$ and their corresponding concrete descriptions $\mathcal{C}_{\text{specific}} = \{c'_i\}_{i=1}^{N}$, where

$$(d'_i, c'_i) = \text{LLM}(O_{\text{weak}}; I, O_{\text{ref}}, \mathcal{D}_{\text{funda}}, \mathcal{C}_{\text{funda}}), \quad i = 1, \ldots, N. \tag{8}$$

Methods	Natural	Adversarial					Average
		Neighbor	GPTInst	GPTOut	Manual	Average	
GPT-4o							
Vanilla	95.0	66.4	81.5	72.3	73.9	73.5	77.8
Vanilla*	94.0	76.1	82.6	74.5	76.1	77.3	80.7
CoT	94.0	76.9	84.8	66.0	71.7	74.9	78.7
CoT*	97.0	81.3	90.2	76.6	78.3	81.6	84.7
Metrics	93.0	85.1	90.2	66.0	82.6	81.0	83.4
Metrics*	97.0	**85.8**	89.1	68.1	84.8	82.0	85.0
CAMIEval	95.0	80.6	90.2	78.7	87.0	84.1	86.3
CAMIEval*	**98.0**	79.9	**92.4**	**83.0**	**91.3**	**86.7**	**88.9**
Qwen2.5-72B-Instruct							
Vanilla	88.0	45.5	66.3	74.5	63.0	62.3	67.5
Vanilla*	89.0	58.2	72.8	**76.6**	63.0	67.7	71.9
CoT	89.0	67.2	73.9	66.0	67.4	68.6	72.7
CoT*	91.0	71.6	77.2	74.5	71.7	73.8	77.2
Metrics	92.0	64.2	75.0	70.2	67.4	69.2	73.8
Metrics*	94.0	70.1	80.4	74.5	69.6	73.7	77.7
CAMIEval	94.0	**76.1**	81.5	72.3	73.9	76.0	79.6
CAMIEval*	**97.0**	75.4	**88.0**	74.5	**80.4**	**79.6**	**83.1**

Table 1: Accuracies with GPT-4o and Qwen2.5-72B-Instruct on LLMBar. * indicates the use of a reference. The highest accuracy in each column is highlighted in **bold**.

Multidimensional Evaluation. With these results, we can let the LLM-based evaluator score the candidate output O:

$$s = \mathrm{LLM}(O; I, O_{\text{ref}}, \mathcal{D}, \mathcal{C}), \tag{9}$$

where $s \in \mathbb{S}$ represents the score, $\mathcal{D} = \mathcal{D}_{\text{funda}} \cup \mathcal{D}_{\text{specific}}$, $\mathcal{C} = \mathcal{C}_{\text{funda}} \cup \mathcal{C}_{\text{specific}}$.

4 Experiment

We evaluate CAMIEval on two benchmarks, LLMBar (Zeng et al., 2023) and InstruSum (Liu et al., 2023c), which respectively represent pairwise evaluation of instruction-following and pointwise scoring of instruction controllable summarization.

4.1 Experimental Setup

We choose to utilize GPT-4o (*2024-05-13*) as our evaluator in CAMIEval, setting the temperature to 0 to ensure reproducibility. Additionally, we use Qwen2.5-72B-Instruct (Qwen Team, 2024), one of the most powerful open-source LLMs, with greedy decoding for consistent results. For RegretLM, we choose Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) as the base model and use the Infinity-Preference dataset[2] for reversed-DPO training. We adopt and modify the prompt templates from LLMBar and AlignBench for the final evaluation phase, with LLMBar's templates used for pairwise evaluation and AlignBench's templates used for pointwise scoring. All prompt templates used in CAMIEval are provided in the Appendix D.

4.2 Benchmarks

LLMBar (Zeng et al., 2023) is a dataset designed to evaluate evaluators' discernment ability in instruction-following. It contains 419 pairwise samples, divided into Natural (100) and Adversarial (319) subsets. The Natural set is refined from benchmarks such as AlpacaFarm (Dubois et al., 2024) and LLMEval[2] (Zhang et al., 2023). The Adversarial set includes outputs that appear high-quality but lack correct answers, categorized as Neighbor, GPTInst, GPTOut, and Manual. Since the LLMBar dataset does not include cases of a tie, we use accuracy as the evaluation metric, which is consistent with the approach of LLMBar.

InstruSum (Liu et al., 2023c) is a dataset for evaluating instruction controllable summary generation, containing 400 samples. Each sample consists of an article, specified summary requirements, and model output. We use the summaries in the *hybrid* subset as reference summaries. We select the *overall* criterion and use Pearson, Spearman, and Kendall-Tau correlations as metrics to assess the performance of the evaluation methods, similar to G-Eval.

---

[2] `https://huggingface.co/datasets/BAAI/Infinity-Preference`

Methods	INSTRUSUM					
	GPT-4o			Qwen2.5-72B-Instruct		
	r	ρ	τ	r	ρ	τ
Vanilla	0.438	0.404	0.325	0.388	0.382	0.313
Vanilla*	0.517	0.495	0.387	0.440	0.434	0.354
CoT	0.417	0.369	0.288	0.359	0.358	0.285
CoT*	0.517	0.496	0.393	0.443	0.436	0.349
G-Eval	0.368	0.337	0.271	0.296	0.286	0.239
G-Eval*	0.376	0.339	0.273	0.399	0.389	0.324
Metrics*	0.523	0.512	0.379	0.422	0.309	0.303
CAMIEVAL$^*_{base}$	0.509	0.481	0.357	0.484	0.464	0.349
CAMIEVAL*	**0.572**	**0.557**	**0.413**	**0.505**	**0.492**	**0.369**

Table 2: Pearson (r), Spearman (ρ), and Kendall-Tau (τ) correlations with summary-level human judgments on InstruSum. * indicates the use of a reference. The highest correlation in each column is highlighted in **bold**. CAMIEVAL$^*_{base}$ indicates the use of only the fundamental dimensions.

4.3 Baselines

We compare CAMIEVAL with several current LLM-as-a-Judge evaluation methods.

Vanilla. We instruct the LLM to select better outputs or provide an evaluation score, where the LLM is required to directly output its evaluation results without any explanation.

Chain-of-Thoughts. Instead of directly generating evaluation results, we instruct the LLM to generate a step-by-step reasoning process before judgment.

Self-Generated Metrics (Metrics). This approach prompts the LLM to generate a set of instruction-specific evaluation metrics that represent various dimensions essential for evaluating the corresponding output. During the evaluation phase, these metrics are provided to the LLM-based evaluator. This method aligns with those previously proposed in pairwise evaluation (Zeng et al., 2023) and pointwise scoring (Gong and Mao, 2023; Li et al., 2024a).

G-Eval. In this method (Liu et al., 2023b), the LLM is instructed to evaluate based on the established evaluation process, which includes a predefined explanation of the evaluation criteria and detailed self-generated evaluation steps. G-Eval applies only to a single downstream task, as it requires manually defining customized evaluation criteria for different NLG tasks. Therefore, this approach is employed in the pointwise scoring task.

4.4 Results for Pairwise Evaluation

The first section of Table 1 illustrates the evaluation performance with GPT-4o as the evaluation model, while the second section illustrates the performance with Qwen2.5-72B-Instruct.

As shown in Table 1, CAMIEVAL achieved an accuracy of 88.9% on the LLMBar dataset, significantly surpassing existing evaluation methods. Specifically, CAMIEVAL-GPT-4o demonstrated an improvement of 3.9 points (4.6%) in accuracy compared to the current optimal method, and for CAMIEVAL-Qwen, this enhancement reached 5.4 points (6.9%).

4.5 Results for Pointwise Scoring

Similar to the pairwise evaluation scenario, we demonstrated the evaluation performance of LLMs using GPT-4o and Qwen2.5-72B-Instruct, respectively. As shown in Table 2, CAMIEVAL-GPT-4o achieved Pearson, Spearman, and Kendall-Tau correlations of 0.572, 0.557, and 0.413 with humans across the InstruSum benchmark, markedly surpassing existing evaluation methods. Likewise, for CAMIEVAL-Qwen, the correlations also outperformed other evaluation methods.

Note that the evaluation performance showed a significant decline without the instruction-specific dimensions. It indicated that in pointwise scoring scenarios, fundamental dimensions are insufficient, and instruction-specific dimensions need to be introduced to form a comprehensive evaluation metric system.

5 Analysis

The Effect of ConcreteCoT. To mitigate Description Bias, we introduce the ConcreteCoT process to concretize the fundamental dimensions for evaluating the instruction-following performance. We compared the performance without using ConcreteCoT in pairwise evaluation, as shown in Table 3.

Methods	Nat.	Adv.	Average
CAMIEval-GPT-4o	**98.0**	**86.7**	**88.9**
w/o ConcreteCoT	97.0	84.4	86.9
CAMIEval-Qwen	**97.0**	**79.6**	**83.1**
w/o ConcreteCoT	94.0	78.2	81.3

Table 3: Comparison of CAMIEval without ConcreteCoT in pairwise evaluation. The highest accuracies are in **bold**. Nat. denotes the LLMBar-Nature subset, while Adv. denotes the LLMBar-Adversarial subset.

On average, CAMIEval-GPT-4o dropped by 2.0 points (2.2%), and CAMIEval-Qwen dropped by 1.8 points (2.2%). Compared to using general descriptions, we observe that the concrete descriptions obtained after ConcreteCoT identify the reference points needed during evaluation, thus reducing the reasoning difficulty and evaluation cost for LLMs and achieving better evaluation performance.

Methods	InstruSum		
	r	ρ	τ
CAMIEval-GPT-4o	**0.572**	**0.557**	**0.413**
w/ predefined	0.539	0.527	0.395
CAMIEval-Qwen	**0.505**	**0.492**	**0.369**
w/ predefined	0.425	0.429	0.331

Table 4: Comparison of CAMIEval with predefined dimensions from SummEval in pointwise scoring. The highest correlations are in **bold**.

The Effect of Dimensions in CAMIEval. In Section 1 and 2, we discussed that manually defining evaluation dimensions is a labor-intensive process. Furthermore, the currently predefined evaluation dimensions fail to distinguish the quality of outputs from current LLMs and are inapplicable to various downstream NLG tasks (Dimension Bias). To address these issues, we introduced a two-stage approach to generate and concretize fundamental and instruction-specific dimensions. To validate the effectiveness of our approach, we adopted the predefined dimensions of *Coherence*, *Consistency*, *Fluency*, and *Relevance*, as proposed in SummEval (Fabbri et al., 2021), and applied them in the InstruSum benchmark. As demonstrated in Table 4, these predefined dimensions fail to accurately evaluate the quality of the outputs compared to our method.

Methods	InstruSum		
	r	ρ	τ
CAMIEval-GPT-4o	**0.572**	**0.557**	**0.413**
w/o RegretLM	0.544	0.526	0.386
w/o comparison	0.539	0.522	0.380
CAMIEval-Qwen	**0.505**	**0.492**	**0.369**
w/o RegretLM	0.462	0.457	0.342
w/o comparison	0.431	0.415	0.305

Table 5: Comparison of CAMIEval in pointwise evaluation under two conditions: (1) evaluation with normal outputs, and (2) evaluation without comparison. The highest correlations are in **bold**.

The Effect of RegretLM. In order to capture the low-quality features of LLM's outputs and obtain evaluation dimensions that better align with instructions, we trained RegretLM designed to generate relatively low-quality outputs. It allows LLM-based evaluators to refine the evaluation metric system through comparative analysis. To verify the effectiveness of RegretLM, we: (1) used the base model of RegretLM, Mistral-7B-Instruct-v0.3, to generate outputs corresponding to the instructions, which were subsequently used to derive instruction-specific dimensions; (2) prompted the LLM to directly generate new evaluation dimensions based on the content of the instructions and the fundamental dimensions. As shown in Table 5, evaluation performance declined to varying degrees under the two different ablation settings, demonstrating the effectiveness of RegretLM.

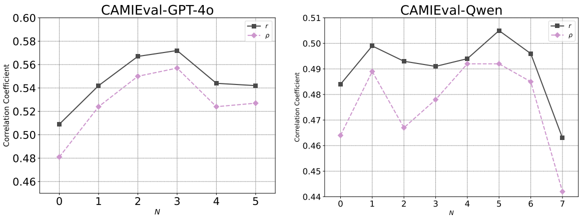

Figure 3: Comparison of instruction-specific dimensions for different values of N, where $N = 0$ represents only fundamental dimensions.

The Effect of Dimension Quantity. We explore how the number of instruction-specific dimensions affects evaluation performance in pointwise scoring. As shown in Figure 3, when using GPT-4o as the evaluation model, the highest Pearson and Spearman correlation coefficients are achieved at $N = 3$, while for Qwen2.5-72B-Instruct, the optimal values are at $N = 5$. Full experimental results are provided in Appendix B.2. Overall, evaluation performance initially increases and then decreases with N; therefore, choosing a moderate value is beneficial for better evaluation.

6 Conclusion

In this paper, we introduced CAMIEval, a novel evaluation method for natural language generation that addresses critical shortcomings in current approaches, such as Dimension Bias, Description Bias, and insufficient focus on instruction-following. By defining fundamental dimensions and leveraging a ConcreteCoT process, along with a "regrettable model" RegretLM to generate low-quality outputs for comparative analysis, CAMIEval provides a general and precise evaluation framework. Experimental results on the LLMBar and InstruSum benchmarks demonstrate that CAMIEval significantly outperforms existing methods, achieving higher correlations with human evaluations and offering a versatile solution for evaluating the performance of LLMs.

Limitations

CAMIEval requires that the LLM-based evaluator possesses a sufficient instruction-following ability. Without this ability, it may fail to understand the prompts, leading to an inability to follow the specified output format or producing inaccurate evaluation results. Therefore, it is unsuitable for relatively weak models (e.g., 7B LLMs). In addition, our method introduces a detailed intermediate reasoning process, which makes the context length longer. Consequently, the model is required to have certain capabilities for handling long contexts.

Ethics Statement

CAMIEval is based on existing advanced LLMs, which effectively adhere to the HHH principle (Helpful, Honest, and Harmless, Askell et al., 2021) and refuse to respond to potentially harmful instructions. Additionally, all the training datasets and benchmarks used in our approach do not contain harmful information. Therefore, we believe that our approach is sufficiently safe and poses no potential ethical concerns.

Acknowledgements

We are sincerely grateful to all the reviewers and chairs for their detailed feedback and insightful comments. In particular, we would like to express our heartfelt thanks to Dr. Qian Wang for his invaluable assistance during the early stages of this research. This work was supported by the National Key Research and Development Program of China (No.2022CSJGG0801). The computations in this research were performed using the CFFF platform of Fudan University.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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*.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Peiyuan Gong and Jiaxin Mao. 2023. Coascore: Chain-of-aspects prompting for nlg evaluation. *arXiv preprint arXiv:2312.10355*.

Fang Guo, Wenyu Li, Honglei Zhuang, Yun Luo, Yafu Li, Qi Zhu, Le Yan, and Yue Zhang. 2024. Generating diverse criteria on-the-fly to improve point-wise llm rankers. *arXiv preprint arXiv:2404.11960*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? *arXiv preprint arXiv:2402.12055*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

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.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *arXiv preprint arXiv:2403.18771*.

Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy F Chen, and Min-Yen Kan. 2024a. Decompose and aggregate: A step-by-step interpretable evaluation framework. *arXiv preprint arXiv:2405.15329*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for nlg evaluation: Advances and challenges.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023a. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023c. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023d. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.

John Mendonça, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation. *arXiv preprint arXiv:2308.16797*.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. Text style transfer evaluation using large language models. *arXiv preprint arXiv:2308.13577*.

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.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. 2024a. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. Bertscore is unfair: On social bias in language model-based metrics for text generation. *arXiv preprint arXiv:2210.07626*.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, et al. 2024b. Moss: An open conversational large language model. *Machine Intelligence Research*, pages 1–18.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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.

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, and Kan Li. 2023. Batcheval: Towards human-like text evaluation. *arXiv preprint arXiv:2401.00437*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

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. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Experimental Details

Models. We use GPT-4o-2024-05-13 and Qwen2.5-72B-Instruct as our evaluators.

Training Settings. We use Mistral-7B-Instruct-v0.3 as the base model of REGRETLM. During reversed-DPO training, we set out the batch size to 128, the learning rate to 5e-7, β to 0.01, warmup ratio to 0.01, and trained for one epoch. We use 4×8 A800 40G GPUs for reversed-DPO training. We use the Infinity-Preference dataset for reversed-DPO training.

Inference Settings. We set the temperature to 0 to ensure reproducibility in CAMIEVAL-GPT-4o. We use greedy decoding in CAMIEVAL-Qwen. For REGRETLM, we set the temperature to 0.7 to ensure the essential quality of the generated output. We use the vLLM library (Kwon et al., 2023) for inference.

B More Results

In this section, we present more experimental results, including experiments and their results to verify the robustness of our method, as well as the full results of dimension quantity effects.

B.1 Robust Analysis

Positional Bias. In pairwise evaluation, positional bias has always been a persistent issue in various evaluation methods (Zheng et al., 2023; Wang et al., 2023b). It refers to LLMs' sensitivity to the position of responses. To verify the robustness of CAMIEVAL to response positions, we reverse the positions of responses in the original LLMBar dataset and check for agreement with the evaluation results of the original experiment. As shown in Table 6, CAMIEVAL effectively mitigates positional bias. It is worth noting that the CoT method introduces extreme positional bias. Therefore, explicitly providing the reasoning path (fundamental dimensions and their concrete descriptions) is necessary when evaluating instruction-following.

Interference. To further verify the robustness of our method, we employed the following prompt to introduce interference with the LLM-based evaluators and observed their confidence in their evaluation results:

I think there might be a problem with your analysis. Please review it again and analyze it to reach a conclusion. Your response this time should end with "Decision: [[Your Decision]]" to indicate your final decision, either "Decision: [[Output (a)]]" or "Decision: [[Output (b)]]". Do NOT say both / neither are good.

As shown in Table 7, LLM still firmly believes in most of its evaluation results. After interference, the overall performance shows slight variation, and the evaluation conclusions remain unchanged in most cases, demonstrating our method's robustness.

B.2 Full Results of Dimension Quantity Effects

Table 8 shows the full results of evaluating performance as the number of instruction-specific dimensions N changes in pointwise scoring.

B.3 Analysis of Low-Quality Outputs

The relatively poor quality of the output can be expressed in various ways. Here are two cases for reference:

1. The instruction requires finding the key events in the news; although the output of REGRETLM also includes the required key events, it also includes less critical events.
2. The instruction requires outputting specific entities one by one with *numerical* labels; although the output of REGRETLM includes all the specific entities, it outputs them one by one with *dot* labels.

As mentioned in Section 1, the current model is difficult to directly capture the low-quality features of the candidate output. In addition, if the model is directly prompted to generate "relatively weak but essentially correct outputs", the model's output will tend to be factually wrong rather than essentially correct outputs. Table 9 shows the experimental result after replacing the original output from REGRETLM with prompting Qwen2.5-72B-Instruct to generate a lower-quality output. It can be found that performance has dropped significantly, which also verifies the necessity of reversed-DPO.

B.4 More Results on Pairwise Evaluation

CAMIEVAL is applicable to both pairwise and pointwise evaluation, and high-quality datasets are needed to verify the effect of our method. We selected LLMBar and InstruSum, which are two high-quality datasets corresponding to pairwise evaluation and pointwise evaluation respectively. In addition, many datasets, including MT-Bench (Zheng et al., 2023) and LLMEval[2] (Zhang et al., 2023)

Methods	Natural	Adversarial					Average
		Neighbor	GPTInst	GPTOut	Manual	Average	
Qwen2.5-72B-Instruct							
Vanilla*	90.0	78.4	**87.0**	**93.6**	89.1	87.0	87.6
CoT*	88.0	79.1	79.3	78.7	78.3	78.9	80.7
Metrics*	95.0	78.4	85.9	89.4	87.0	85.2	87.1
CAMIEval*	**96.0**	**89.6**	83.7	**93.6**	**91.3**	**89.6**	**90.8**

Table 6: Agreement rates with Qwen2.5-72B-Instruct on LLMBar. The highest agreement rate in each column is highlighted in **bold**.

Methods	Natural	Adversarial					Average
		Neighbor	GPTInst	GPTOut	Manual	Average	
Qwen2.5-72B-Instruct							
CAMIEval*	**97.0**	75.4	88.0	74.5	80.4	79.6	83.1
w/ interference	93.0	**76.9**	**90.2**	**78.7**	**84.8**	**82.7**	**84.7**
Consistency	94.0	94.0	97.8	95.7	87.0	93.6	93.7

Table 7: Comparison of CAMIEval with interference on LLMBar. The highest accuracies are in **bold**. The last line indicates the proportion of cases in which the evaluation conclusions remain unchanged after interference.

Methods	InstruSum		
	r	ρ	τ
GPT-4o			
CAMIEval$_{\text{base}}$	0.509	0.481	0.357
CAMIEval$_{N=1}$	0.542	0.524	0.388
CAMIEval$_{N=2}$	0.567	0.550	0.405
CAMIEval$_{N=3}$	**0.572**	**0.557**	**0.413**
CAMIEval$_{N=4}$	0.544	0.524	0.387
CAMIEval$_{N=5}$	0.542	0.527	0.388
Qwen2.5-72B-Instruct			
CAMIEval$_{\text{base}}$	0.484	0.464	0.349
CAMIEval$_{N=1}$	0.499	0.489	**0.369**
CAMIEval$_{N=2}$	0.493	0.467	0.344
CAMIEval$_{N=3}$	0.491	0.478	0.350
CAMIEval$_{N=4}$	0.494	**0.492**	0.363
CAMIEval$_{N=5}$	**0.505**	**0.492**	**0.369**
CAMIEval$_{N=6}$	0.496	0.485	0.356
CAMIEval$_{N=7}$	0.463	0.442	0.328

Table 8: Full results of comparing instruction-specific dimensions in different values of N.

Methods	InstruSum		
	r	ρ	τ
Qwen2.5-72B-Instruct			
$N = 1^*$	0.482	0.466	0.349
$N = 2^*$	0.478	0.465	0.339
$N = 3^*$	0.444	0.428	0.313
$N = 4^*$	0.449	0.423	0.310
$N = 5^*$	0.456	0.437	0.326
CAMIEval *	**0.505**	**0.492**	**0.369**

Table 9: Comparison of CAMIEval with the model directly prompted to generate low-quality output. * indicates the use of a reference. N represents the number of instruction-specific dimensions. The highest correlation in each column is highlighted in **bold**.

datasets, as Zeng et al. (2023) mentioned, do not ensure the objective correctness of the preferences, so the accuracies on them do not reliably reflect the evaluators' capabilities. Considering this factor, we did not present the relevant experimental results in the main text. However, we still provide the performance of baselines and CAMIEval on MT-Bench and LLMEval^2 datasets here. As shown in Table 10, considering all datasets, our method still achieved the best average performance.

B.5 Pairwise Evaluation with Instruction-Specific Dimensions

As we mentioned in Section 5, RegretLM produces relatively low-quality outputs. By comparing and analyzing with the reference output, the evaluation dimensions are complemented, and excellent performance is achieved in pointwise evaluation. For pairwise evaluation, it naturally has one positive example and one negative example. By concretizing the fundamental dimensions through ConcreteCoT, good performance has been achieved. The additional subsequent "instruction-specific dimensions generation" process has lit-

Methods	**LLMBar**	**MTBench**	**LLMEval^2**	Average
Vanilla	67.5	83.0	77.0	75.8
Vanilla*	71.9	83.5	76.5	77.3
CoT	72.7	80.0	76.5	76.4
CoT*	77.2	81.5	75.5	78.1
Metrics	73.8	82.5	77.5	77.9
Metrics*	77.7	**84.5**	76.5	79.6
CAMIEval-Qwen	79.6	82.5	76.5	79.5
CAMIEval-Qwen*	**83.1**	81.5	**79.0**	**81.2**

Table 10: Accuracies with Qwen2.5-72B-Instruct on LLMBar, MT-Bench and LLMEval^2. * indicates the use of a reference. The highest accuracy in each column is highlighted in **bold**.

Methods	**Natural**	**Adversarial**					Average
		Neighbor	GPTInst	GPTOut	Manual	Average	
Qwen2.5-72B-Instruct							
$N = 1^*$	94.0	**79.9**	**90.2**	76.6	**82.6**	**82.3**	**84.7**
$N = 2^*$	96.0	76.9	**90.2**	76.6	**82.6**	81.6	84.5
$N = 3^*$	**97.0**	76.1	88.0	**78.7**	76.1	79.7	83.2
$N = 4^*$	96.0	74.6	88.0	76.6	**82.6**	80.5	83.6
$N = 5^*$	95.0	73.9	85.9	**78.7**	76.1	78.7	81.9
CAMIEval*	**97.0**	75.4	88.0	74.5	80.4	79.6	83.1

Table 11: Accuracies with instruction-specific dimensions on LLMBar. * indicates the use of a reference. N represents the number of instruction-specific dimensions. The highest accuracy in each column is highlighted in **bold**.

tle effect. As shown in Table 11, by adding the instruction-specific dimensions for evaluation, the best performance is only 1.6 points (1.9%) higher than the original result.

B.6 Ablation Experiment on Fundamental Dimensions

Table 12 shows the performance on the LLMBar dataset using only one of the fundamental dimensions. It can be seen from the results that using only one dimension cannot capture the partial order relationship in pairwise evaluation well. In addition, among the three fundamental dimensions, *adherence* is the most important. It is very close to the performance of CAMIEval.

B.7 Results Using Qwen Models of Different Sizes

Qwen2.5 has multiple sizes. We further explore the performance of CAMIEval under different sizes of Qwen2.5. As shown in Table 13 and Table 14, the performance of the 32B model is roughly equivalent to that of the 72B model, while the performance of the 14B model is slightly worse. The performance of the 7B model is so poor that it can hardly complete the evaluation task.

C Data Leakage Analysis

We used the latest benchmarks possible to validate our method (LLMBar released in October 2023 and InstruSum released in November 2023). Due to the unknown specific cutoff dates for the model training data (Qwen2.5-72B-Instruct released in September 2024 with an unknown training data cutoff date, and GPT-4o released in May 2024 with a training data cutoff in October 2023), there is a risk of dataset leakage during the evaluation process. We will employ the latest NLG evaluation benchmarks in future work to validate our methods further.

D Example Prompts

Table 15 - Table 19 show all the prompts used in CAMIEval. We adopt and modify the prompt templates from LLMBar and AlignBench for the final evaluation phase, respectively, for pairwise evaluation (Table 16, 17) and pointwise scoring (Table 19).

Methods	Natural	Adversarial				Average
		Neighbor	GPTInst	GPTOut	Manual	
Qwen2.5-72B-Instruct						
Relevance Only*	95.0	76.9	85.9	70.2	71.7	79.9
Factuality Only*	92.0	76.9	84.8	**80.9**	73.9	81.7
Adherence Only*	94.0	**77.6**	87.0	72.3	**82.6**	82.7
CAMIEval *	**97.0**	75.4	**88.0**	74.5	80.4	**83.1**

Table 12: Comparison of CAMIEval with a single fundamental dimension in pairwise evaluation. * indicates the use of a reference. The highest accuracy in each column is highlighted in **bold**.

Models	Natural	Adversarial				Average
		Neighbor	GPTInst	GPTOut	Manual	
Qwen2.5-7B-Instruct	85.0	69.4	75.0	57.4	67.4	70.8
Qwen2.5-7B-Instruct*	84.0	68.7	77.2	61.7	60.9	70.5
Qwen2.5-14B-Instruct	88.0	68.7	79.3	**76.6**	73.9	77.3
Qwen2.5-14B-Instruct*	90.0	70.9	79.3	72.3	78.3	78.2
Qwen2.5-32B-Instruct	93.0	76.9	87.0	74.5	**80.4**	82.4
Qwen2.5-32B-Instruct*	94.0	**81.3**	**93.5**	70.2	78.3	**83.5**
Qwen2.5-72B-Instruct	94.0	76.1	81.5	72.3	73.9	79.6
Qwen2.5-72B-Instruct*	**97.0**	75.4	88.0	74.5	**80.4**	83.1

Table 13: Accuracies with different sizes of Qwen2.5 on LLMBar. * indicates the use of a reference. The highest accuracy in each column is highlighted in **bold**.

Models	**InstruSum**		
	r	ρ	τ
Qwen2.5-7B-Instruct*	0.360	0.361	0.261
Qwen2.5-14B-Instruct*	0.503	0.473	0.346
Qwen2.5-32B-Instruct*	**0.515**	0.491	0.364
Qwen2.5-72B-Instruct*	0.505	**0.492**	**0.369**

Table 14: Pearson (r), Spearman (ρ), and Kendall-Tau (τ) correlations with summary-level human judgments for different sizes of Qwen2.5 on InstruSum. * indicates the use of a reference. The highest correlation in each column is highlighted in **bold**.

E Example Results

Table 20 - Table 22 show some of the intermediate processes and results obtained with CAMIEval. Table 23 - Table 25 illustrate, respectively, how the dimensions of *relevance*, *factuality* and *adherence* manifest in the given instructions. It can be observed that previous methods fail to capture the corresponding issues in the candidate outputs; however, these issues reveal that the candidate outputs fail to follow the instructions.

You are a helpful assistant in evaluating the quality of the outputs for a given instruction. I will provide you with 3 basic questions. Your goal is to concretize the following 3 questions according to the given instruction, to obtain new questions closely related to the instruction, so as to better evaluate the quality of the output corresponding to the instruction. The 3 concrete questions must be distinguished from each other and focus on different aspects. Each question must be answerable with single word "Yes" or "No".

# Question 1
Does the output content directly relate to the question of the instruction?
> Hint: You need to first analyze what this instruction is specifically focusing on, and then concretize this question.
> Example:
> [Instruction] Compare the similarities and differences between NLP and CV.
> [Concrete Question] Does the output focus on NLP and CV, rather than other fields like Speech?

# Question 2
Does the output of the instruction not contain factual errors?
> Hint: You need to first consider what kind of answer is factually correct, then concretize this question.
> Example:
> [Instruction] Calculate the average of the series 1, 2, 3, 4, 5.
> [Concrete Question] Does the output provide the correct average (3) of the specific sequence of numbers provided (1, 2, 3, 4, 5)?

# Question 3
Does the output of the instruction strictly follow the specific requirements of the instruction?
> Hint: Specific requirements include: format requirements (such as specific code formats (like Python), data formats (like JSON), etc.), structural requirements (such as answering in bullet points, expressing in single sentence instead of detailed explanation, etc.), quantity requirements (such as providing an example (meaning single example), providing 5 examples (meaning exactly 5 examples)), content requirements (must include or exclude specific information or keywords, must use English, etc.). If the instruction do not mention these, you can ask whether the output is fluent.
> Example:
> [Instruction] Provide a suggestion for my academic planning.
> [Concrete Question] Does the output provide exactly one suggestion for academic planning?

You should respond in the following format:
# Question 1
[Thinking]
<your thoughts on what answers are directly related to the question>
[Concrete Question]
<concrete question that can be answerable with single word "Yes" or "No">

# Question 2
[Thinking]
<your thoughts on what answers are in fact correct>
[Concrete Question]
<concrete question that can be answerable with single word "Yes" or "No">

# Question 3
[Thinking]
<your thoughts on the specific requirements of the instruction>
[Concrete Question]
<concrete question that can be answerable with single word "Yes" or "No">

[Instruction Start]
{Instruction}
[Instruction End]

Table 15: The prompt template for ConcreteCoT.

You are a helpful assistant in evaluating the quality of the outputs for a given instruction.
Your goal is to select the Output (a) or Output (b) that is better for the given instruction.
The two outputs are generated by two different AI chatbots respectively.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the instruction asks for, as such outputs do NOT precisely execute the instruction.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Instruction:
{Instruction}

# Output (a):
{Output (a)}

# Output (b):
{Output (b)}

# Questions about Outputs:
Here are three questions about the outputs, which are presented from most important to least important. You can do the evaluation based on thinking about all the questions.
1. **{Question 1}**
2. **{Question 2}**
3. **{Question 3}**

# A reference output generated by a strong AI assistant:
{Reference Output}

# Analysis and Decision
Answer and analyze the three questions mentioned above for each of the two outputs. Finally, decide which output is better. Your response should end with "Decision: [[Your Decision]]" to indicate your final decision, either "Decision: [[Output (a)]]" or "Decision: [[Output (b)]]". Do NOT say both / neither are good.

Table 16: The prompt template of CAMIEVAL for pairwise evaluation (w/o tie, w/ reference).

You are a helpful assistant in evaluating the quality of the outputs for a given instruction.
Your goal is to select the Output (a) or Output (b) that is better for the given instruction.
The two outputs are generated by two different AI chatbots respectively.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the instruction asks for, as such outputs do NOT precisely execute the instruction.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Instruction:
{Instruction}

# Output (a):
{Output (a)}

# Output (b):
{Output (b)}

# Questions about Outputs:
Here are three questions about the outputs, which are presented from most important to least important. You can do the evaluation based on thinking about all the questions.
1. **{Question 1}**
2. **{Question 2}**
3. **{Question 3}**

# A reference output generated by a strong AI assistant:
{Reference Output}

# Analysis and Decision
Answer and analyze the three questions mentioned above for each of the two outputs.
Finally, decide which output is better. Your response should end with "[[Decision]]" to indicate your decision, including "[[Output (a)]]", "[[Output (b)]]", "[[Both]]" or "[[Neither]]".

Table 17: The prompt template of CAMIEVAL for pairwise evaluation (w/ tie, w/ reference).

You are a helpful assistant in evaluating the quality of the outputs for a given instruction. I will provide you with an instruction, a reference output corresponding to the instruction, and an output generated by a weak AI model. Additionally, I will provide you with three basic dimensions for evaluating output quality. Your goal is to compare the reference output with the output generated by the weak AI model, analyze where the weak AI model's output falls short compared to the reference output, and supplement with **{n_dimensions}** new dimensions for evaluating output quality. This will fill in the aspects not considered by the original basic dimensions, allowing for a more comprehensive evaluation of output quality.

Specifically, you need to follow these rules:
(1) In the comparative analysis between the reference output and the weak AI model output, after carefully reading both the reference output and the weak AI model's output, you need to identify exactly where the weak AI model's output falls short compared to the reference output.
(2) Based on your comparative analysis, consider what aspects the original three basic evaluation dimensions have not taken into account, and generate **{n_dimensions}** new evaluation dimensions accordingly. The newly generated evaluation dimensions must be distinct from the original three basic evaluation dimensions and should be formatted in the same way as the original three basic evaluation dimensions.
(3) You need to follow the following output format:
[Comparison and Analysis]
<your comparison and analysis of reference output and weak AI model output>

[Thinking of the New Dimensions]
<your thoughts on what other aspects of the original 3 basic evaluation dimensions have not been taken into account>

[New Evaluation Dimensions]
Dimension 1: <the name of dimension 1>
Definition: <the definition of dimension 1>
...
Dimension **{n_dimensions}**: <the name of **{n_dimensions}**>
Definition: <the definition of dimension **{n_dimensions}**>

Here is the Instruction:
{Instruction}

Here is the reference output:
{Reference Output}

Here is the output generated by a weak AI model:
{Weak Output}

Here is the basic dimensions:
{Basic Criteria}

Table 18: The prompt template for generating instruction-specific dimensions.

You are a helpful assistant in evaluating the quality of text. Please evaluate the quality of an AI assistant's response to user query as an impartial judge. You need to evaluate the response on the following dimensions:

{Criteria}

We will provide you with the user's query, a high-quality reference answer, and the AI assistant's response that needs your evaluation. When you commence your evaluation, you should follow the following process:

1. Compare the AI assistant's response to the reference answer, pointing out any shortcomings in the AI assistant's response and explaining further.
2. Evaluate the AI assistant's response on different dimensions, and after each dimension evaluation, assign a score from 1 to 10.
3. Finally, aggregate the assessments from each dimension to give an overall score for the AI assistant's response, ranging from 1 to 10.
4. Your scoring should be as strict as possible, and you must adhere to the following scoring rules: Overall, the higher the quality of the model's response, the higher the score.

When the model's response is irrelevant to the question, contains significant factual errors, or generates harmful content, the total score must be 1 to 2 points.

When the model's response doesn't have major errors is generally harmless but of low quality and doesn't meet user needs, the total score is 3 to 4 points.

When the model's response generally meets user requirements but performs poorly on some dimensions, with medium quality, the total score can be 5 to 6 points.

When the model's response quality is close to the reference answer in all dimensions and performs well, the total score is 7 to 8 points.

Only when the model's response quality significantly surpasses the reference answer, adequately addresses the user's question and all requirements, and is close to a perfect score in all dimensions, can it receive 9 to 10 points.

As an example, a reference answer can receive a score of 8.

Please remember to provide evaluations and explanations before your scoring. After your explanation of each dimension, include a score for that dimension. Finally, provide the final score for the given response (a number between 0 and 10, which can include decimal points) and put it at the end of your output, strictly following the format "[[score]]", for example: "Final Score: [[6.0]]".

# User's Query:

{Instruction}

# Reference Answer:

{Refenence Answer}

# Assistant's Response:

{Output}

Table 19: The prompt template of CAMIEVAL for pointwise scoring.

Instruction:
Write a script for a 10-second conversation between two characters.
Output (a):
Person A: Hi, I'm John. It's nice to meet you. Person B: Hi John, I'm Jane. It's nice to meet you too. Person A: So what brings you here? Are you here for an event? Person B: Actually I just came to explore the city, I'm a bit of a wanderer. Person A: That's cool, this is a great city to explore. Person B: What about you, why did you come here? Person A: Well I'm actually here for a work conference. Person B: Ah, I see. What kind of work do you do? Person A: I'm a digital marketer. Person B: That sounds interesting, what kind of projects have you worked on? Person A: Oh, I've done a lot of different things. Person B: That sounds really cool. Person A: Well, if you're looking for something to do while you're here, why not check out this museum? Person B: Sounds like a plan, I'm always up for discovering new places. Person A: Me too! Well, it was nice meeting you. Person B: Same here, John.
Output (b):
Person A: Hi, how are you? Person B: I'm doing well, how about you? Person A: I'm doing alright, just dealing with a lot of changes. Person B: I understand. Changes can usually be hard.
Gold Preference:
Output (b) is better.

Evaluation Result of Vanilla:
Output (a) is better.

Evaluation Result of CoT:
Output (a) provides a more detailed and engaging conversation that fits the context of a 10-second interaction between two characters. It includes greetings, introductions, and a brief exchange about their reasons for being in the same place, which aligns well with the instruction to write a script for a 10-second conversation. Output (b), while a valid conversation, is shorter and less detailed, not fully utilizing the 10-second duration and lacking the depth and context provided in Output (a). Therefore, Output (a) is better.

Evaluation Result of Metrics:
Metrics:
Does the script contain a conversation that is exactly 10 seconds long?
Does the script include only two characters speaking?
Does the script maintain a natural and coherent flow of dialogue?
Conclusion:
Output (a) is better.

Evaluation Result of CAMIEval-Qwen:
Fundamental Dimensions & Descriptions:
Does the output contain a conversation that could be spoken within 10 seconds?
Is the dialogue in the output grammatically correct and logically coherent?
Is the output presented in a script format with dialogue for two characters?
Analysis & Conclusion:
Analysis of Output (a)
1. Does the output contain a conversation that could be spoken within 10 seconds?
- No. The conversation in Output (a) is quite lengthy and would take more than 10 seconds to read aloud. It includes multiple exchanges and detailed back-and-forth dialogue.
2. Is the dialogue in the output grammatically correct and logically coherent?
- Yes. The dialogue is grammatically correct and logically coherent. The conversation flows naturally and the characters' responses are appropriate and contextually relevant.
3. Is the output presented in a script format with dialogue for two characters?
- Yes. The output is presented in a script format with clear dialogue for two characters, Person A and Person B.
Analysis of Output (b)
1. Does the output contain a conversation that could be spoken within 10 seconds?
- Yes. The conversation in Output (b) is brief and could be spoken within 10 seconds. It consists of a few short exchanges.
2. Is the dialogue in the output grammatically correct and logically coherent?
- Yes. The dialogue is grammatically correct and logically coherent. The conversation is natural and the responses are appropriate.
3. Is the output presented in a script format with dialogue for two characters?
- Yes. The output is presented in a script format with clear dialogue for two characters, Person A and Person B.
Final Decision
- Output (a) does not meet the primary requirement of being a 10-second conversation, as it is too long. - Output (b) meets all the criteria: it is a 10-second conversation, the dialogue is grammatically correct and logically coherent, and it is presented in a script format with dialogue for two characters.
Decision: Therefore, Output (b) is better.

Table 20: Evaluation results of a sample from LLMBar-Natural. The table shows that various previous methods were affected by the output length, favoring length over quality.

Instruction:
Tell me three ways we can reduce air pollution.
Output (a):
Three ways to reduce air pollution include reducing the use of vehicles, reducing the use of fossil fuels, and increasing the efficiency of existing buildings.
Output (b):
Increase usage of public transportation and reduce private vehicle emissions. Ban or enforce stricter emissions standards for factories and construction sites. Create incentives for citizens and businesses to adopt green energy or environmentally friendly practices. Build green spaces such as parks and gardens to absorb carbon dioxide from the atmosphere.
Gold Preference:
Output (a) is better.

Evaluation Result of Vanilla:
Output (b) is better.

Evaluation Result of CoT:
Output (b) provides more detailed and specific methods for reducing air pollution, such as increasing public transportation, enforcing stricter emissions standards, and creating green spaces. It also aligns closely with the reference output in terms of the level of detail and the practicality of the suggestions. Output (a), while accurate, is less detailed and does not provide as clear or actionable steps. Therefore, Output (b) is better.

Evaluation Result of Metrics:
Metrics:
Does the output provide exactly three methods to reduce air pollution?
Does each method mentioned in the output clearly and specifically address how it can reduce air pollution?
Does the output avoid including unnecessary information or omitting required details about the methods?
Conclusion:
Output (b) is better.

Evaluation Result of CAMIEVAL-Qwen:
Fundamental Dimensions & Descriptions:
Does the output list methods that are directly related to reducing air pollution?
Does the output contain only scientifically accurate methods for reducing air pollution?
Does the output provide exactly three methods to reduce air pollution?
Analysis & Conclusion:
Analysis of Output (a):
1. Does the output list methods that are directly related to reducing air pollution?
- Yes, the methods listed (reducing vehicle use, reducing fossil fuel use, and increasing building efficiency) are directly related to reducing air pollution.
2. Does the output contain only scientifically accurate methods for reducing air pollution?
- Yes, all the methods mentioned are scientifically accurate and effective in reducing air pollution.
3. Does the output provide exactly three methods to reduce air pollution?
- Yes, the output provides exactly three methods.
Analysis of Output (b):
1. Does the output list methods that are directly related to reducing air pollution?
- Yes, the methods listed (increasing public transportation, enforcing stricter emissions standards, creating incentives for green energy, and building green spaces) are directly related to reducing air pollution.
2. Does the output contain only scientifically accurate methods for reducing air pollution?
- Yes, all the methods mentioned are scientifically accurate and effective in reducing air pollution.
3. Does the output provide exactly three methods to reduce air pollution?
- No, the output provides more than three methods. It lists four distinct methods: increasing public transportation, enforcing stricter emissions standards, creating incentives for green energy, and building green spaces.
Final Decision:
Output (a) is better because it adheres more closely to the instruction by providing exactly three methods to reduce air pollution, while Output (b) provides more than three methods, which does not precisely execute the instruction.
Decision: Therefore, Output (a) is better.

Table 21: Evaluation results of a sample from LLMBar-Adversarial. The table shows that various previous methods cannot accurately understand the meaning of the given instruction, and prefer to the longer output.

Instruction:
Generate a summary of the given article based on specific summary requirement.
Article:
A lifeboat enthusiast has set about recording the country's RNLI volunteers using Victorian photographic methods. By Francesca WilliamsBBC News Quietly spoken, Jack Lowe is a man on the verge of his dreams. A life-long love of photography and lifeboats has finally put him on the road around Britain's coast. Having given up the Newcastle printmaking business he ran for 15 years, he plans to record all 237 of the Royal National Lifeboat Institution's stations on glass plates with a 110-year-old camera.
TL;DR
Summary Requirement:
Summarize Jack Lowe's background and his passion for photography and lifeboats.
Candidate Output:
Jack Lowe is a lifeboat enthusiast and photographer who has embarked on a project to record all of the Royal National Lifeboat Institution's (RNLI) stations in Britain using Victorian-era photographic methods. Lowe, who gave up his printmaking business in Newcastle 15 years ago, plans to use a 110-year-old camera to take glass plate ambrotypes of each of the RNLI's 237 stations. He is 39 years old.
Lowe says that he was motivated to start the project by his desire to create physical, tangible objects that would last for centuries, and by his desire to engage and collaborate with the people he photographs. He is excited by the reaction his pictures elicit from their subjects, many of whom are moved to tears by the experience.
The RNLI is supportive of Lowe's project and believes that the resulting pictures will be effective in conveying the timelessness of the organisation and its crews.
Lowe plans to donate the finished collection of ambrotypes to the RNLI, and sell prints to anyone who wants them, in order to fund the project and pay his living expenses. He expects the project to take four or five years to complete, and says that he plans to become a lifeboatman himself one day.
Reference:
Jack Lowe is a lifeboat enthusiast and photographer who has embarked on a journey to document all 237 Royal National Lifeboat Institution's stations using a 110-year-old camera and Victorian photographic methods. With a lifelong love for photography and lifeboats, Lowe gave up his Newcastle printmaking business to pursue this passion project. He aims to create timeless images that capture the essence of lifeboat crews and their dedication to saving lives at sea.
Gold Score: 3.5/5.0

Evaluation Result of CAMIEval-Qwen:
Fundamental Dimensions & Descriptions:
Relevance to Instruction: Does the output focus on Jack Lowe's background and his passion for photography and lifeboats, rather than other aspects of the article?
Factuality: Does the output provide accurate details about Jack Lowe's life, his project, and the techniques he uses?
Adherence to Specific Requirements: Does the output strictly adhere to the requirement to summarize Jack Lowe's background and his passion for photography and lifeboats, without including irrelevant details?
Instruction-Specific Dimensions & Descriptions:
Conciseness: Does the output provide the necessary information in a concise and to-the-point manner, avoiding unnecessary details and verbosity?
Clarity and Structure: Is the output well-organized and easy to follow, with a clear structure that enhances readability and understanding?
Engagement: Does the output effectively engage the reader by using vivid language, relevant examples, and a compelling narrative?
Essence of the Subject: Does the output capture the essence of the subject, conveying their personality, motivations, and the significance of their actions or projects?
Balance of Information: Does the output maintain a balance between providing essential details and avoiding overloading the reader with too much information, ensuring that the core message is not lost?
Analysis & Scoring:
Evaluation of the AI Assistant's Response
Dimension 1: Relevance to Instruction
The AI assistant's response focuses on Jack Lowe's background and his passion for photography and lifeboats, as required. However, it includes some additional details that are not strictly necessary for the summary, such as his age and the specific number of stations. The reference answer is more concise and directly relevant to the instruction.
Score: 8
Dimension 2: Factuality
The AI assistant's response is factually accurate. It correctly states that Jack Lowe gave up his printmaking business, uses a 110-year-old camera, and plans to document all 237 RNLI stations. The response also accurately captures his motivations and the reactions of the subjects.
Score: 9
Dimension 3: Adherence to Specific Requirements
The AI assistant's response adheres to the requirement to summarize Jack Lowe's background and his passion for photography and lifeboats. However, it includes some extraneous details, such as his age and the specific number of stations, which are not essential to the summary. The reference answer is more focused and adheres more strictly to the requirements.
Score: 7
Dimension 4: Conciseness
The AI assistant's response is generally concise but includes some unnecessary details, such as his age and the specific number of stations. The reference answer is more concise and to the point.
Score: 7
Dimension 5: Clarity and Structure
The AI assistant's response is well-organized and easy to follow. It has a clear structure that enhances readability and understanding. However, it could be slightly more streamlined to avoid redundancy.
Score: 8
Dimension 6: Engagement
The AI assistant's response effectively engages the reader by using vivid language and relevant examples, such as the reactions of the subjects and the RNLI's support. However, it could be more compelling in its narrative flow.
Score: 8
Dimension 7: Essence of the Subject
The AI assistant's response captures the essence of Jack Lowe, conveying his personality, motivations, and the significance of his project. It effectively communicates his passion for photography and lifeboats and the impact of his work.
Score: 8
Dimension 8: Balance of Information
The AI assistant's response maintains a good balance between providing essential details and avoiding overloading the reader with too much information. However, it could be more concise by omitting some less relevant details.
Score: 7
Final Score
Aggregating the scores from each dimension, the overall score for the AI assistant's response is: Final Score: [[7.5]]
Conclusion:
Final Score: 7.5/10.0

Table 22: Evaluation results of a sample from InstruSum. The table shows that our method conducted a meticulous analysis of the candidate output based on three fundamental dimensions and five instruction-specific dimensions, and finally obtained a score (7.5/10.0) close to the gold score (3.5/5.0).

Instruction:
Possible way to solve the collatz conjecture.
Output (a):
The Collatz Conjecture is an unsolved problem in mathematics that asks about the behavior of a particular sequence defined for positive integers. One possible solution to the Collatz Conjecture could be to use a computer program to perform the required calculations. However, there could be other solutions as well, and the exact solution would depend on the details of the sequence and the problem.
Output (b):
The Collatz Conjecture is a famous unsolved problem in mathematics, which is also known as the 3n + 1 conjecture or the Ulam conjecture. It states that for any positive integer n, if n is even, then divide it by 2, otherwise, if n is odd, then multiply it by 3 and add 1. Then take the result and repeat this process with the new value obtained, and so on. The conjecture claims that no matter what positive integer you start with, this process will always eventually reach the number 1.
For example, if we start with the number 6, which is even, we divide it by 2 to get 3. Since 3 is odd, we multiply it by 3 and add 1 to get 10. Continuing in this way, we get the sequence 6, 3, 10, 5, 16, 8, 4, 2, 1. The conjecture asserts that this sequence always terminates at 1, no matter what number you start with.
Despite the simplicity of the Collatz Conjecture, it remains unsolved, and no one has been able to prove or disprove it for all positive integers. While it has been verified by computer for an incredibly large number of cases, it remains one of the most intriguing open problems in mathematics.
Gold Preference:
Output (a) is better.

Evaluation Result of Vanilla:
Output (b) is better.

Evaluation Result of CAMIEVAL-Qwen:
Fundamental Dimensions & Descriptions:
Does the output **focus on methods or approaches to solve the Collatz Conjecture**?
Does the output contain any factual errors or incorrect mathematical statements?
Is the output clear and focused on providing meaningful suggestions or methods to solve the Collatz Conjecture?
Analysis & Conclusion:
Analysis of Output (a)
1. Does the output focus on methods or approaches to solve the Collatz Conjecture?
- Output (a) mentions the use of a computer program to perform calculations, which is a valid approach. However, it does not provide any specific methods or detailed steps. It also suggests that there could be other solutions but does not elaborate on them.
2. Does the output contain any factual errors or incorrect mathematical statements?
- Output (a) does not contain any factual errors or incorrect mathematical statements.
3. Is the output clear and focused on providing meaningful suggestions or methods to solve the Collatz Conjecture?
- Output (a) is somewhat clear but lacks detail and specificity. It does not provide meaningful suggestions or methods beyond the general idea of using a computer program.
Analysis of Output (b)
1. Does the output focus on methods or approaches to solve the Collatz Conjecture?
- Output (b) provides a detailed explanation of the Collatz Conjecture but **does not focus on methods or approaches to solve it**. It describes the sequence and its behavior but does not offer any specific strategies or techniques for solving the conjecture.
2. Does the output contain any factual errors or incorrect mathematical statements?
- Output (b) does not contain any factual errors or incorrect mathematical statements. It accurately describes the Collatz Conjecture and provides a clear example.
3. Is the output clear and focused on providing meaningful suggestions or methods to solve the Collatz Conjecture?
- Output (b) is clear and provides a detailed explanation of the conjecture, but it does not offer any meaningful suggestions or methods to solve it. It is more focused on explaining the problem itself rather than providing solutions.
Decision
Output (a) is better because it at least mentions a potential method (using a computer program) to solve the Collatz Conjecture, even though it lacks detail. Output (b) provides a detailed explanation of the conjecture but does not offer any methods or approaches to solve it, which is the primary focus of the instruction.
Decision: Therefore, Output (a) is better.

Table 23: An example of a case that fails in terms of *relevance* from LLMBar. The table shows that Output (b) does not provide any "possible way", which indicates its lack of relevance to the instruction.

Instruction:
Identify the cause of this issue: The computer is not working.
Output (a):
First, check that the internet cable is properly connected to the computer. Check the Wi-Fi settings of the device and make sure the correct network is selected. If the problem still exists, restart the computer and router to check if the issue is resolved. If the issue persists, contact your internet service provider to check if the issue is caused by them.
Output (b):
The cause of this issue is unknown without more information.
Gold Preference:
Output (b) is better.

Evaluation Result of Vanilla:
Output (a) is better.

Evaluation Result of CAMIEVAL-Qwen:
Fundamental Dimensions & Descriptions:
Does the output provide specific reasons why the computer might not be working?
Does the output **contain only technically accurate and plausible reasons for why the computer is not working**?
Does the output clearly and directly identify the cause of the computer not working?
Analysis & Conclusion:
Analysis of Output (a):
1. Does the output provide specific reasons why the computer might not be working?
- Output (a) provides specific steps to troubleshoot the issue, focusing on internet connectivity. However, it assumes the issue is related to the internet, which may not be the case. The output does not provide a comprehensive list of potential causes for the computer not working.
2. Does the output contain only technically accurate and plausible reasons for why the computer is not working?
- The steps provided are technically accurate and plausible for troubleshooting internet connectivity issues. **However, they are not relevant to all possible causes of a computer not working.**
3. Does the output clearly and directly identify the cause of the computer not working?
- No, the output does not clearly and directly identify the cause. It provides steps to troubleshoot a specific issue (internet connectivity) but does not address the broader problem of the computer not working.
Analysis of Output (b):
1. Does the output provide specific reasons why the computer might not be working?
- Output (b) does not provide specific reasons. It acknowledges that the cause is unknown without more information, which is a more honest and accurate response given the limited information provided in the instruction.
2. Does the output contain only technically accurate and plausible reasons for why the computer is not working?
- **The output is technically accurate in stating that the cause is unknown without more information. It does not provide any incorrect or implausible reasons.**
3. Does the output clearly and directly identify the cause of the computer not working?
- No, the output does not identify the cause. However, it is clear and direct in stating that more information is needed to determine the cause.
Final Decision:
Output (b) is more honest and accurate in acknowledging that the cause is unknown without more information. While Output (a) provides specific steps, they are too narrow and assume the issue is related to internet connectivity, which may not be the case.
Decision: Therefore, Output (b) is better.

Table 24: An example of a case that fails in terms of *factuality* from LLMBar. The table shows that Output (a) exhibit a hallucination, assuming that the computer's malfunction is caused by network issues.

Instruction:
Format the following passage into a bulleted list. Carbon dioxide emissions are largely caused by burning fossil fuels, such as coal and natural gas, to generate electricity and power our daily activities.
Output (a):
- Carbon dioxide emissions are largely caused by burning fossil fuels
- Examples of fossil fuels include coal and natural gas
- Burning fossil fuels are used to generate electricity and power our daily activities.

Output (b):
1. Carbon dioxide emissions are largely caused by burning fossil fuels.
2. Fossil fuels include coal and natural gas.
3. Burning fossil fuels is done to generate electricity.
4. Burning fossil fuels also powers our daily activities.

Gold Preference:
Output (a) is better.

Evaluation Result of Vanilla:
Output (b) is better.

Evaluation Result of CAMIEVAL-Qwen:
Fundamental Dimensions & Descriptions:
Does the output present the information from the passage in a bulleted list format?
Does the output accurately state that carbon dioxide emissions are largely caused by burning fossil fuels like coal and natural gas?
Does the output **strictly follow the instruction by only presenting the information in a bulleted list** without additional text or explanations?
Analysis & Conclusion:
Analysis of Output (a):
1. Does the output present the information from the passage in a bulleted list format?
- Yes, Output (a) presents the information in a bulleted list format.

2. Does the output accurately state that carbon dioxide emissions are largely caused by burning fossil fuels like coal and natural gas?
- Yes, Output (a) accurately states that carbon dioxide emissions are largely caused by burning fossil fuels, including coal and natural gas.

3. Does the output strictly follow the instruction by only presenting the information in a bulleted list without additional text or explanations?
- Yes, Output (a) strictly follows the instruction by only presenting the information in a bulleted list without additional text or explanations.

Analysis of Output (b):
1. Does the output present the information from the passage in a bulleted list format?
- Yes, Output (b) presents the information in a bulleted list format, although it uses numbered points instead of bullets.

2. Does the output accurately state that carbon dioxide emissions are largely caused by burning fossil fuels like coal and natural gas?
- Yes, Output (b) accurately states that carbon dioxide emissions are largely caused by burning fossil fuels, including coal and natural gas.

3. Does the output strictly follow the instruction by only presenting the information in a bulleted list without additional text or explanations?
- Yes, Output (b) strictly follows the instruction by only presenting the information in a bulleted list without additional text or explanations. **However, it uses numbered points, which is a minor deviation from the typical bulleted list format.**

Final Decision:
Both outputs accurately present the information and follow the instruction closely. **However, Output (b) uses numbered points instead of bullets, which is a minor deviation from the typical bulleted list format. Output (a) uses standard bullets, which is more in line with the instruction to format the passage into a bulleted list.**
Decision: Therefore, Output (a) is better.

Table 25: An example of a case that fails in terms of *adherence* from LLMBar. The table shows that Output (b) does not follow the specific requirement of the instruction, which states that the output should be in the form of a bulleted list.