Large Language Models Are Active Critics in NLG Evaluation

Anonymous ACL submission

Abstract

The conventional paradigm of using large language models (LLMs) for natural language generation (NLG) evaluation relies on pre-defined task definitions and evaluation criteria, positioning LLMs as "passive critics" that strictly follow developer-provided guidelines. However, human evaluators often apply implicit criteria, and their expectations in practice can vary widely based on specific end-user needs. Consequently, these rigid evaluation methods struggle to adapt to diverse scenarios without extensive prompt customization. To address this, we introduce **ACTIVE-CRITIC**, a novel LLM-based evaluator that transforms LLMs into "active critics" capable of adapting to diverse NLG tasks using limited example data. ACTIVE-CRITIC consists of two stages: (1) self-inferring the target NLG task and relevant evaluation criteria, and (2) dynamically optimizing prompts to produce human-aligned scores along with detailed justifications. Our experiments show that ACTIVE-CRITIC can generate nuanced, context-aware evaluation criteria, enabling it to achieve superior alignment with human judgments across multiple tasks.

1 Introduction

Recent advances in language technologies have accelerated the development of natural language generation (NLG) systems, benefiting a variety of downstream applications such as text summarization (Fabbri et al., 2021), dialogue generation (Mehri and Eskenazi, 2020), and storytelling (Guan et al., 2021). However, despite the rapid progress in NLG systems, reliable techniques for automatic evaluation of NLG systems still lay far behind, primarily due to the inherent challenges posed by the open-ended nature of NLG and the diverse demands of different stakeholders. This gap, in return, undermines the reliability of machine-generated content in real-world applications.

Traditional NLG evaluation methods typically focus on a specific criterion and require human-written references for comparison (Li et al., 2024). Commonly considered criteria include reference similarity (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019; Yuan et al., 2021), text fluency (Kann et al., 2018; Mutton et al., 2007), human likeness (Song et al., 2025; Jiang et al., 2019), and information adequacy (Adlakha et al., 2024). Moving beyond single-aspect metrics, recent studies propose to use a universal large language model (LLM) as a judge to score machine-generated texts across multiple criteria in diverse NLG tasks, either by fine-tuning (Zhong et al., 2022; Jiang et al., 2023; Xu et al., 2023; Ke et al., 2023) or by prompting an LLM for assessment (Chiang and Lee, 2023a; Gong and Mao, 2023; Lin et al., 2023). To address the high cost of human annotation and potential biases introduced by limited references, researchers have further developed reference-free LLM-based evaluations (Fu et al., 2024; Liu et al., 2023a; Li et al., 2023; Jia et al., 2023).

Despite recent progress, two major concerns remain: the reliance on *pre-defined evaluation criteria* and *fixed prompt* crafted by developers, forcing LLM evaluators to adhere strictly to developers' expectations rather than real users. While recent studies (Liu et al., 2024b; Li et al., 2025; Liu et al., 2024a) have explored prompting LLMs to generate evaluation criteria automatically, these methods still rely on pre-defined task descriptions, requiring substantial manual effort to tailor prompts for each NLG task. In contrast, real-world evaluation is shaped by more nuanced and implicit preferences, varying significantly across different stakeholders. Human evaluators naturally adopt diverse evaluation criteria according to their unique perspectives or roles (Liu et al., 2024b; Clark et al., 2021; Celikyilmaz et al., 2020). Moreover, even under the same criterion, different stakeholders may map response quality to scores differently, leading to varied judgments for the same response. Consequently, there are potential risks in constructing criteria and related prompts beforehand without first accounting for human evaluation preferences.

To overcome the above limitations, we propose a novel evaluation approach, i.e., **ACTIVE-CRITIC**, that instructs an LLM to actively derive an evaluation protocol purely from human-scored data examples. Our approach includes two stages: (1) adaptively inferring the target NLG task and identifying its underlying evaluation criteria that matter most to end users, and (2) dynamically optimizing prompts to produce human-aligned judgments across diverse NLG scenarios. To enhance trustworthiness, ACTIVE-CRITIC also generates detailed text justifications alongside its scoring. The comparison between related methods and ACTIVE-CRITIC is shown in Table 1.

We have conducted experiments across diverse NLG tasks using various base LLMs. The results show that the ACTIVE-CRITIC consistently achieves a noticeably higher correlation with human judgments, indicating its ability to adapt effectively to different NLG evaluation tasks according to different evaluation criteria. Our approach requires as few as 5 human-scored data to obtain a strong correlation with humans, with performance steadily improving as the dataset grows. Further analysis highlights that the task inference stage contributes more to ACTIVE-CRITIC's performance than the scoring stage, and ACTIVE-CRITIC can effectively identify nuanced, context-aware criteria beyond pre-defined ones. In summary, our method offers three key benefits:

- **Self-adaptive evaluation.** ACTIVE-CRITIC can infer any NLG evaluation task, recover human judgment criteria, and make justified assessments directly from data, eliminating the need for pre-defined task descriptions, fixed evaluation criteria, or manual prompt engineering.
- **Accurate judgment alignment.** Our two-stage design guides LLMs to mimic human judgment step by step, yielding interpretable justifications while achieving state-of-the-art alignment with human assessments against strong baselines.
- **Generic for diverse LLMs and NLG tasks.** Our method operates independently of specific LLMs and evaluation tasks. Our results on four LLM backbones across four NLG tasks showcase its broad applicability.

Method	Criteria Align.	Scoring Align.	Expl.
InstructScore	pre-defined	Instruction-tuned	Yes
Auto-J	pre-defined	Instruction-tuned	Yes
TIGERScore	pre-defined	Instruction-tuned	Yes
UniEval	pre-defined	Instruction-tuned	No
GPTScore	pre-defined	fixed prompt	No
G-Eval	pre-defined	fixed prompt	No
METAMETRICS	pre-defined	fixed prompt	No
HD-Eval	task-specific	fixed prompt	Yes
DnA-Eval	task-specific	fixed prompt	Yes
AutoCalibrate	task-specific	fixed prompt	No
AC (Ours)	**user-centered**	**prompt optimization**	**Yes**

Table 1: Comparison of our ACTIVE-CRITIC (AC) with common evaluation methods, including InstructScore (Xu et al., 2023), Auto-J (Li et al., 2023), TIGERScore (Jiang et al., 2023), UniEval (Zhong et al., 2022), GPTScore (Fu et al., 2024), G-Eval (Liu et al., 2023a), METAMETRICS (Winata et al., 2024), HD-Eval (Liu et al., 2024b), DnA-Eval (Li et al., 2025), AutoCalibrate (Liu et al., 2024a). Align. → Alignment, Expl. → Explainability.

2 Related Work

NLG Evaluation Overview. Existing NLG evaluation methods include early human-centric approaches (Mellish and Dale, 1998), untrained metrics (Papineni et al., 2002; Lin, 2004; Lavie and Denkowski, 2009), and more recent machine-learned metrics (Sennrich et al., 2015; Zhang et al., 2019; Yuan et al., 2021; Kim et al., 2023), often focusing on single-criterion design (Liu et al., 2023b; Wang et al., 2023b). Recent efforts have shifted toward unified LLM-based frameworks for multi-criteria evaluation (Chiang and Lee, 2023a; Liu et al., 2024a; Gong and Mao, 2023; Liu et al., 2024b; Li et al., 2025), which our work builds upon.

LLM-based NLG Evaluation. Current unified frameworks enhance generalizability either by prompting LLMs with criteria-centered templates (Fu et al., 2024; Liu et al., 2023a; Lin and Chen, 2023; Chiang and Lee, 2023b; Li et al., 2025; Liu et al., 2024b; Yuan et al., 2023) or fine-tuning them on multi-scenario benchmarks (Zhong et al., 2022; Li et al., 2023; Wang et al., 2023a; Ke et al., 2023; Kim et al., 2023; Hu et al., 2024; Xiao et al., 2023; Gao et al., 2024). While prompting is cost-efficient, it is sensitive to manually crafted prompts and often assumes fixed task-specific criteria. several latest works have explored instructing LLMs to generate evaluation criteria (Liu et al., 2024b; Li et al., 2025) or scoring rubrics (Liu et al., 2024a) based on pre-defined context like the target NLG task description. In essence, criteria generation in these studies implicitly assumes that each NLG task has a fixed set of evaluation criteria. In contrast, we argue that different end-user needs may lead to varying emphases, even for the same NLG task, resulting in criterion and/or rubric variation. To address this, our approach takes a user-centered perspective, instructing the LLM for NLG evaluations through self-inference of all relevant contexts.

Dynamic Prompt Optimization. Existing prompt optimization methods can be divided into two categories based on their inference depth. Single-layer optimization methods, such as APE (Zhou et al., 2023), APO (Pryzant et al., 2023), OPRO (Yang et al., 2023), and IPC (Levi et al., 2024), focus on optimizing prompts within a single stage, limiting their adaptability to complex tasks. In contrast, multi-layer optimization methods, like DSPy (Khattab et al., 2023) and MIPRO (Opsahl-Ong et al., 2024), refine prompts across multiple stages, supporting more comprehensive reasoning. We design a correlation-based comparison to optimize multi-stage NLG evaluation tasks.

3 Notations and Problem Definition

Our goal is to develop a highly adaptive NLG evaluation approach that can dynamically align with diverse end-user preferences to make explainable judgments across diverse NLG scenarios. Specifically, given a small set of source-response-quality tuples $\mathcal{D} = \{(x_i, y_i, r_i)\}_{i=1}^{N}$ annotated by humans based on their hidden criteria $\mathcal{C} = \{c_1, ..., c_k\}$, we aim to build an LLM-based reference-free evaluator $E(x', y')$. This evaluator learns from the annotated dataset $\mathcal{D}$ to infer task-relevant information, including the target NLG task description T and the evaluation criteria $\hat{\mathcal{C}} = \{\hat{c}_1, \ldots, \hat{c}_m\}$. Using this inferred information, it can estimate the quality score $\hat{r}$ of the source-response pair (x', y'), along with a free-text justification $\hat{e}$. Here, x_i denotes the i-th input text from the original NLG task, while y_i denotes the corresponding response generated by an NLG system and r_i is the quality score of y_i. We denote LLM(`[prompt]`) $\rightarrow$ `[response]` as the response generation by LLM given a prompt.

4 ACTIVE-CRITIC

Overview. Figure 1 shows the overall workflow of ACTIVE-CRITIC. With the motivation that an ideal unified evaluation framework should flexibly capture and align to human preferences across diverse generation scenarios, both by uncovering user-prioritized evaluation criteria and by making human-aligned judgments, we design a user-centered evaluation framework structured in two stages. The first stage is *task inference* (§4.1), where we instruct an LLM to predict task-related information(i.e. task and critria) by actively reviewing a small set of human-rated data examples. Through this analysis of the human-rated data, we expect the model to self-infer the details of the target evaluation task and the implicit criteria used by human annotators. The second stage is *scoring alignment* (§4.2), where we aim to align the LLM evaluator with human scoring based on the predicted evaluation criteria. Specifically, we design a dynamic prompt optimization method to automatically select the optimal few-shot examples, $\mathcal{D}_{\text{demo}}$, from $\mathcal{D}$ which enables the LLM evaluator to achieve human-aligned scoring through in-context prediction.

4.1 Task Inference

The Task Inference stage, depicted on the left side of Figure 1, focuses on identifying two key components for NLG evaluation: (1) *task description* and (2) *criteria definition*. This stage uses the LLM to analyze the dataset $\mathcal{D}_{\text{train}}$, infer the characteristics of the NLG task, and establish relevant evaluation criteria without human intervention.

Task Description. This module instructs the LLM to formulate an accurate task description T by reviewing examples in $\mathcal{D}_{\text{train}}$ and identifying key information that characterizes the target NLG task (e.g., summarization, storytelling) for evaluation. Considering that LLM's context length limit may not fit in all examples in $\mathcal{D}_{\text{train}}$, we split these examples into N mini-batches, and generate one task description T_n from each mini-batch $\mathcal{D}_{\text{train},n}$. That is, $\text{LLM}(f_t(\mathcal{D}_{\text{train},n})) \rightarrow T_n,\ \forall n \in [1, N]$, where f_t is a prompt template shown in Table 14 in the appendix. The final task description T is generated by the LLM through the ensemble of all task descriptions $\{T_n\}_{n=1}^{N}$ over all mini-batches.

Criteria Definition. After establishing the task description T, we apply prompt templates(f_{init}, f_{obs}, and f_{refine})[1], each instantiated with T, to guide the LLM in iteratively inferring evaluation criteria from $\mathcal{D}_{\text{train}}$. As shown in Algorithm 1, the generation process operates over mini-batches and follows three steps: generating initial observations Obs_0

[1]We instruct the LLM to output a criteria set in the JSON format, as shown in Table 15 in the appendix.

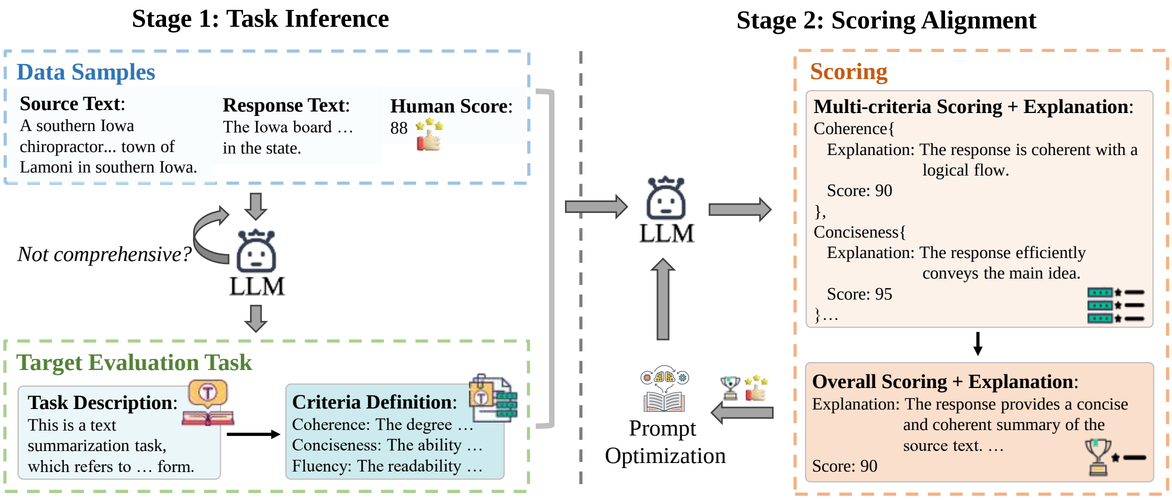

Figure 1: Overview of ACTIVE-CRITIC, including two stages: (1) task inference, where the LLM is instructed to derive the target NLG evaluation task description and relevant criteria from data samples, and (2) scoring alignment, allowing the LLM to generate multi-criteria and overall quality scores along with accompanying explanations.

and criteria C_0 based on the first mini-batch $\mathcal{D}_0$, observing subsequent batches to assess prior findings (C_t, Obs_t) and generate updated observations Obs_{t+1}, and progressively refining the criteria set C_{t+1} based on the latest observation Obs_{t+1} to better reflect the underlying human judgment patterns.

To enhance efficiency, we instruct the LLM to decide whether to stop early based on the comprehensiveness of the generated task description and criteria set after processing each mini-batch. Specifically, early stopping is triggered when the LLM outputs "COMPLETE" five times consecutively, or when it reaches the maximum limit of 25 iterations, whichever comes first. Once the stopping condition is met, we obtain the final criteria set, denoted as $\hat{C} = \{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_m\}$.

4.2 Scoring Alignment

Our second stage, as shown on the right side of Figure 1, focuses on aligning the LLM evaluator with human scoring judgments by automatically optimizing the evaluation prompts. Inspired by prior research that harnesses the potential of LLMs by breaking down complex tasks into simpler ones (Wei et al., 2022; Khot et al., 2023), we hypothesize that starting with fine-grained, criteria-specific scoring can help the model further derive an accurate overall quality score. With this intuition in mind, we structure the scoring stage into two modules: (1) *Multi-criteria Scoring with Explanation (McS-E)*, followed by (2) *Overall Scoring with Explanation (OS-E)*.

Multi-criteria Scoring with Explanation (McS-E). In this module, we use the LLM to assess the model output y_i based on the criteria set $\hat{C} = \{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_m\}$ obtained from the *task inference* stage (§4.1). Specifically, for each input-output pair (x_i, y_i), the LLM is instructed to estimate a score $\hat{r}_{ij}$ and a corresponding explanation e_{ij} according to each criterion $\hat{c}_j \in \hat{C}$:

$$\text{LLM}(x_i, y_i, f_{\text{McS-E}}(T, C, \mathcal{D}_{\text{demo}})) \rightarrow \hat{R}_i \quad (1)$$

$$\hat{R}_i = \{(\hat{r}_{ij}, \hat{e}_{ij}), \ \forall \hat{c}_j \in \hat{C}\} \quad (2)$$

where the output uses a JSON format, indicating a set of score-explanation pairs $\hat{R}_i$ for all criteria in $\hat{C}$ and $\mathcal{D}_{\text{demo}}$ is a set of demonstration examples randomly selected from the training set $\mathcal{D}_{\text{train}}$. This mechanism ensures that the evaluation is both quantitative and interpretable, offering insights into the rationale behind each score. The prompt template $f_{\text{McS-E}}(T, C, \mathcal{D}_{\text{demo}})$ is designed to enable scoring across multiple criteria simultaneously, accounting for the interconnections between them. This design enables a fine-grained evaluation, where each criterion is treated both individually and in connection with the others, providing detailed explanations that enhance the interpretability of the scoring process.

Overall Scoring with Explanation (OS-E). After scoring the individual criteria, we use a prompt template $f_{\text{OS-E}}$ to instruct the LLM to synthesize these scores $\{\hat{r}_{i1}, ...\hat{r}_{im}\}$ into an overall quality score $\hat{r}_i$, and an explanation e_i that provides a comprehensive justification for the final decision.

$$\text{LLM}(x_i, y_i, f_{\text{OS-E}}(T, \hat{R}_i, \mathcal{D}_{\text{demo}})) \rightarrow \hat{r}_i, \hat{e}_i \quad (3)$$

Prompt Optimization. Given the sensitivity of LLMs' in-context prediction performance to the few-shot examples $\mathcal{D}_{\text{demo}}$ in the prompt, we further propose an automatic prompt optimization strategy to iteratively select the optimal $\mathcal{D}^*_{\text{demo}}$ to refine the prompts. Specifically, given two lists of overall quality scores across all examples in $\mathcal{D}_{\text{train}}$—one predicted by the LLM, i.e., $\hat{r} = [\hat{r}_1, \ldots, \hat{r}_N]$ from Eq. (3), and the other annotated by humans, i.e., $r = [r_1, \ldots, r_N]$—we design an objective function to maximize the correlation between these two score lists. To mitigate potential biases caused by relying on a single correlation measurement, we calculate the sum of three widely-used correlation coefficients: Pearson (γ), Spearman (ρ), and Kendall (τ) with equal weights:

$$\mathcal{Q}(\hat{r}, r) = \gamma(\hat{r}, r) + \rho(\hat{r}, r) + \tau(\hat{r}, r) \quad (4)$$

$$\mathcal{D}^*_{\text{demo}} = \arg \max_{\mathcal{D}_{\text{demo}} \subset \mathcal{D}} \mathcal{Q}(\hat{r}, r) \quad (5)$$

where $\mathcal{D}_{\text{demo}}$ is the optimal few-shot demonstration examples $\mathcal{D}_{\text{demo}}$ selected from $\mathcal{D}_{\text{train}}$. To approximately solve the above maximization problem, we repeat K time for the evaluations of Eq. (3) using different randomly sampled $\mathcal{D}_{\text{demo}}$, and select the best $\mathcal{D}^*_{\text{demo}}$ that maximizes $\mathcal{Q}(\hat{r}, r)$. This optimization also enhances robustness by reducing the influence of outlier examples in $\mathcal{D}_{\text{train}}$, leading to more stable and human-aligned predictions.

5 Experiment Settings

Benchmarks Following prior work (Zhong et al., 2022; Fu et al., 2024; Liu et al., 2023a), we evaluate our method on four popularly-used benchmarks. These datasets cover diverse topics (e.g., politics, sports, restaurants, etc.) across four NLG tasks (i.e., summarization (Fabbri et al., 2021), dialogue generation (Mehri and Eskenazi, 2020), data-to-text generation (Wen et al., 2015), and storytelling (Guan et al., 2021)), aiming to construct a robust testbed to access ACTIVE-CRITIC. Details of each benchmark are provided in Appendix D.1.

We standardize all benchmarks into a uniform format that includes: (1) the machine-generated responses for evaluation, (2) the source input used by the generation systems for response generation, and (3) the human scores assessing response quality. Following prior work (Mahmoudi, 2023; Kocmi and Federmann, 2023; Lin and Chen, 2023; Chen et al., 2023), we normalize all human scores to a unified 0–100 scale to address the inconsistency of rating scales across NLG datasets (e.g., 1–5 in SummEval, 1–6 in SFRES).

Meta-evaluation We establish ACTIVE-CRITIC using four widely adopted backbone models: two open-source LLMs (Orca2-13B and LLaMA3-8B) and two closed-source LLMs (GPT-3.5 and GPT-4)[2] across four diverse NLG tasks. We test two variants of ACTIVE-CRITIC (AC) in this study: (1) **AC-Coarse** performs a coarse-grained, explainable evaluation by prompting the LLM to infer task-related information and directly produce an overall score along with an explanation for each test case. This process considers all inferred criteria at once during scoring alignment. (2) **AC-Fine** provides a fine-grained, explainable evaluation. Similar to AC-Coarse, it begins with task inference, but during scoring alignment, it assesses the input test case against each criterion individually, offering detailed explanations for each score. The overall quality score is then generated by combining the evaluations across all criteria. Appendix D.2 provides the details of implementation.

Baselines and Metrics We compare ACTIVE-CRITIC with a variety of state-of-the-art publicly accessible NLG evaluation methods. The baselines are grouped into two categories: (1) fine-tuning-based methods including Auto-J (Li et al., 2023), UniEval (Zhong et al., 2022), InstructScore (Xu et al., 2023), TIGERScore (Jiang et al., 2023) and TIGERScore (continued), which continues training from the original TIGERScore using the same 25% training split; and (2) prompting-based methods, including GPTScore (Fu et al., 2024), G-eval (Liu et al., 2023a), HD-Eval (Liu et al., 2024b), METAMETRICS (Winata et al., 2024) and four selected base LLMs under the zero-shot manner, implemented following (Mahmoudi, 2023). Following prior work (Fu et al., 2024; Jiang et al., 2023), we use GPTScore-src to refer to the source-hypothesis scoring type. To ensure fair comparison, all methods are evaluated on the same 75% testing split. Among them, TIGERScore (continued), HD-Eval, and METAMETRICS are iteratively trained on the same 25% training split.

Regarding metrics, we use three correlation coefficients to assess the evaluation consistency between machine-based evaluators and humans: Pearson (γ) (Mukaka, 2012), Spearman (ρ) (Zar, 2005) and Kendall-Tau (τ) (Kendall, 1938).

[2] We used GPT-3.5-turbo-1106 and gpt-4-turbo version for the experiments.

	SummEval			TopicalChat			SFRES			OpenMEVA (ROC)			Average
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	
InstructScore	0.3496	0.2703	0.203	0.2691	0.2774	0.2423	0.2039	0.1502	0.133	0.2936	0.2772	0.1658	0.2363
Auto-J	0.1345	0.1457	0.1149	0.4681	0.459	0.3714	0.1315	0.1053	0.0869	0.3896	0.3704	0.3065	0.257
TIGERScore	0.458	0.3694	0.2937	0.3785	0.4401	0.3458	0.1898	0.1246	0.1075	0.451	0.4413	0.3356	0.3279
UniEval	0.5457	0.4914	0.3707	0.5133	0.5448	0.4134	0.3247	0.2791	0.2081	0.4501	0.4408	0.3119	0.4078
GPTScore-src (FLAN-T5)	0.4043	0.3584	0.2696	0.2313	0.2437	0.1792	0.2819	0.2082	0.1618	0.2283	0.2265	0.1534	0.2456
Zero-shot (LLaMA3-8B)	0.4104	0.3857	0.2809	0.5197	0.5242	0.4018	0.2138	0.196	0.152	0.4141	0.3676	0.2808	0.3456
Zero-shot (Orca2-13B)	0.5447	0.4916	0.3999	0.5542	0.5512	0.4476	0.3068	0.23	0.1842	0.4809	0.4695	0.358	0.4182
G-eval (GPT-3.5)	0.4687	0.4504	0.3745	0.5427	0.5597	0.4501	0.2464	0.1956	0.1591	0.362	0.3408	0.1982	0.3624
Zero-shot (GPT-3.5)	0.453	0.385	0.292	0.5503	0.5436	0.4231	0.2823	0.2274	0.1828	0.4229	0.397	0.3	0.3716
G-eval (GPT-4)	0.6323	0.5697	0.437	0.6921	0.6975	0.596	0.3412	0.2868	0.2133	0.3901	0.3622	0.2732	0.4576
Zero-shot (GPT-4)	0.5943	0.5038	0.4055	0.6659	0.656	0.4937	0.3301	0.2823	0.2284	0.5627	0.4928	0.3777	0.4661
Iterative training on 25% training split													
TIGERScore (continued)	0.4832	0.3865	0.3059	0.454	0.4863	0.3976	0.2238	0.1568	0.137	0.4808	0.4799	0.3604	0.3593
HD-Eval (LLaMa3-8B)	0.5104	0.4765	0.3796	0.4725	0.5087	0.413	0.18	0.1546	0.1407	0.4261	0.427	0.3371	0.3688
HD-Eval (Orca2-13B)	0.523	0.5076	0.4036	0.5933	0.5994	0.4689	0.3307	0.2595	0.2124	0.4766	0.4818	0.3517	0.434
Ours:													
AC-COARSE (LLaMA3-8B)	0.5307	0.4972	0.3958	0.4873	0.5246	0.4259	0.1853	0.1594	0.1451	0.4394	0.4403	0.3477	0.3816
AC-FINE (LLaMA3-8B)	0.5334	0.502	0.401	0.5321	0.5379	0.4045	0.2265	0.2245	0.169	0.4506	0.4436	0.3625	0.399
AC-COARSE (Orca2-13B)	0.5386	0.5227	0.4156	**0.611**	0.6173	**0.4845**	**0.3612**	**0.2981**	**0.2393**	0.4908	0.4962	0.3622	0.4531
AC-FINE (Orca2-13B)	**0.6301**	**0.5486**	**0.4299**	0.6023	**0.6214**	0.4713	0.324	0.2834	0.2289	**0.5259**	**0.5363**	**0.4109**	**0.4677**
HD-Eval (GPT-3.5)	0.5628	0.5058	0.3927	0.603	0.6185	0.4926	0.3385	0.2694	0.2275	0.4583	0.4106	0.35	0.4358
HD-Eval (GPT-4)	0.6034	0.531	0.4274	0.6719	0.6932	0.5336	0.3577	0.284	0.2318	0.5584	0.5137	0.3902	0.483
METAMETRICS	0.6512	0.5843	0.4528	0.722	0.7338	0.6095	0.343	0.2849	0.2276	0.4268	0.4127	0.2975	0.4788
Ours:													
AC-COARSE (GPT-3.5)	0.6569	0.5368	0.4178	0.6425	0.6171	0.4855	0.3585	0.2846	0.2374	0.4185	0.3766	0.2981	0.4442
AC-FINE (GPT-3.5)	0.653	**0.6016**	**0.4745**	0.6718	0.6703	0.5156	0.3616	0.2833	0.2342	0.4693	0.4527	0.3442	0.4777
AC-COARSE (GPT-4)	0.6561	0.5371	0.4277	0.7264	**0.7815**	0.6133	0.343	**0.2878**	**0.2395**	0.5366	0.5226	0.4039	0.5063
AC-FINE (GPT-4)	**0.6926**	0.5723	0.462	**0.7789**	0.7753	**0.6212**	**0.363**	0.2809	0.236	**0.5877**	**0.5581**	**0.4249**	**0.5294**

Table 2: Correlation between LLM-based unified evaluators and human judgments on overall quality per instance across four NLG tasks. We compare Pearson (γ), Spearman (ρ) and Kendall-Tau (τ) correlation, respectively. The best performance per indicator is highlighted in bold, and the second-highest results are underlined. We implemented and tested all the methods with p-value < 0.05.

6 Results and Analysis

6.1 How well does ACTIVE-CRITIC perform?

Table 2 displays the correlation between unified evaluators and human judgments. Overall, ACTIVE-CRITIC noticeably outperforms the corresponding prompting-based baselines and state-of-the-art fine-tuning-based evaluators, where our variants built on Orca2-13B and GPT-4 achieve the highest correlation in the methods using open- and close-source LLM, respectively. Comparing two variants of ACTIVE-CRITIC per LLM, we find that the fine-level variant consistently achieves higher alignment with human scores, outperforming the coarse one by $\sim$2% in average correlation. These results show that our approach can effectively enhance LLMs' potential to capture human-centric assessment nuances in diverse scenarios and make more human-aligned judgments. Moreover, prompting the LLM to assess each criterion individually and then aggregate the scores benefits ACTIVE-CRITIC's decision-making.

We also validate the generalizability of ACTIVE-CRITIC's self-inferred evaluation prompts on unseen data within a similar NLG scenario. We use the prompts generated by ACTIVE-CRITIC on SummEVAL examples to evaluate responses from the Newsroom (Grusky et al., 2018) benchmark, similarly focusing on news article summarization but using diverse strategies. As shown in Appendix E, our approach outperforms state-of-the-art baselines, demonstrating strong generalization.

Further examining our approach's stability across base LLMs, we observe that ACTIVE-CRITIC consistently achieves a noticeable improvement, with an average gain of $\sim$6.8% correlation over the zero-shot baseline for each base model. This indicates its effectiveness, regardless of the chosen base LLM. Although ACTIVE-CRITIC generally obtains greater enhancements when employing a stronger base LLM, it is noteworthy that ACTIVE-CRITIC built on Orca2-13B performs comparably to its GPT-4 counterpart on SFRES and OpenMEVA (ROC). Considering the computational cost and evaluator performance, we primarily focus on the Orca2-based AC-Fine for further analysis.

6.2 Ablation Study

Dependence on Human-scored Data. To examine the impact of labeled data size on ACTIVE-CRITIC's performance, we varied the size of the

	SummEval			TopicalChat			SFRES			OpenMEVA (ROC)			Average
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	
Ours (AC-Fine)	**0.6301**	**0.5486**	**0.4299**	<u>0.6023</u>	**0.6214**	<u>0.4713</u>	<u>0.324</u>	<u>0.2834</u>	<u>0.2289</u>	**0.5259**	**0.5363**	**0.4109**	**0.4677**
w/o Task Description	0.5825	0.4826	0.3552	0.4949	0.5057	0.4211	0.2683	0.2017	0.168	0.3846	0.3802	0.2918	0.3781
w/o Criteria Definition	0.5726	0.522	0.4062	0.5533	0.5368	0.4451	0.293	0.2715	0.1907	0.4176	0.4237	0.326	0.4132
w/o McS-E	0.5386	<u>0.5227</u>	<u>0.4156</u>	**0.611**	<u>0.6173</u>	**0.4845**	**0.3612**	**0.2981**	**0.2393**	0.4908	<u>0.4962</u>	0.3622	<u>0.4531</u>
w/o OS-E	<u>0.6106</u>	0.5129	0.3908	0.5639	0.5615	0.4464	0.3165	0.2405	0.1899	<u>0.509</u>	0.4931	<u>0.3632</u>	0.4332

Table 3: Ablation study of key modules in ACTIVE-CRITIC.

feed examples from 5-shot to 5%, 15%, and the full 25% of each benchmark. Figure 2 shows the results. While ACTIVE-CRITIC improves as the labeled data size increases, it can achieve a decent correlation with human evaluators using as few as five human-rated examples. Among four tasks, ACTIVE-CRITIC is more sensitive to labeled data size in TopicalChat and SummEval than in OpenMeva and SFRES. The former two benchmarks involve longer contexts and diverse topics, while the latter focus on specific topics with shorter contexts, making the first two tasks more complex. Our observations suggest that ACTIVE-CRITIC requires more labeled data for evaluating complex NLG tasks compared to simpler ones.

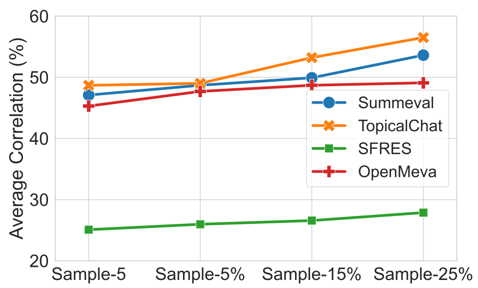

Figure 2: Average correlation between Orca2-based ACTIVE-CRITIC and human judgments with varying label sizes. Results for each correlation coefficient are provided in Appendix F

Module Contribution. Table 3 shows the individual contribution of each module in ACTIVE-CRITIC. Note that the variant w/o criteria inference uses the original predefined criteria from each benchmark for further computation. This results in a notable performance drop (∼5%), highlighting that our automatically inferred criteria achieve better alignment with human judgments than the manually defined ones. Additionally, In the variant w/o OS-E, we calculated the overall quality score per test case by averaging the multiple criteria-specific scores generated by McS-E. The larger performance drop in the variant w/o OS-E, compared to the one w/o McS-E, indicates that the LLM-generated overall quality score contributes more meaningfully than simply averaging the criteria-specific scores. Interestingly, on the SFRES dataset, the w/o McS-E variant slightly outperforms the full model, likely due to the narrow score distribution (∼80% of ratings between 4–6), where multi-criteria scoring may over-amplify subtle differences and reduce alignment with human judgment.

Impact of Optimization. We compare ACTIVE-CRITIC's performance by removing its dynamic prompt optimization for scoring and, furthermore, eliminating mini-batch iterations during task inference. As shown in Figure 3, there is a drop in ACTIVE-CRITIC's performance when removing scoring prompt optimization, with a further decline when only using a single mini-batch of labeled data for task inference, suggesting that both strategies contribute to ACTIVE-CRITIC for making optimal decisions. Interestingly, ACTIVE-CRITIC is more sensitive to scoring optimization in the fine-level setting of SummEval and the coarse-level setting of SFRES, highlighting its greater impact in these scenarios. In contrast, mini-batch iterations have limited effect in SummEval, indicating that ACTIVE-CRITIC can infer the evaluation task effectively with minimal training data.

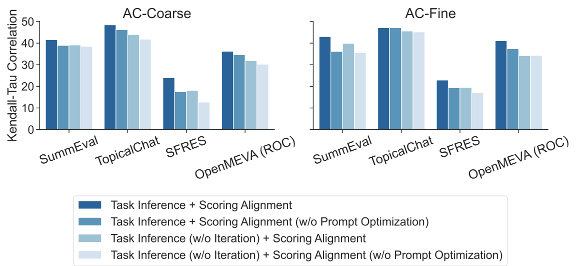

Figure 3: Impact of prompt optimization on scoring and mini-batch iterations on task inference (Kendall-Tau %). See Appendix G for Pearson and Spearman results.

LLM-inferred Criteria Analysis. Moving forward from quantitative analysis, we examine the LLM-inferred criteria in depth. Table 6 shows an illustrative comparison between the criteria generated by ACTIVE-CRITIC and those pre-defined by

Method + Mode	Inferred Criteria
AC (OO)	Relevance, Conciseness, Coherence, Accuracy, Completeness, Objectivity, Fluency and Correctness
AC (FO)	Fluency, Grammaticality, Spelling and Punctuation, Surface-level Coherence, Sentence Structure
HD-Eval (OO)	Sentence ordering, Discourse structure, Topic focus, Factuality, Entity consistency, Temporal consistency, Grammar, Spelling, Readability, Coverage, Redundancy, Novelty, Factual Consistency, Factual Coverage, Factual Relevance, Factual Accuracy, Temporal Coverage, Temporal Relevance
HD-Eval (FO)	Relevance, Coverage, Accuracy, Coherence, Conciseness, Readability, Fluency, Grammar and Syntax, Clarity, Consistency, Synthesis, Novelty, Length Appropriateness, Complexity of Vocabulary, Smoothness, Ease of Understanding, Syntax Accuracy, Spelling, Subject-Verb Agreement, Structure and Organization

Table 4: Evaluation criteria inferred by our ACTIVE-CRITIC (AC) and HD-EVAL under the Original-Overall (OO) and Fluency-as-Overall (FO) modes. HD-Eval (FO) does not provide initial dimensions (e.g., coherence, consistency, fluency, and relevance). Detailed evaluation criteria are provided in Table 18.

humans in SummEval. We find that our approach incorporates more nuanced criteria (i.e., "clarity", "conciseness", "coverage", and "engagement") beyond the four pre-defined aspects. Moreover, each criterion is paired with a clear definition to specify its distinct characteristics. For example, the human-defined "coherence" starts with a high-level description like "well-structured and well-organized", while the LLM's definition tends to be more concrete, e.g., "the summary flows logically".

6.3 Can ACTIVE-CRITIC capture user-centered evaluation criteria?

We simulate a user preference scenario by selecting a subset from SummEval where *fluency* is rated high ($\geq$ 4.5) and other dimensions (coherence, consistency, relevance) are low ($<$ 3.0). This yields 180 examples (45 train / 135 test), under two settings: (1) **Original-Overall (OO)** using the original human-provided overall scores, and (2) **Fluency-as-Overall (FO)** using fluency replaces the overall score as the supervision signal. We compare ACTIVE-CRITIC with a strong baseline (HD-Eval) under these two modes.

Qualitative Analysis. As shown in Table 4, ACTIVE-CRITIC adapts its inferred criteria according to the supervision mode, capturing general dimensions(e.g., relevance, fluency) in OO and focusing on linguistic quality (e.g., grammaticality, sentence structure) in FO, demonstrating clear sensitivity to user-defined preferences. In contrast, HD-Eval produces a broad but relatively fixed set of criteria in both modes, combining fluency-related and unrelated dimensions.

Quantitative Analysis. As shown in Table 5, we observe that correlation scores for both ACTIVE-CRITIC and HD-Eval drop notably under the FO setting compared to the OO, suggesting that fluency-focused evaluation is inherently more challenging to model. Nevertheless, ACTIVE-CRITIC consistently outperforms the HD-Eval baseline across both settings ($\sim$12% in OO and $\sim$4% in FO), indicating its robustness and adaptability in aligning with user-set preferences.

Method	γ	ρ	τ	AVE
HD-Eval (OO)	0.3395	0.3556	0.2921	0.3291
AC (OO)	**0.3927**	**0.4168**	**0.3014**	**0.3703**
HD-Eval (FO)	0.1529	0.1628	0.1607	0.1588
AC (FO)	**0.1821**	**0.1787**	**0.1648**	**0.1657**

Table 5: Quantitative results of ACTIVE-CRITIC and HD-EVAL under the Original-Overall (OO) and Fluency-as-Overall (FO) modes.

7 Conclusion

We proposed ACTIVE-CRITIC, a novel LLM-based NLG evaluation protocol that relies solely on lightweight human-scored data. Unlike existing machine-based evaluators that depend on human-predefined task-related information for assessment, ACTIVE-CRITIC self-identifies the target evaluation task and nuanced evaluation criteria purely from the data for making judgments. This paradigm shift will enhance the adaptability of ACTIVE-CRITIC, enabling it to flexibly capture the varying priority expectations of different end-users across diverse generation scenarios. Our approach reduces the need for intensive manual efforts to design task-specific criteria and extensive prompt engineering. Experiments across four distinct NLG tasks demonstrate LLMs' potential as active critics, achieving higher correlation with human judgments compared to baselines. Fine-level criteria-specific scoring, paired with explanations, prompts the LLM to engage more deeply with the test cases, leading to improved overall quality scoring.

8 Limitation

Below, we make an elaborate discussion about the current limitations of ACTIVE-CRITIC and outline potential directions for future improvement:

1) Due to resource constraints, our meta-evaluation experiments primarily focused on four representative NLG tasks and four backbone LLMs (including both open- and closed-source models). While this setup effectively demonstrated the utility of our framework, applying ACTIVE-CRITIC across a wider range of tasks, benchmarks, and language models would provide a more comprehensive picture of its generalizability and robustness. We believe expanding this scope is a key step toward developing broadly applicable evaluation protocols for NLG.

2) We prompt the LLM to generate explanations alongside each predicted score, offering a basic level of transparency. While this is a useful first step, it falls short of providing structured or fine-grained reasoning. Future efforts could focus on integrating more explicit explanation mechanisms to improve the interpretability and trustworthiness of the evaluation process.

3) Our early stopping strategy, though effective in practice, is based on simple heuristics designed for general use. In reality, task-specific or dataset-specific dynamics may benefit from more adaptive stopping criteria or data-driven thresholding methods. Investigating automated alternatives could improve both efficiency and adaptability across a broader range of applications.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374.

Cheng-Han Chiang and Hung-Yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

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.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

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*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free nlg evaluation language model with flexibility and interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. In *Proceedings* of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11017–11031.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

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

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

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115.

Elad Levi, Eli Brosh, and Matan Friedmann. 2024. Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, et al. 2023. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy Chen, and Min-Yen Kan. 2025. Dna-eval: Enhancing large language model evaluation through decomposition and aggregation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2277–2290.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yixin Liu, Alexander Richard Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024a. Calibrating llm-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024b. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. *arXiv preprint arXiv:2402.15754*.

Ghazaleh Mahmoudi. 2023. Exploring prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 219–227.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4):349–373.

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

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *arXiv preprint arXiv:2406.11695*.

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.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025. Assessing the human likeness of AI-generated counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3547–3559, Abu Dhabi, UAE. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. 2023a. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665.

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.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Genta Indra Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Tanti Wijaya. 2024. Metametrics: Calibrating metrics for generation tasks using human preferences. *arXiv preprint arXiv:2410.02381*.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994.

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

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

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations(ICLR)*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A An example of evaluation protocol and prompt on SummEval

A.1 An example of input data

This section shows an example of data (x_i, y_i, r_i) from SummEval in Table 10.

A.2 An Example of Output

This section shows an example of output with our AC-Fine method in table 11.

A.3 Instruction component of the evaluation protocol

This section presents the instruction I in evaluation protocol Φ, which is also the output of the task inference module, in table 12

A.4 In-context exemplar of the evaluation protocol

This section presents the in-context exemplar D_{demo} in evaluation protocol Φ in table 13

A.5 Prompt Template

This section presents prompt templates in multiple stages: (1) Task Description (Table 14), (2) Criteria Definition (Table 15), (3) Multi-Criteria Scoring with Explanation (Table 16), and (4) Overall Scoring with Explanation (Table 17).

B Criteria Generation Algorithm

In the initial generation step, we use f_{init} to prompt the LLM with the first mini-batch $\mathcal{D}_0$, producing an initial criteria set C_0 and corresponding observations Obs_0:

$$LLM(f_{\text{init}}(\mathcal{D}_0)) \rightarrow Obs_0, C_0 \tag{6}$$

In the observation step, given the prompt f_{obs}, the LLM reviews new batch data $\mathcal{D}_{t+1}$ and assess prior findings (C_t, Obs_t), and generate updated observations Obs_{t+1}:

$$LLM(f_{\text{obs}}(Obs_t, C_t, \mathcal{D}_{t+1})) \rightarrow Obs_{t+1} \tag{7}$$

In the refinement step, where we apply f_{refine} to generate a revised criteria set C_{t+1} based on the latest observation Obs_{t+1}:

$$LLM(f_{\text{refine}}(C_t, Obs_{t+1})) \rightarrow C_{t+1} \tag{8}$$

Algorithm 1 Criteria generation algorithm

Require: Dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i, r_i)\}_{i=1}^{N}$, number of mini-batches N
Require: Model LLM, prompts $\{f_{\text{init}}, f_{\text{obs}}, f_{\text{refine}}\}$, stop condition stop$(\cdot)$

```
D_n ← {D_train,n}, ∀n ∈ [1, N]                    ▷ Data on n-th batch
Obs_0, C_0 ← LLM(f_init(D_0))                     ▷ Initial generation
for iteration t = 0, 1, ... do
    Obs_{t+1} ← LLM(f_obs(Obs_t, C_t, D_{t+1}))   ▷ Observation
    if stop(Obs_t, t) then                        ▷ Stop condition
        break
    else
        C_{t+1} ← LLM(f_refine(C_t, Obs_t)))      ▷ Refine
    end if
end for
return C_{t+1}
```

(a) AC-Fine	(b) Human
Coherence: The degree to which the summary flows logically and cohesively, with clear connections between the main points. **Conciseness**: The ability of the summary to convey all necessary information in a succinct and efficient manner. **Coverage**: The extent to which the summary captures the main events and details from the source text without omitting crucial information. **Accuracy**: The faithfulness of the summary to accurately reflect the main points and details of the source text. **Fluency**: The readability and naturalness of the language used in the summary, with smooth transitions between ideas and paragraphs. **Relevance**: The relevance of the summary to the main topic and the inclusion of only pertinent information from the source text. **Clarity**: The clarity and comprehensibility of the summary, with clear and precise language used to convey the main points. **Engagement**: The ability of the summary to captivate and engage the reader, drawing them into the main events and details effectively.	**Coherence**: the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic. **Consistency**: the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. **Fluency**: the summary should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read. **Relevance**: The summary should include only important information from the source document.

Table 6: An illustrative example of the generated evaluation criteria on SummEval, either generated by an ACTIVE-CRITIC (a) or predefined by humans (b). The highlighted text in blue are additional criteria generated by the machine compared to the human-defined ones.

C Detailed Evaluation Criteria

D Details of Experiment Settings

D.1 Details of Benchmark

- **SummEval** (Fabbri et al., 2021): 1,600 machine-generated summaries of CNN/Daily-Mail articles were rated by both expert and layman judges on coherence, consistency, fluency, relevance, and overall quality.

- **Topical-Chat** (Mehri and Eskenazi, 2020): A knowledge-grounded, open-domain dialogue dataset consisting of 60 conversations, each paired with 6 responses (2 by humans and 4 by machines). Responses are human-evaluated on overall quality across five dimensions: naturalness, coherence, engagingness, groundedness, and understandability.

- **SFRES** (Wen et al., 2015): A data-to-text generation benchmark with 1,181 instances, focusing on generating free-text utterances from structured restaurant information. Annotators rated the overall quality of each instance based on informativeness and naturalness.

- **OpenMEVA (ROC)** (Guan et al., 2021): 1,000 open-ended commonsense stories generated by various models trained upon the ROCStories corpus. Annotators rate each story based on fluency, creativity, and coherence.

D.2 Details of Parameter Setting and Implementation

Considering that no predefined split was available in these benchmark datasets, popularly used in LLM-based NLG evaluation, we follow prior work (Liu et al., 2024b; Winata et al., 2024) and adopt a 25%-75% train-test split to build ACTIVE-CRITIC. During task inference, we set the number of mini-batches to 25, with a batch size of 5. The LLM is instructed to generate one task description and a set of evaluation criteria per mini-batch. To enhance tuning efficiency, we allow the LLM to decide when to stop early, capping the number of task descriptions and criteria sets at 5. For the scoring stage, we run 11 epoches of prompt optimization. The number of in-context exemplars used per epoch is 3 for SummEval and TopicalChat, and 8 for SFRES and OpenMeVA (ROC), with the difference due to varying input text lengths across tasks. All parameter settings are based on empirical testing of sequential values to determine optimal configurations.

Our experiments were carried out using two NVIDIA V100 GPU cards. For prompt optimization in the scoring stage, we utilized the “BootstrapFewShotWithRandomSearch" method in DSPy (Khattab et al., 2023) as the optimizer, which leverages random search to generate examples.

E Generalization to Unseen Datasets

Ideally, we expect the ACTIVE-CRITIC-generated evaluation prompts can be directly used for NLG system assessment in a similar future NLG scenario. To assess the generalizability of these prompts, we use the prompts generated by ACTIVE-CRITIC based on SummEval examples to assess unseen cases in Newsroom (Grusky et al., 2018). This dataset comprises 60 news articles and their corresponding summaries generated by 7 summarization systems. Each summary is paired with an overall quality score provided by human annotators. Table 7 displays the results. Our ACTIVE-CRITIC noticeably outperforms baselines by $\sim$10% correlation on average, indicating ACTIVE-CRITIC's generalizability.

Method	γ	ρ	τ	AVE
TIGERScore	0.3731	0.41	0.3075	0.3635
UniEval	0.4485	0.4505	0.325	0.408
G-eval (gpt3.5)	0.3853	0.4053	0.3012	0.3639
GPT-3.5 (zero-shot)	0.504	0.561	0.430	0.4983
AC-FINE (GPT3.5)	**0.6382**	**0.6444**	**0.4949**	**0.5925**
GPT-4 (zero-shot)	0.6583	0.6649	0.4957	0.6063
AC-FINE (GPT4)	**0.7466**	**0.7111**	**0.5474**	**0.6684**

Table 7: Generalization results of ACTIVE-CRITIC on Unseen Datasets.

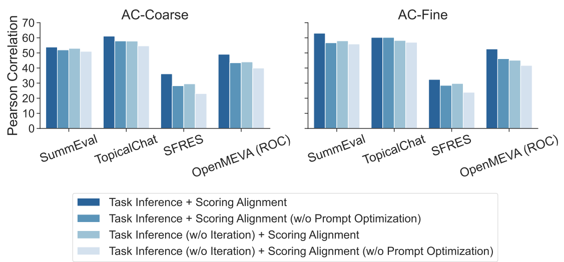

Figure 4: Effectiveness of Optimization. We report the Pearson (γ) correlation coefficient for our two optimal experimental variants: AC-Coarse and AC-Fine.

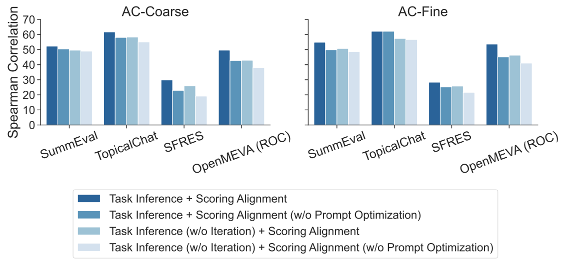

Figure 5: Effectiveness of Optimization. We report the Spearman (ρ) correlation coefficient for our two optimal experimental variants: AC-Coarse and AC-Fine.

F Additional Results of ACTIVE-CRITIC's Dependence on Human-scored Data

G Impact of Optimization by Pearson and Spearman Correlation

H Explainability Analysis

H.1 Helpfulness of Explanations

To assess the impact of explanations generated by ACTIVE-CRITIC, we compared our protocol's performance with versus without explanations, at both coarse and fine levels of evaluations. Figure 7 shows the results based on the Kendall-Tau correlation. We also provide the results of Pearson and Spearman correlation in Figure 8 and Figure 9 respectively.

As shown in Figures 7, ACTIVE-CRITIC with explanations consistently demonstrates a higher correlation with human judgments than the version without explanations. Notably, the difference in correlation is greater for the fine-level ACTIVE-CRITIC compared to the coarse-level variant. These findings suggest that generating explanations for scoring helps the base LLM engage more effectively in the evaluation process, resulting in stronger alignment with human judgments. In particular, fine-level explanations for each model-inferred criterion are especially effective in boosting the model's engagement and improving evaluation accuracy.

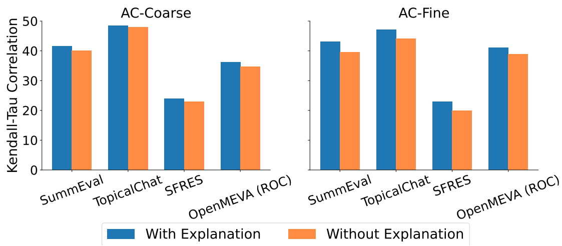

Figure 7: Effectiveness of Explanation in Kendall-Tau correlation (%).

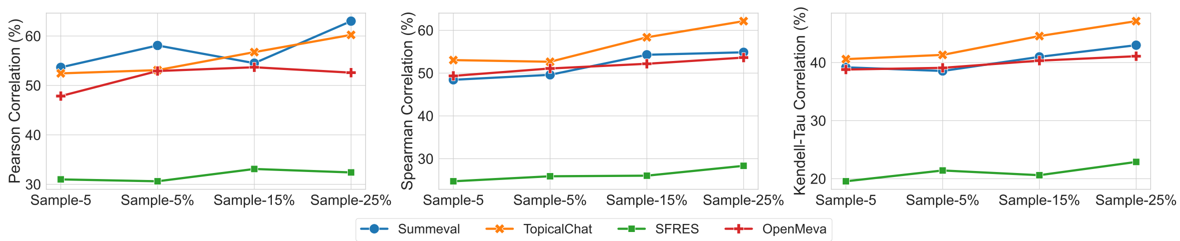

Figure 6: Results of ACTIVE-CRITIC's dependence on human-scored data by Pearson, Spearman, and Kendell-Tau, respectively.

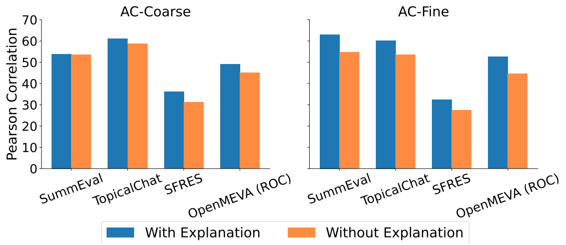

Figure 8: Effectiveness of Explanation in Pearson (%).

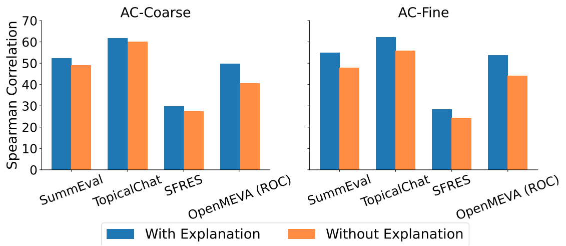

Figure 9: Effectiveness of Explanation in Spearman (%).

Dimension Rate→	**Clarity** Yes (%)	**Relevance** Yes (%)	**Score Consistency** Yes (%)	**Accuracy** Yes (%)
Coherence	99.11	92	95.78	85.33
Conciseness	98.67	91.78	96.89	88.89
Coverage	98.82	91.33	97.56	96.89
Accuracy	98.22	92.22	95.56	98
Fluency	99.56	98.89	96	96.67
Relevance	98.89	99.11	98.44	95.56
Clarity	98	94.22	93.56	95.78
Engagement	99.33	94.67	93.33	91.11
Overall Quality	98.44	98.44	97.33	98
Average	98.78	94.74	96.05	94.03

Table 8: Human evaluation of criterion-specific explanations on SummEval samples.

Dimension Rate→	**Aspect-to-Overall Alignment** Yes (%)	**Differentiability** Yes (%)	**Usefulness** (1-5)
Overall	95.11	90	4.515

Table 9: Human evaluation of overall explanations on SummEval samples, emphasizing (1) the alignment of the overall explanation with criterion-specific ones, (2) explanations' differentiability across vary-quality cases, and (3) explanations' overall usefulness per case.

H.2 Human Evaluation of Explanations.

We also employ three proficient English-speaking annotators to evaluate the quality of the scoring explanations generated by ACTIVE-CRITIC on a random sample of 150 test cases from SummEval. Our assessment consisted of four parts, with details provided below. First, for each individual explanation per case, each annotator rated the quality based on: (1) clarity of the statement, (2) relevance to the target criterion, (3) alignment with the corresponding score, and (4) accuracy within the context of the test case (e.g., correctness in matching the source text). Further emphasizing the overall scoring explanation per case, we asked annotators to assess its alignment with the criteria-specific explanations, and its differentiability across cases of varying quality, respectively. Finally, we asked annotators to provide an overall rating on a scale of 1-5 based on the usefulness of all generated explanations per case. To validate the reliability of human annotations, following prior work (Fabbri et al., 2021), we calculated intercoder reliability by Krippendorff's alpha (Krippendorff, 2011). The 0.6534 Kappa coefficient indicates substantial agreement among annotators.

As shown in Table 8, the individual explanations demonstrate comparatively high quality across four dimensions, with average scores of 98.78% for clarity, 94.74% for relevance, 96.05% for score consistency, and 94.03% for information accuracy. As shown in Table 9, the overall explanations generally align with the criteria-specific ones (95.11%), and 90% of the overall explanations effectively differentiate case quality. With an average rating of ∼4.5 out of 5 on the generated explanations across sampled testing cases, the result shows that explanations generated by ACTIVE-CRITIC are of good quality and useful to explain the resulting scores.

Source (x_i)
A southern Iowa chiropractor accused of accepting sex as payment for his services and performing exorcisms on patients has surrendered his state license. The Iowa Board of Chiropractic released a report Wednesday detailing charges against Charles Manuel, of Lamoni. Manuel signed an agreement last month admitting his misdeeds and pledging not to apply for reinstatement for at least 10 years. Patient satisfaction: A chiropractor in Iowa has surrendered his license to practice and admitted to swapping services for sex and performing exorcisms on some patients. Nonetheless, he's received outstanding evaluations for patient satisfaction on HealthGrades.com The agreement requires that Manuel prove any circumstances surrounding the revocation of his license no longer exist before he can resume practicing chiropractic in the state. Those circumstances included bartering sex for services with some patients. Manuel also recommended that patients stop taking medication he prescribed to them. A woman who answered a call to Manuel's home from the Des Moines Register declined to comment on the case. A woman at his former practice said he hadn't worked there for some time. A Lamoni address listed on the Yelp page for Manuel's practice appears to be a home on a residential street. While Maneul has received just three patient survey responses on HealthGrades.com, those responses were quite positive. The disgraced chiropractor received a perfect five out of five stars in patient satisfaction. Strange Practice: Charles Manuel, who admitted wrongdoing to the Iowa Board of Chiropractic, listed his practice's addresses on this residential street in the small agricultural town of Lamoni in southern Iowa.
Response (y_i)
The iowa board of chiropractic released a report wednesday detailing charges against charles manual, of lamoni. Manuel signed an agreement last month admitting his misdeeds and pledging not to apply for reinstatement for at least 10 years . The agreement requires that manuel prove any circumstances surrounding the revocation of his license no longer exist before he can resume practicing chiropractic in the state.
Human Judgment Score (r_i)
88.33333333333332

Table 10: An example of data (x_i, y_i, r_i) from SummEval

Multi-Criteria Scoring with Explanation

Coherence: {
explanation: The response is coherent and logically flows from one point to the next, with clear connections between the main points.
score: 90
}
Conciseness: {
explanation: The response is concise and efficiently conveys the main events and details from the source text.
score: 90
}
Coverage: {
explanation: The response covers the main events and details from the source text without omitting crucial information.
score: 85
}
Accuracy: {
explanation: The response accurately reflects the main points and details of the source text.
score: 90
}
Fluency:{
explanation: The response is fluent, with smooth transitions between ideas and paragraphs, and natural language used throughout.
score: 95
}
Relevance: {
explanation: The response is highly relevant to the main topic and includes pertinent information from the source text.
score: 95
}
Clarity: {
explanation: The response is clear and comprehensible, with precise language used to convey the main points effectively.
score: 90
}
Engagement: {
explanation: The response effectively draws the reader into the main events and details, making it highly engaging.
score: 95
}

Overall Scoring with Explanation
Explanation: The response provides a concise and coherent summary of the source text, capturing the main events and details without omitting crucial information. The summary is accurate, relevant, and clear, with smooth transitions between ideas and paragraphs, making it highly engaging and readable. The coverage of the main events and details is comprehensive, and the language used is clear and precise. Overall, the response effectively conveys the main points of the source text in a compelling and engaging manner. **Overall Score:** 90.0

Table 11: An example of output with our AC-Fine method

Task Description T:
The evaluation task is text summarization, which refers to evaluating abstractive summaries generated from source texts. The relationship between 'Source' and 'Response' is that the responses consistently provide a summary of the main events or details described in the source text and accurately reflect the main points of the source text in a summarized form.
Critiria C:
"Coherence": "The degree to which the summary flows logically and cohesively, with clear connections between the main points." "Conciseness": "The ability of the summary to convey all necessary information in a succinct and efficient manner." "Coverage": "The extent to which the summary captures the main events and details from the source text without omitting crucial information." "Accuracy": "The faithfulness of the summary to accurately reflect the main points and details of the source text." "Fluency": "The readability and naturalness of the language used in the summary, with smooth transitions between ideas and paragraphs." "Relevance": "The relevance of the summary to the main topic and the inclusion of only pertinent information from the source text." "Clarity": "The clarity and comprehensibility of the summary, with clear and precise language used to convey the main points." "Engagement": "The ability of the summary to captivate and engage the reader, drawing them into the main events and details effectively."

Table 12: An example of instruction I in evaluation protocol Φ

AC-Fine Output Example:

Source: “Paul Merson has restarted his row with Andros Townsend... Any bad feeling between the pair seemed to have passed but Merson was unable to resist having another dig at Townsend after Tottenham drew at Turf Moor."
Response: “Paul merson has restarted his row with andros townsend .. in the 83rd minutefor tottenham as they drew 0-0 against burnley."
Multiple Evaluation Criteria:
Coherence: The degree to which the summaryflows logically and cohesively, with clearconnections between the main point.
Conciseness: The ability of the summary to convey all necessaryinformation in a succinctand efficient manner.
...
Score Of Each Criterion In JSON:
Coherence: {
Explanation: The response is somewhat coherent, but it jumps between different events and details without clear connections between them.
Score: 60
}
...
Explanation: The response provides a concise summary ... to provide a more compelling and logically flowing summary.
Score of overall: 75

$\mathcal{D}_{train}$ **Example:**
Source: Chelsea have made an offer for FC... The initial five-year deal is the biggest in the club ’s history , with the Blues now considering a two-week pre-season tour of Japan this summer. **Response:** Chelsea have made an offer for fc ... in muto is not connected to the 200million sponsorship deal they signed with japanese company yokohama rubber in February. **"Score of Overall":** 91.66666666666666

Table 13: An example of in-context exemplar D_{demo}

Given several examples from an NLG evaluation dataset where each entry consists of a 'Source' text and its corresponding 'Response', along with a score that evaluates the response quality. Please write observations about trends that hold for most or all of the samples. I will also provide you with some previous observations I have already made. Please add your observations or if you feel the observations are comprehensive say 'COMPLETE'. Some areas you may consider in your observations: content and structure, scenario, task, evaluation objective, evaluation criteria, etc. It will be useful to make an educated guess as to the nature of the task this dataset will enable. Don't be afraid to be creative. $\{$*examples*$\}$ $\{$*prior observations*$\}$
Given a series of observations I have made and some description about this NLG evaluation dataset. 1. Identify the type of evaluation task. Possible tasks include: machine translation, text summarization, data-to-text generation, dialogue generation, image description, text simplification, story generation, paraphrase generation, textual entailment, reasoning, etc. 2. What this evaluation task refers to evaluating. 3. Output the relationship between 'Source' and 'Response' in this task in 1-3 sentences. 4. Given a summary in fill [ ]: The evaluation task is [ ], which refers to evaluating [ ] generated from [ ]. The relationship between 'Source' and 'Response' is [ ]. $\{$*observations*$\}$ $\{$*prior task description*$\}$

Table 14: Prompt template on Task Description

Given several examples from ${*task type*} evaluation dataset. Each example contains a Source text, a generated Response, and a Human Score.
Please write observations about the relationship between the response quality and the human score based on the given examples, and infer which evaluation criteria the human annotators most likely focused on when assigning these scores.
Analyze the trends across examples. Return the observed patterns and main evaluation criteria.
Examples:${*examples*}
Output in the following format:
{
"Criteria": {
"Criterion A": "Short description of Criterion A",
"Criterion B": "Short description of Criterion B",
... },
"Observations": "Observed patterns and insights about how human annotators likely evaluated the responses based on the examples." }

Given several examples from ${*task type*} evaluation dataset. Each example contains a Source text, a generated Response, and a Human Score. Your task is to review the existing observations and evaluation criteria. I will provide you with: - A batch of new examples - Previously generated observations - A current set of evaluation criteria believed to reflect human scoring Please assess whether: 1. The existing observations sufficiently capture patterns in the data. 2. The evaluation criteria accurately represent the dimensions human annotators focus on when assigning scores. If you feel the observations and criteria are already comprehensive and accurate, say 'COMPLETE'. Otherwise, provide a paragraph of new insights or refinements that should be considered. ${*examples*} ${*prior observations*} ${*current evaluation criteria*}
Given a set of evaluation criteria and new observations. Your task is to refine the current evaluation criteria based on the new observations. Output a refined set of 4 to 10 evaluation criteria in JSON format: criterion as key, description as value. ${*evaluation criteria*} ${*new observations*}

Table 15: Prompt template on Criteria Definition

${*Task Description*}
Your task is to evaluate the response on multiple evaluation criteria with respect to the source on a continuous scale from 0 to 100, and explain your process for scoring each criterion. Rate the response on multiple evaluation criteria and give a brief explanation in a JSON format by filling in the placeholders in [ ].

${*In-context exemplar*}

${*Source*}
${*Response*}
${*Multiple Evaluation Criteria*}

Output format:
Score Of Each Criterion In JSON:

```
{
Coherence: {
        Explanation: “[your explanation]”,
        Score: “[score from 0 to 100: 0 - No logic, 100 - Perfectly coherent]” },
Conciseness: {
        Explanation: “[your explanation]”,
        Score: “[score from 0 to 100: 0- Overly verbose, 100- Highly efficient]” },
...
}
```

Table 16: Prompt template on Multi-Criteria Scoring with Explanation

${*Task Description*}
Your task is to rate the quality of the response, based on the source and the scores for different criteria of the response on a continuous scale from 0 to 100, where 0 means ‘completely irrelevant and unclear’ and 100 means ‘perfectly relevant, clear, and engaging.’ IMPORTANT!! Only output the score as an ‘int’ and nothing else.
“Also explain your process to get this score to response. Also please perform error Analysis of given response. What should we change to have a better result?"

${*In-context exemplar*}

${*Source*}
${*Response*}
${*Score Of Different Criteria*}

Output format:

Explanation:
Score Of Overall:

Table 17: Prompt template on Overall Scoring with Explanation

AC (OO):
Relevance: Assesses whether the summary captures the main points and key information from the original article, focusing on the most important content and omitting trivial details. Conciseness: Evaluates if the summary conveys essential information using few words, avoiding redundancy and unnecessary elaboration. Coherence: Measures the logical flow and clarity of the summary, ensuring that ideas are well-organized and transitions between sentences are smooth. Accuracy: Checks whether the summary faithfully reflects the facts and nuances of the source, without introducing misinformation or hallucinations. Completeness: Determines whether all critical aspects of the original article are included, avoiding omissions of important content. Objectivity: Evaluates the neutrality of the summary, ensuring it presents information without personal opinions or bias, unless such tone is inherent in the source. Fluency and Correctness: Assesses the summary's grammatical correctness, spelling, punctuation, and overall readability, ensuring it reads naturally and professionally.
AC (FO):
Fluency: The summary should read smoothly and naturally, without awkward or disjointed phrasing. Grammaticality: The summary should follow standard grammar rules, with no major grammatical errors. Spelling and Punctuation: The summary should be free of spelling mistakes and should use punctuation correctly. Surface-level Coherence: The sentences should be logically ordered and connected, forming a cohesive whole. Sentence Structure: The summary should contain well-formed sentences with proper syntax and variation.
HD-Eval (OO):
Sentence ordering: how well the sentences in the summary follow a natural and logical order. Discourse structure: how well the summary uses discourse markers (such as however, therefore, etc.) to indicate the relations between sentences. Topic focus: how well the summary maintains a consistent topic throughout. Factuality: how well the summary preserves the factual information from the original article without introducing errors or distortions. ... Temporal Relevance: how well the summary is relevant to the source document in terms of the temporal information it presents.
HD-Eval (FO):
Relevance: Does the summary include the most important points of the original text? This checks if the summary focuses on the core topics and not on peripheral details. Coverage: How well does the summary cover key aspects of the original text? This involves checking if all significant points are mentioned. Accuracy: Are the facts and figures mentioned in the summary correct and consistent with the original text? Coherence: Does the summary flow logically from one point to another? This evaluates the logical sequence and connection between ideas. ... Structure and Organization: Is the summary well-organized and logically structured?

Table 18: Detailed evaluation criteria inferred by our ACTIVE-CRITIC (AC) and HD-EVAL under different modes

Human Eval for Explainations

I will provide you with instances from the SummEval dataset, each randomly selected and categorized into three score ranges: 0-50, 51-80, and 81-100, with 10 instances per category. Each instance includes a detailed evaluation of a summary response to a source text. The evaluation covers several dimensions: coherence, conciseness, coverage, accuracy, fluency, relevance, clarity, and engagement, accompanied by detailed explanations and scores for each. The overall quality is also assessed.

Your task is to **assess the explanations in these instances using the provided criteria below**. Please begin your evaluation now. Keep the document open at all times and consult it as necessary to guide your assessment of the specific evaluation criteria.

Instance Number

Copy the instance number, for example, (0-50)_1

____

►Please read the explanation for each dimension in '**Explanation**' carefully, and judge whether each explanation is unambiguous and easy to understand.

Clarity: Is the explanation unambiguous and easy to understand?

Yes: The explanation is concise, clear, and free of confusing terminology or expressions.

No: The explanation contains ambiguity or confusing terms that make it hard to understand.

	Yes	No
Coherence	○	○
Conciseness	○	○
Coverage	○	○
Accuracy	○	○
Fluency	○	○
Relevance	○	○
Clarity	○	○
Engagement	○	○
Overall Quality	○	○

► Please read the explanation for each dimension in '**Explanation**' carefully, and judge whether each explanation reflects and closely relates to its evaluation dimension.

Relevance: Does the explanation accurately reflect and closely relate to its evaluation dimension?

Yes: The explanation accurately reflects and closely relates to the evaluation dimension.

No: The explanation does not accurately reflect or closely relate to the evaluation dimension.

	Yes	No
Coherence	○	○
Conciseness	○	○
Coverage	○	○
Accuracy	○	○
Fluency	○	○
Relevance	○	○
Clarity	○	○
Engagement	○	○
Overall Quality	○	○

► Please read the explanation and score for each dimension in '**Explanation**' carefully, and judge whether each explanation reflects the assigned score.

Explanation and Score Alignment: Does the explanation appropriately reflect the assigned score, and can the user understand the reason for the assigned score through the explanation?

Yes: The explanation content clearly reflects the assigned score, and the user can understand the reason for the score.

No: The explanation content does not clearly reflect the assigned score, and the user cannot understand the reason for the score.

	Yes	No
Coherence	○	○
Conciseness	○	○
Coverage	○	○
Accuracy	○	○

Fluency	○	○
Relevance	○	○
Clarity	○	○
Engagement	○	○
Overall Quality	○	○

►Please read the '**Source**' and '**Explanation**' carefully, and judge whether each explanation matches the source.

Accuracy: Does the explanation match the source?

Yes: The explanation matches the source text, accurately reflecting the source data or facts, with no hallucinations.

No: The explanation does not match the source, containing inaccuracies or hallucinations.

	Yes	No
Coherence	○	○
Conciseness	○	○
Coverage	○	○
Accuracy	○	○
Fluency	○	○
Relevance	○	○
Clarity	○	○
Engagement	○	○
Overall Quality	○	○

►Please read the '**Explanation**' carefully and judge from an overall perspective whether the overall explanation aligns with the explanations for each dimension.

Overall Alignment: Does the overall explanation align with the explanations for each dimension?

Yes: The overall explanation is consistent with each dimension's explanation and avoids any contradictory meanings.

No: The overall explanation is inconsistent with the explanations for each dimension and contains contradictory meanings.

	Yes	No
Overall Alignment	○	○

►Please read the '**Explanation**' carefully and judge from an overall perspective whether the explanation clearly differentiates the current score segment from others.

Score Segment Differentiation: Does the explanation clearly differentiate the current score segment from others?

Yes: The explanation shows the unique characteristics of its score segment and distinguishes it from other segments, ensuring clear and transparent scoring.

No: The explanation does not clearly show the unique traits of its score segment and fails to distinguish it from other segments, which may cause confusion in scoring.

	Yes	No
Overall Alignment	○	○

Overall: Review all your previous evaluations and give an overall score for the explanation text in the current instance.

○**1:** Very poor quality, most aspects need significant improvement.

○**2:** Poor quality, several key aspects need improvement.

○**3:** Average quality, some aspects are good, but others need improvement.

○**4:** Good quality, most aspects meet standards with minor improvements needed.

○**5:** Excellent quality, all aspects are outstanding and consistent.