Survey

LLM-based NLG Evaluation: Current Status and Challenges

Mingqi Gao*, Xinyu Hu*, Xunjian Yin, Jie Ruan,
Xiao Pu, and Xiaojun Wan

Wangxuan Institute of Computer Technology, Peking University
gaomingqi@pku.edu.cn, huxinyu@pku.edu.cn, xjyin@pku.edu.cn,
ruanjie@stu.pku.edu.cn, puxiao@stu.pku.edu.cn, wanxiaojun@pku.edu.cn

Corrected Proof

Evaluating natural language generation (NLG) is a vital but challenging problem in natural language processing. Traditional evaluation metrics mainly capturing content (e.g., n-gram) overlap between system outputs and references are far from satisfactory, and large language models (LLMs) such as ChatGPT have demonstrated great potential in NLG evaluation in recent years. Various automatic evaluation methods based on LLMs have been proposed, including metrics derived from LLMs, prompting LLMs, fine-tuning LLMs, and human–LLM collaborative evaluation. In this survey, we first give a taxonomy of LLM-based NLG evaluation methods, and discuss their pros and cons, respectively. Lastly, we discuss several open problems in this area and point out future research directions.

1. Introduction

The evaluation of natural language generation (NLG) is an important but challenging issue. The lack of a single standard answer and the presence of multiple quality criteria make evaluating NLG more challenging than other NLP tasks. For example, in news summarization, a good summary should capture the key information from the source document, remain faithful to the source document, and be expressed in logically coherent and fluent language, but there is not a single "correct" way to achieve this. The inherent difficulty of NLG evaluation means that human evaluation is always needed and regarded as the gold standard. However, due to the high cost and time-consuming nature of human evaluation, automatic evaluation metrics remain indispensable and play a crucial role in model development. Over the past two decades, many automatic evaluation metrics such as BLEU (Papineni et al. 2002) and BARTScore (Yuan, Neubig, and Liu 2021) have been developed, but none have been fully satisfactory. Some studies

* Equal contribution.

Action Editor: Kevin Duh. Submission received: 26 November 2024; revised version received: 24 February 2025; accepted for publication: 14 March 2025.

https://doi.org/10.1162/coli_a_00561

© 2025 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

(Sai et al. 2021; He et al. 2023) have highlighted their deficiencies in robustness, such as insensitivities, biases, or even loopholes when evaluating challenging texts.

Recently, large language models (LLMs) have emerged and demonstrated unprecedented capacities in following instructions, understanding content, and generating text, which inspires researchers to utilize LLMs for NLG evaluation. Although this is a research direction that only emerged in 2023, the past year has seen an enormous amount of relevant work. While there have been surveys on automatic evaluation metrics or human evaluation practices in NLG evaluation (Celikyilmaz, Clark, and Gao 2020; Hämäläinen and Al-Najjar 2021; Zhou et al. 2022; Sai, Mohankumar, and Khapra 2023; Gehrmann, Clark, and Sellam 2023; Zhou, Ringeval, and Portet 2023), none of them addresses the LLM-based evaluation approach, and a comprehensive survey of this area is urgently needed.

The survey will mainly focus on research on LLM-based approaches for NLG evaluation, which involves language models with over one billion parameters. Moreover, it mainly considers the typical scope of NLG tasks where both input and output are natural languages including machine translation, text summarization, story generation, dialogue response generation, data-to-text, text simplification, paraphrase generation, grammatical error correction, and creative writing. Broader areas like evaluation of LLMs are not included (Zhuang et al. 2023; Chang et al. 2024) because this work focuses on LLMs used for evaluation, rather than the evaluation of LLMs' capabilities. We search the literature on Google Scholar with an end date of June 2024 with keywords. Because this is a new direction that emerged in 2023, a considerable number of arXiv preprints are included in addition to papers published in *ACL venues or other related venues. About 100 pieces of work will be included. To maintain focus, this article neither discusses datasets and benchmarks in NLG evaluation (Gehrmann et al. 2021; Kim et al. 2024a) nor analyzes evaluation metrics statistically (Ni'mah et al. 2023; Xiao et al. 2023).

As shown in Figure 1, we categorize related studies into four categories according to how LLMs are utilized for NLG evaluation:

- **LLM-derived Metrics** (§ 2): developing or deriving evaluation metrics from embeddings or generation probabilities of LLMs.
- **Prompting LLMs** (§ 3): directly inquiring of existing LLMs via specific prompts and processes designed for evaluation.
- **Fine-tuning LLMs** (§ 4): using labeled evaluation data to fine-tune existing LLMs and improving their NLG evaluation capabilities.
- **Human–LLM Collaborative Evaluation** (§ 5): leveraging distinctive strengths of both human evaluators and LLMs to achieve robust and nuanced evaluations through human–LLM collaboration.

LLMs have driven NLG evaluation toward a more human-centered direction, and the four categories we propose reflect this evolution: LLM-derived metrics are a continuation of traditional evaluation metrics and can only handle coarse-grained evaluation; prompting and fine-tuning methods enable users to express flexible evaluation requirements in natural language; collaborative evaluation takes it a step further, making it possible for humans and LLMs to leverage their strength respectively. We will review each type of evaluation method and discuss the pros and cons, respectively. Last but not

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

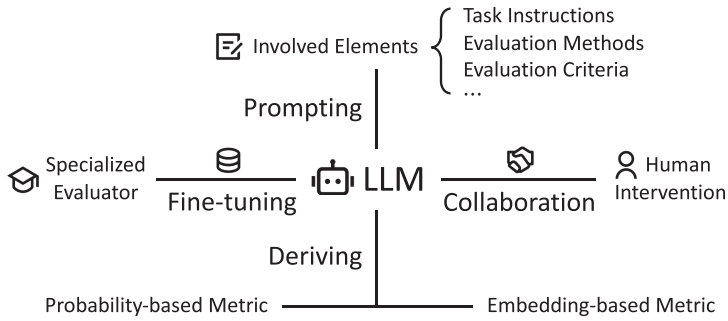

Figure 1
Schematic representation of our proposed four categories of LLM-based NLG evaluation.

least, we will provide our suggestions and conclusions, and discuss future directions in this area (§ 6).

It is worth stating that since LLM-based evaluation has shown unprecedented generality across NLG tasks, we do not summarize the literature for each task separately. Nevertheless, we will draw a list documenting all the approaches in this survey, indicating which NLG tasks each approach has been experimented on.

2. LLM-derived Metrics

LLM-derived metrics can be viewed as a continuation of early model-based NLG evaluation metrics such as BERTScore and BARTScore, replacing traditional pre-trained language models with stronger LLMs. Such work can be categorized into two main types: embedding-based metrics (Es et al. 2023) and probability-based metrics. The latter can be further divided into two categories based on different ways of using probabilities: directly converting the probabilities into scores (Fu et al. 2023a; Varshney et al. 2023) and leveraging the variation in probabilities under changed conditions (Jia et al. 2023; Xie et al. 2023).

2.1 Embedding-based Metrics

The embedding-based methods, like BERTScore, generally utilize representations of language models and thus compute the semantic similarity between the reference and the target text to evaluate, with different possible ways of implementation. However, unlike traditional embedding-based evaluation metrics, which require references, many LLM-based embedding evaluation metrics do not. This is because their application scenarios and implementation methods differ from those of traditional metrics. For example, when Es et al. (2023) evaluate the answer relevance of Retrieval Augmented Generation, given the original question q and the answer Y to be evaluated, they first prompt the LLM to generate n possible questions q_i for Y. Then, the relevance of Y is represented by the average similarity between q_i and q, denoted as $\sum_{i=1}^{n} \text{sim}(q_i, q)$, where $\text{sim}(q_i, q)$ refers to the cosine similarity of the embeddings of q_i and q. The embedding is generated by OpenAI *text-embedding-ada-002*, which can efficiently convert text into a 1536-dimensional vector, capturing semantic information and ensuring that similar texts are positioned close to each other in the vector space. Furthermore, Sheng et al. (2024) developed a more sophisticated method based on embeddings from the

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

open-source decoder-only LLM, utilizing Principal Component Analysis to adapt it for both pointwise scoring and pairwise comparison.

2.2 Probability-based Metrics

To better utilize the knowledge inherent in language models, probability-based methods like BARTScore formulate text generation evaluation as conditional probability comparison, positing that the better the quality of the target text, the higher the likelihood that models should be able to generate it. Recently, GPTScore (Fu et al. 2023a) has established tailored evaluation templates for each aspect to effectively guide multiple LLMs for NLG evaluation, including GPT3 (Brown et al. 2020), OPT (Zhang et al. 2022), and FLAN (Chung et al. 2022). The core idea of GPTScore is that a good generative language model is more likely to assign higher probabilities to high-quality text generated in response to a given instruction and context. Specifically, given a generative large language model θ, context information X (such as a source document), output text $Y = \{y_1, y_2, \ldots, y_m\}$ containing m tokens to be evaluated, and instruction I that specifies the requirement for the LLMs to generate text that can flexibly correspond to different evaluation aspects (e.g., *generating a factually consistent summary* for the aspect of consistency), GPTScore is defined as:

$$\text{GPTScore}(X, Y, I, \theta) = \sum_{i=1}^{m} \log P(y_i | y_{<i}, X, I, \theta)$$

Similarly, Murugadoss et al. (2024) score the task output Y to be evaluated by its perplexity under the corresponding large language model θ, given only the task context X. They believe this approach is unbiased by prompts, which transparently measures alignment with model training data. Furthermore, such methods have also been applied to the hallucination detection of the LLM-generated text (Varshney et al. 2023) with three different attempts for calculating the probability score.

On the other hand, some works leverage the variation in probabilities under changed conditions as the evaluation metric. FFLM (Jia et al. 2023) proposes to evaluate the faithfulness of the target text by calculating a combination of probability changes based on the intuition that the generation probability of a given text segment increases when more consistent information is provided, and vice versa. Similarly, DELTA-SCORE (Xie et al. 2023) measures the quality of different story aspects according to the likelihood difference between pre- and post-perturbation states with LLMs including GPT-3.5 (text-davinci-003) that provide logits. They believe that the sensitivity to specific perturbations indicates the quality of related aspects, and their experiments demonstrate the effectiveness of their approach.

2.3 Pros and Cons

Traditional NLG evaluation approaches always fall short due to their surface-form similarity when the target text and reference convey the same meaning but use different expressions. In contrast, LLM-derived metrics offer a remedy for the limitation and demonstrate stronger correlations with human judgments benefiting from the evolving modeling techniques. However, the flaws within LLMs can lead to some issues, as introduced in the following.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Robustness. Some research has investigated the robustness of LLM-derived metrics and found that they lack robustness in different attack scenarios. Specifically, He et al. (2023) develop a set of stress tests to assess the robustness of various model-based metrics on some common NLG tasks. They show a catalogue of the blind spots and potential errors identified that are not detected by different metrics.

Efficiency. Compared with traditional metrics, LLM-derived evaluation methods are more time-consuming and require more computational resources, especially when adopting LLMs with quite large parameter scales. To address this, Eddine et al. (2022) propose an approach to learning a lightweight version of LLM-derived metrics, and some fast LLM inference and serving tools like popular vLLM (Kwon et al. 2023) have been launched. vLLM improves memory utilization during inference through the PagedAttention algorithm, as well as the optimized memory management and batching strategies, thereby increasing LLMs' generation throughput. However, closed-source LLMs often do not make their parameters, representations, or logits public and available, thus making it impossible to apply LLM-derived methods to them.

Fairness. Sun et al. (2022) assess the social bias across various metrics for NLG evaluation on six sensitive attributes: race, gender, religion, physical appearance, age, and socioeconomic status. Their findings reveal that model-based metrics carry noticeably more social bias than traditional metrics. Relevant biases can be categorized into two types: intrinsic bias encoded within pre-trained language models and extrinsic bias injected during the computation of similarity. Therefore, current LLM-derived methods may have similar issues.

3. Prompting LLMs

The remarkable generation abilities of LLMs have expanded the possibilities for NLG evaluation. For a long time, human evaluation has been viewed as the gold standard for NLG evaluation. Recently, some studies claim that LLMs are on par with crowd-sourcing annotators in several tasks (Törnberg 2023; Gilardi, Alizadeh, and Kubli 2023; Ostyakova et al. 2023; Cegin, Simko, and Brusilovsky 2023). This raises questions about whether LLMs could replace human evaluators. Studies in this area often involve feeding LLMs with detailed prompts that include both instructions and the text to be evaluated, with LLMs producing the evaluation outcomes. An example of prompting LLMs is shown in Figure 2. From this example, we can see that such a prompt is quite similar to the guidelines given to human evaluators. The main differences between this prompting method for LLMs and LLM-derived metrics are twofold: (1) LLM-derived metrics generally do not involve highly human-like prompts that require the LLM to perform an evaluation. (2) The evaluation results from prompting LLMs are typically generated directly by the LLM, whereas LLM-derived metrics require further transformation from embeddings and probabilities. We will describe existing works according to the five elements that they mainly focus on:

- **Evaluation Methods:** The way the evaluation results of LLM evaluators are obtained, such as scoring and comparison.
- **Task Instructions:** How LLM evaluators should read or manipulate different parts to complete the annotation.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Prompt

Please act as an impartial and helpful evaluator for natural language generation (NLG). You will be given a news article and a summary written for it. Your task is to evaluate the quality of the summary strictly based on the given evaluation criterion. Please make sure you read and understand these instructions carefully.

Evaluation Criterion:
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. Annotators were also asked to penalize summaries that contained hallucinated facts.

News Article:
Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley on Sunday. 'Just been watching the game, did you miss the coach? #RubberDub #7minutes,' Merson ...

Summary:
Paul Merson was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley. Andros Townsend scored the Tottenham midfielder in the 89th minute ...

Evaluation Form:
Begin the evaluation by providing your analysis concisely and accurately, and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better).

Response

The summary contains several inconsistencies with the facts presented in the article. Firstly, it states that "Paul Merson was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley," which is incorrect as it was Andros Townsend who was brought on in that context, not ...
Rating: 2

Figure 2
An example of prompting LLMs to evaluate the aspect of consistency of the summary. There are role and interaction, task instructions, evaluation criteria, input content, and evaluation methods in the prompt, as well as the evaluation results, including the rating and explanation generated by LLMs.

Corrected Proof

- **Input Content:** The target text to be evaluated and other required content. Other required content including source documents, references, and external knowledge is provided as needed.
- **Evaluation Criteria:** The general definition of how good or bad the text to be evaluated is in a particular aspect of quality, e.g., fluency, faithfulness.
- **Role and Interaction:** The roles LLM evaluators play in the evaluation and the interactions between them.

3.1 Evaluation Methods

Diverse evaluation methods have been used in prompting LLMs to obtain their preferences for the text to be evaluated: scoring, comparison, ranking, Boolean QA, and error analysis (Table 1).

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Table 1
Representative studies on prompting LLMs for NLG evaluation.

Related Work	Evaluation Method	NLG Task
Chiang and Lee (2023a)	Scoring	Story Generation
Wang et al. (2023a)	Scoring	Summarization, Data-to-text & Story Generation
Kocmi and Federmann (2023b)	Scoring	Translation
Lin and Chen (2023)	Scoring	Dialogue
Mendonça et al. (2023)	Scoring	Dialogue
Naismith, Mulcaire, and Burstein (2023)	Scoring	Discourse Generation
Liusie, Manakul, and Gales (2023)	Scoring & Comparison	Summarization, Dialogue & Data-to-text
Wang et al. (2023d)	Comparison	Personalized Text Generation
Ji et al. (2023)	Ranking	Open-end Text Generation
Liu et al. (2023c)	Scoring, Ranking & Comparison	Summarization
Wang, Funakoshi, and Okumura (2023)	Boolean QA	Question Generation
Manakul, Liusie, and Gales (2023)	Boolean QA	Fact Verification
Guan et al. (2023)	Boolean QA	Fact Verification
Es et al. (2023)	Boolean QA	Retrieval Augmented Generation
Kocmi and Federmann (2023a)	Error Analysis	Translation
Lu et al. (2023)	Error Analysis	Translation
Chang et al. (2023)	Error Analysis	Summarization

Scoring. Scoring is the most commonly used evaluation method in human evaluation for NLG (van der Lee et al. 2021), and it is naturally applied to LLM-based evaluation. Chiang and Lee (2023a) have conducted relevant studies early, using a Likert scale from 1 to 5 to evaluate story generation and adversarial attacks with InstructGPT (Ouyang et al. 2022) and ChatGPT,[1] showing that the evaluation results of LLMs are consistent with expert human evaluators. Kocmi and Federmann (2023b) find that GPT-3.5 and GPT-4 achieve the state-of-the-art accuracy of evaluating translation quality compared to human labels with a rating scale from 1 to 5 or 0 to 100, outperforming all the results from the metric shard task of WMT22 (Freitag et al. 2022). Furthermore, Wang et al. (2023a) experiment on five datasets across summarization, story generation, and data-to-text; ChatGPT with similar rating scales achieves the state-of-the-art or comparative correlations with human judgments in most settings, compared with prior metrics. Similar conclusions are also observed in open-domain dialogue response generation (Lin and Chen 2023). Besides English, Mendonça et al. (2023) show that ChatGPT with simple rating prompts is a strong evaluator for multilingual dialogue evaluation, surpassing prior metrics based on encoders.

Comparison. Different from absolute scoring, comparison refers to choosing the better of the two. Luo, Xie, and Ananiadou (2023) and Gao et al. (2023) use ChatGPT to compare the factual consistency of two summaries. AuPEL (Wang et al. 2023d) evaluates

1 https://openai.com/blog/chatgpt/.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

personalized text generation from three aspects in the form of comparison with the PaLM 2 family (Anil et al. 2023). According to Liusie, Manakul, and Gales (2023), pairwise comparison is better than scoring when medium-sized LLMs (e.g., FlanT5 [Chung et al. 2022] and LLaMa2 [Touvron et al. 2023]) are adopted as evaluators.

Ranking. Ranking can be viewed as an extended form of comparison. In comparison, only two examples are involved at a time, whereas in ranking, the order of more than two examples needs to be decided at once. Ji et al. (2023) use ChatGPT to rank five model-generated responses across several use cases at once, indicating the ranking preferences of ChatGPT align with those of humans to some degree. Similarly, GPTRank is a method to rank summaries in a list-wise manner (Liu et al. 2023c). Moreover, Liu et al. (2023b) compare different evaluation methods in LLM-based summarization including scoring, comparison, and ranking, showing that the optimal evaluation method for each backbone LLM may vary.

Boolean QA. Boolean QA requires LLMs to answer "Yes" or "No" to a question. It is adopted more in scenarios where human annotations are binary, such as grammaticality (Hu et al. 2023), faithfulness of summaries and statements (Luo, Xie, and Ananiadou 2023; Gao et al. 2023; Es et al. 2023; Hu et al. 2023), factuality of generated text (Fu et al. 2023b; Guan et al. 2023; Manakul, Liusie, and Gales 2023), and answerability of generated questions (Wang, Funakoshi, and Okumura 2023).

Error Analysis. Error Analysis refers to the evaluation of a text by looking for errors that occur in the text according to a set of predefined error categories. Multidimensional Quality Metrics (MQM) (Jain et al. 2023) is an error analysis framework prevalent in machine translation evaluation. According to MQM, Lu et al. (2023) and Kocmi and Federmann (2023a) use ChatGPT or GPT-4 to automatically detect translation quality error spans. BooookScore (Chang et al. 2023), an LLM-based evaluation metric, assesses the coherence of book summaries by identifying eight types of errors.

3.2 Task Instructions

In human evaluation, task instruction usually comes in the form of a task description or evaluation steps. They can also exist at the same time. The task description states the annotation in a more general way, and the evaluation steps, which can be considered as Chain-of-Thought, explicitly describe what to do at each step. In the context of prompting LLMs for NLG evaluation, we discuss three broad categories of influences: various templates of prompts (Leiter et al. 2023; Kim et al. 2023a; Kotonya et al. 2023; He, Zhang, and Roth 2023), in-context examples (Jain et al. 2023; Kotonya et al. 2023; Hasanbeig et al. 2023), and whether LLMs are required to provide analyses or explanations (Chiang and Lee 2023b; Naismith, Mulcaire, and Burstein 2023).

Form and Requirements. Several studies from an Eval4NLP 2023 shared task (Leiter et al. 2023) have explored task instructions in various settings. For example, Kim et al. (2023a) conduct experiments on different templates and lengths of task descriptions and evaluation steps, finding that providing clear and straightforward instructions, akin to those explained to humans, is more effective. Kotonya et al. (2023) generate task instructions with LLMs or improve existing task instructions with LLMs. Moreover, Leiter and Eger (2024) conduct a larger-scale prompt exploration for the evaluation of machine translation and summarization based on the Eval4NLP 2023 shared task.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Somewhat differently, He, Zhang, and Roth (2023) evaluate generative reasoning using LLMs by asking them first to generate their own answers, and then conduct a quantitative analysis of the text to be evaluated. Additionally, explicit evaluation requirements and output formats are typically included in the instructions, and the evaluation results are extracted using regular expression matching. Early LLMs may sometimes provide unrecognizable evaluation results or refuse to conduct evaluation due to their limited instruction-following capabilities (Gao et al. 2023). This issue can be mitigated through multiple sampling or setting random outputs, and it basically does not exist in the currently more advanced and powerful LLMs.

Analysis and Explanations. LLMs are able to include analysis or explanation in their evaluations, which is a key point that distinguishes them from previous automatic evaluation metrics. Early explorations into prompting LLMs for NLG evaluation mostly do not examine the impact of whether LLMs are required to analyze and explain the evaluation results. However, Chiang and Lee (2023b) explore different types of evaluation instructions in summarization evaluation and dialogue evaluation, finding that explicitly asking large models to provide analysis or explanation achieve higher correlation with human judgments. Besides, the quality of the analysis and explanation generated by LLMs itself requires additional manual evaluation (Leiter et al. 2023). Naismith, Mulcaire, and Burstein (2023) compare the explanations written by humans and generated by GPT-4 and conduct a simple corpus analysis on the generated explanations, finding that GPT-4 has strong potential to produce ratings that are comparable to human ratings on discourse coherence, accompanied by clear rationales.

In-context Examples. Similarly to other fields, sometimes demonstrations are needed when prompting LLMs for NLG evaluation. Specifically, Jain et al. (2023) use only in-context examples as task instructions, relying on LLMs to evaluate the quality of summaries. In scenarios where task descriptions or evaluation steps are included, Kotonya et al. (2023) compare the performance of LLMs as evaluators in both zero-shot and one-shot settings, finding that one-shot prompting does not bring improvements. Moreover, Hasanbeig et al. (2023) improve the performance of LLM evaluators by updating the in-context examples iteratively.

3.3 Input Content

The types of input content mainly depend on the evaluation criteria and are relatively fixed. For most task-specific evaluation criteria, such as the faithfulness of a summary (Luo, Xie, and Ananiadou 2023; Gao et al. 2023), the source document is needed in addition to the target text to be evaluated. For task-independent criteria, such as fluency (Hu et al. 2023; Chiang and Lee 2023b), only the text to be evaluated needs to be provided, though many studies also provide the source document (Wang et al. 2023a; Liusie, Manakul, and Gales 2023). Other types of input content can be provided as required by the specific task. Kocmi and Federmann (2023b) use two different settings when evaluating machine translation: providing references and not providing references and find that GPT-4 without references can also outperform all existing reference-based metrics. Guan et al. (2023) provide relevant facts and context when evaluating whether a text conforms to the facts. Exceptionally, Shu et al. (2023) add the output of other automatic evaluation metrics to the input of the LLM.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

3.4 Evaluation Criteria

The evaluation targeting specific aspects is used in numerous studies of human evaluation for NLG, such as text summarization, story generation, dialogue, and text simplification. Evaluation criteria, that is, the definitions of aspects, are key in this context. Most evaluation criteria in LLM-based evaluation are directly derived from human evaluation. However, a few studies have attempted to let LLMs generate or improve evaluation criteria. Liu et al. (2023e) use a few human-rated examples as seeds to let LLMs draft some candidate evaluation criteria, and then further filter them based on the performance of LLMs using these criteria on a validation set, to obtain the final evaluation criteria. Kim et al. (2023c) designed an LLM-based interactive evaluation system, which involves using LLMs to review the evaluation criteria provided by users, including eliminating ambiguities in criteria, merging criteria with overlapping meanings, and decomposing overly broad criteria. Additionally, Ye et al. (2023a) propose a hierarchical aspect classification system with 12 subcategories, demonstrating that under the proposed fine-grained aspect definitions, human evaluation and LLM-based evaluation are highly correlated. Additionally, the chain-of-aspects approach improves LLMs' ability to evaluate on a specific aspect by having LLMs score on some related aspects before generating the final score (Gong and Mao 2023).

3.5 Role and Interaction

We include in this section the evaluation strategies that either use the same LLMs in different ways or involve different LLMs (Bai et al. 2023; Li, Patel, and Du 2023; Cohen et al. 2023). The former can be further divided into chain-style (Yuan et al. 2024; Fu et al. 2023b; Hu et al. 2023) and network-style interactions (Chan et al. 2023; Zhang et al. 2023b; Saha et al. 2023; Wu et al. 2023).

Chain-style Interaction. Inspired by human evaluators, Yuan et al. (2024) have LLMs score a batch of examples to be evaluated each time. Specifically, the evaluation process is divided into three stages: analysis, ranking, and scoring. Similar to QA-based evaluation metrics (Durmus, He, and Diab 2020), Fu et al. (2023b) assess the faithfulness of summaries in two stages: treating LLMs as question generators to generate a question from the summary; then having LLMs answer the question using the source document. Differently, when Hu et al. (2023) use GPT-4 to evaluate the faithfulness of summaries, they first ask GPT-4 to extract event units from the summary, then verify whether these event units meet the requirements, and finally judge whether the event units are faithful to the source document.

Network-style Interaction. Unlike chain-style interactions, network-style interactions involve the dispersion and aggregation of information. In network-style interactions, LLMs on the same layer play similar roles. ChatEval (Chan et al. 2023) is a framework for evaluating content through debates among multiple LLMs, with three communication strategies designed among the three types of LLMs: One-By-One, Simultaneous-Talk, and Simultaneous-Talk-with-Summarizer. Zhang et al. (2023b) find that under certain conditions, widening and deepening the network of LLMs can better align its evaluation with human judgments. Saha et al. (2023) propose a Branch-Solve-Merge strategy, assigning LLMs the roles of decomposing problems, solving them, and aggregating answers, thereby improving the accuracy and reliability of evaluations. Wu et al. (2023) assume that different people such as politicians and the general public have

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

different concerns about the quality of news summaries, use LLMs to play different roles in evaluation accordingly, and aggregate the results finally.

Different LLMs. Different from having the same LLM play different roles, some research has used different LLMs (such as GPT-4 and Claude) in their studies. The use of a single LLM as evaluator may introduce bias, resulting in unfair evaluation results. In light of this, Bai et al. (2023) design a decentralized peer-examination method, using different LLMs as evaluators and then aggregating the results. Further, Li, Patel, and Du (2023) let different LLMs serve as evaluators in pairwise comparisons and then have them go through a round of discussion to reach the final result. Additionally, Cohen et al. (2023) evaluate the factuality of texts through the interaction of two LLMs, where the LLM that generated the text acts as the examinee and the other LLM as the examiner.

3.6 Pros and Cons

The benefits of prompting LLMs for NLG evaluation are exciting. First, for the first time, people can express evaluation criteria and evaluation methods in natural language within the prompts given to LLMs, providing great flexibility. Where previously people needed to design specific evaluation metrics for different NLG tasks or even different aspects of a single task, now they only need to modify the prompts for LLMs. Secondly, surprisingly, LLMs have the ability to generate explanations while assessing texts, making this approach somewhat interpretable. Furthermore, in many NLG tasks, prompting LLMs for evaluation has achieved state-of-the-art correlations with human judgments.

However, as many studies have pointed out, this type of approach still has many limitations. Wang et al. (2023b) note that when using ChatGPT and GPT-4 for pairwise comparisons, the order of the two texts can affect the evaluation results, which is known as position bias. To alleviate this issue, Li et al. (2023c) propose a strategy of splitting, aligning, and then merging the two texts to be evaluated into the prompt. Also, LLM evaluators tend to favor longer, more verbose responses (Zheng et al. 2023) and responses generated by themselves (Liu et al. 2023a). Wu and Aji (2023) show that compared with answers that are too short or grammatically incorrect, answers with factual errors are considered better by LLMs. Liu et al. (2023d) demonstrate through adversarial meta-evaluation that LLMs without references are not suitable for evaluating dialogue responses in closed-ended scenarios: They tend to score highly on responses that conflict with the facts in the dialogue history. Zhang et al. (2023a) also present the robustness issues of LLMs in dialogue evaluation through adversarial perturbations. Shen et al. (2023) indicate that LLM evaluators have a lower correlation with human assessments when scoring high-quality summaries. In addition, Hada et al. (2023) state that LLM-based evaluators have a bias towards high scores, especially in non-Latin languages like Chinese and Japanese. Bavaresco et al. (2024) find that the performance of LLM-based evaluators exhibits significant variance depending on the dataset, evaluation criteria, and whether the evaluated texts are human-generated. Beyond these shortcomings of performance, both ChatGPT and GPT-4 are proprietary models, and their opacity could lead to irreproducible evaluation results.

4. Fine-tuning LLMs

As mentioned above, despite the exciting performance of prompting LLMs like ChatGPT and GPT-4 for NLG evaluation, several shortcomings in practice are inevitable, such as high costs, possibly irreproducible results, and potential biases in LLMs. In

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

response, recent research has shifted towards fine-tuning smaller, open-source LLMs specifically for evaluation purposes, aiming to achieve performance close to GPT-4 in NLG evaluation. Representative works of this type include PandaLM (Wang et al. 2023e), Prometheus (Kim et al. 2023b), Prometheus 2 (Kim et al. 2024b), Shepherd (Wang et al. 2023c), TIGERScore (Jiang et al. 2023), INSTRUCTSCORE (Xu et al. 2023), Auto-J (Li et al. 2023a), CritiqueLLM (Ke et al. 2023), JudgeLM (Zhu, Wang, and Wang 2023), Themis (Hu et al. 2024), CompassJudger-1 (Cao et al. 2024), and Self-Taught (Wang et al. 2024). Their main ideas are similar, involving the elaborate construction of high-quality evaluation data, followed by fine-tuning open-source foundation LLMs with specific methods. Nevertheless, there are certain discrepancies in the designs across different works, such as the usage of references and evaluation criteria. We have summarized the key different components of these methods in Table 2 and Table 3 for comparison, and we will elaborate on these in the following sections.

Corrected Proof

Table 2
Comparison of the different key components among the representative methods of fine-tuning LLMs (Part 1).

Method	Data Construction			Foundation LLM
	Instruction Source	Annotator	Scale	
PandaLM	Alpaca 52K	GPT-3.5	300K	LLaMA 7B
Prometheus	GPT-4 Construction	GPT-4	100K	LLaMA-2-Chat 7B & 13B
Prometheus 2	FEEDBACK COLLECTION	GPT-4	200K	Mistral-7B Mixtral-8×7B
Shepherd	Community Critique Data & 9 NLP Tasks Data	Human	1317	LLaMA 7B
TIGERScore	23 Distinctive Text Generation Datasets	GPT-4	48K	LLaMA-2 7B & 13B
INSTRUCTSCORE	GPT-4 Construction	GPT-4	40K	LLaMA 7B
AUTO-J	Real-world User Queries from Preference Datasets	GPT-4	4,396	LLaMA-2-Chat 13B
CritiqueLLM	AlignBench & ChatGPT Augmentation	GPT-4	9,332	ChatGLM-2 6B, 12B & 66B
JudgeLM	GPT4All-LAION, ShareGPT Alpaca-GPT4 & Dolly-15K	GPT-4	100K	Vicuna 7B, 13B & 33B
Themis	NLG-Eval with 58 NLG Evaluation Datasets	Human & GPT-4	67K	LLaMa-3-8B
Self-Taught	Screened WildChat	LLaMa-3-70B	20K	LLaMa-3-70B
CompassJudger-1	Sampling from existing datasets	Mixture	900K	Qwen-2.5 1.5B, 7B, 14B & 32B

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Table 3
Comparison of the different key components among the representative methods of fine-tuning LLMs (Part 2).

Method	Evaluation Method			Reference Required
	Result Mode	Details	Specific Criteria	
PandaLM	Comparison	Reason & Reference	Unified	No
Prometheus	Scoring	Reason	Explicit	Yes
Prometheus 2	Scoring & Comparison	Reason	Explicit	Yes
Shepherd	Overall Judgment	Error Identifying & Refinement	Unified	No
TIGERScore	MQM	Error Analysis	Implicit	No
INSTRUCTSCORE	MQM	Error Analysis	Implicit	Yes
AUTO-J	Scoring & Comparison	Reason	Implicit	No
CritiqueLLM	Scoring	Reason	Unified	Flexible
JudgeLM	Scoring & Comparison	Reason	Unified	Flexible
Themis	Scoring	Reason	Explicit	No
Self-Taught	Comparison	Reason	Unified	No
CompassJudger-1	Scoring & Comparison	Reason	Explicit	No

Corrected Proof

4.1 Data Construction

Diverse data with high-quality annotations is crucial for the fine-tuning of evaluation models, which mainly involves task scenarios, inputs, target texts to evaluate, and evaluation results. Early NLG evaluation research primarily focused on conventional NLG tasks, such as summarization and dialogue generation. Thus, the task scenarios, inputs, and target texts refer to the corresponding NLP task, source inputs of the task, and outputs generated by specialized systems based on task requirements, respectively. Mainstream datasets for these tasks predominantly use human annotators to provide evaluation results, which are often considered reliable.

With the recent rise of LLMs, the spectrum of NLG tasks has been broadened to scenarios of instruction and response that are more aligned with human needs. Traditional tasks like summarization with corresponding source inputs can be viewed as kinds of instructions and requirements. Meanwhile, responses generated by various general LLMs generally serve as the target texts now and require more flexible evaluation so that the performance of different LLMs can be compared, promoting further developments. Therefore, to keep pace with the current advancement of modeling techniques, most evaluation methods have adopted a similar instruction-response scenario.

The primary differences in these works actually lie in the construction of instructions, with the purpose of improving either diversity or reliability for the better generalization ability of the fine-tuned model. PandaLM and JudgeLM entirely sample

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

from common instruction datasets, such as Alpaca 52K, while CritiqueLLM adopts small-scale sampling followed by ChatGPT augmentation. In contrast, Prometheus and INSTRUCTSCORE rely on GPT-4 to generate all the instructions based on seed data, whereas Auto-J and Shepherd use real-world data. Moreover, because large-scale human annotation is impractical, most work utilizes GPT-4 as the powerful annotator, except for PandaLM and Shepherd, which use GPT-3.5 and human annotation on small-scale community data, respectively. Specifically, Themis focuses on NLG tasks and combines existing human evaluations with additional evaluations from GPT-4, selecting more consistent training data. Self-Taught uses the evaluation results from the model to fine-tune itself (LLaMa-3-70B), considering that it already possesses strong capabilities. During the construction, these studies basically all design detailed prompts or guidance and apply heuristic filtering strategies and post-processing methods to mitigate noise. Overall, despite the possible higher quality of human annotation, the corresponding drawback is the difficulty in constructing large-scale datasets, which in turn may hinder adequate model training, while using LLMs for construction is the opposite situation.

4.2 Evaluation Method

As with prompting LLMs, the evaluation methods adopted in these studies are highly diversified, involving different evaluation criteria, result modes, and usages of the reference. Given that current instruction-response scenarios encompass different types of tasks, it is unsuitable to specify unified evaluation criteria as in traditional NLG tasks. However, some work still does it this way, while some other methods let LLM annotators adaptively and implicitly reflect the required criteria in their evaluations, like PandaLM, TIGERScore, and AUTO-J. In particular, AUTO-J has meticulously crafted 332 evaluation criteria, matched to different tasks. Furthermore, Prometheus and Themis explicitly incorporate evaluation criteria into the evaluation instructions, and CompassJudger-1 can work either with or without evaluation criteria, enabling flexible evaluation based on various customized criteria.

More details about the evaluation methods are shown in Table 3. All the methods require models to provide detailed information, such as reasons for their evaluation results. And the MQM mode can achieve more informative error analysis, offering stronger interpretability. Moreover, some methods do not necessarily require references and then have greater value in practice. A more optimal method is to concurrently support both reference-based and reference-free evaluations as JudgeLM and CritiqueLLM.

4.3 Fine-tuning Implementation

The fine-tuning process is implemented by different studies on their selected open-source foundation LLMs, like LLaMA, and respective constructed data, with some targeted settings. Specifically, Prometheus maintains balanced data distributions during fine-tuning, including the length and label. JudgeLM eliminates potential biases by randomly swapping sample pairs to be compared and randomly removing references. INSTRUCTSCORE utilizes GPT-4 to provide error annotations for the intermediate outputs of the fine-tuned model for further supervised reinforcement. And based on some preliminary experiments and manual analysis, TIGERScore determines appropriate ratios of different types of data during fine-tuning, which are claimed to be crucial by them. Moreover, CritiqueLLM implements separately, with and without references, and explores the effects of data and model scale. Themis uses additional rating-guided preference optimization after the fine-tuning process. Specifically, Self-Taught utilizes

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

the evaluation results of the fine-tuned model itself for self-iterative optimization, leading to surprising improvements. Compared with the vanilla fine-tuning setting, these methods have improved the efficiency of model training and the robustness of evaluations.

4.4 Pros and Cons

The shortcomings of prompting LLMs for NLG evaluation can be significantly alleviated by the customized construction of training data and specifically fine-tuned LLMs. For instance, most models in Table 2 have less than 14B parameters, facilitating low-cost inference in practice and good reproducibility, with performance comparable to GPT4. And specific measures can be adopted to prevent certain biases found in GPT4 during different stages, such as randomly changing the order of training pairs for position bias. Furthermore, this type of approach allows for continuous iteration and improvement of the model to address potential deficiencies or emerging issues discovered in future applications.

However, some inherent biases associated with GPT4 may still persist, like self-biases, as the data construction of most methods uses GPT4 for critical evaluation annotation. On the other hand, many studies have chosen open-source foundation LLMs spanning three generations of the LLaMa series. With the recent rapid updates and improvements of open-source LLMs, it is intuitive that utilizing a more powerful foundation LLM should lead to better evaluation performance of the fine-tuned model. However, this means repetitive fine-tuning processes and computational expenses from scratch since directly migrating existing fine-tuned models to the new foundation LLM is challenging.

Additionally, although many existing methods aspire to more flexible and comprehensive evaluation through fine-tuning, demanding excessive evaluation settings may ultimately lead to poor performance or failure in model training, as AUTO-J and CritiqueLLM were found to have difficulties with criteria and references, respectively. However, there are some disagreements here since Prometheus, JudgeLM, and CompassJudger-1 show different results, indicating such a seemingly straightforward fine-tuning process is actually quite complex. Moreover, considering the different evaluation settings in existing work, conducting a horizontal comparison among them is challenging. These issues require further exploration in future research.

5. Human–LLM Collaborative Evaluation

Human evaluation remains the gold standard for NLG due to its ability to capture nuanced aspects of quality. However, it is expensive, time-consuming, and prone to subjective biases (van der Lee et al. 2021; Deriu et al. 2021; Li et al. 2023b). While LLMs offer a promising avenue for automated evaluation, their reliability and correlation with human judgment are still areas of active development (Li et al. 2023c; Liu et al. 2023d). Human–LLM collaborative evaluation seeks to leverage the strengths of both: the nuanced judgment of humans and the scalability and efficiency of LLMs. This section explores emerging paradigms in this collaborative space, focusing on how humans and LLMs can work together to improve the accuracy, efficiency, and trustworthiness of NLG evaluation. This includes collaborative approaches like: traditional evaluation tasks such as scoring and explaining (Zhang, Ren, and de Rijke 2021; Li et al. 2023b), general evaluation tasks such as testing and debugging (Ribeiro and Lundberg 2022), auditing NLG models to ensure fairness (Rastogi et al. 2023), aligning LLM-assisted

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

evaluation of LLM outputs with human preferences (Shankar et al. 2024), and addressing the intricate challenge of scalable oversight (Amodei et al. 2016; Saunders et al. 2022).

5.1 Human-guided LLM Evaluation

Some work (Zhang, Ren, and de Rijke 2021; Li et al. 2023b) focuses on approaches where LLMs perform the primary evaluation task, but with significant guidance and oversight from humans. This guidance can take several forms, from designing detailed evaluation criteria to refining LLM outputs.

One common method is called checklist-based evaluation. A key challenge in open-ended NLG tasks is the lack of consistent evaluation criteria. Li et al. (2023b) address this with COEVAL, a collaborative pipeline where humans design a task-specific checklist. LLMs then use this checklist to generate initial evaluations and explanations, drawing on developments in explainable NLP (Yin and Neubig 2022; Jung et al. 2022; Ribeiro and Lundberg 2022; Ye et al. 2023b). Humans then scrutinize these LLM-generated evaluations, refining scores and explanations. This approach leverages the LLM's ability to process large amounts of text while retaining human oversight to ensure accuracy and reduce outliers. Notably, human review still leads to revisions in approximately 20% of LLM scores, highlighting the importance of human judgment. Furthermore, InteractEval (Chu, Kim, and Yi 2025) combines human and LLM-generated attributes using Think Aloud methods to create questions and produce final prediction scores. Think Aloud methods mean that human experts verbalize their thoughts and LLMs articulate their knowledge to generate text attribute insights using sample texts and evaluation rubrics, which highlights the necessity of effectively combining humans and LLMs in an automated checklist-based text evaluation.

Collaborative assignment is also useful for human-guided LLM evaluation. Zhang, Ren, and de Rijke (2021) propose HMCEval, a framework that frames dialogue evaluation as a sample assignment problem. This approach aims to optimize the allocation of evaluation tasks between humans and machines to maximize accuracy while minimizing human effort. HMCEval achieves high accuracy (99%) with significantly reduced human involvement (half the effort). Additionally, EvalAssist (Ashktorab et al. 2024) can help practitioners refine evaluation criteria using both direct and pairwise assessment strategies. Ashktorab et al. (2024) also examine how users refine their criteria and identify key differences approaches examined, and how users refine their criteria and identified key differences between the two evaluation approaches.

5.2 LLM-assisted Human Evaluation

Some studies (Ribeiro and Lundberg 2022; Rastogi et al. 2023; Pozdniakov et al. 2024) explore scenarios where humans remain the primary evaluators, but LLMs provide assistance to improve efficiency, identify flaws, or audit for biases.

Ribeiro and Lundberg (2022) introduce AdaTest, a system where LLMs generate unit tests to identify bugs in a target NLG model. Human feedback guides the LLM, significantly increasing the effectiveness of bug detection (5x–10x improvement). This demonstrates the power of LLMs in generating diverse test cases, guided by human intuition. In the task of evaluating machine translation systems, Zouhar, Kocmi, and Sachan (2025) assist annotators by pre-filling error annotations with recall-oriented automatic quality estimation, which achieves the effect of reducing the time per span

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

annotation by half while maintaining the same annotation quality level and further cutting the annotation budget by almost 25%.

Addressing biases and irresponsible behavior in LLMs is crucial (Blodgett et al. 2020; Jones and Steinhardt 2022). AdaTest++ (Rastogi et al. 2023), drawing on human-AI collaboration research, facilitates collaborative auditing. Humans leverage their strengths in schematization and hypothesis testing, while LLMs assist in identifying a wide range of failure modes. This collaborative approach uncovered both previously known and under-reported issues.

Evaluating LLMs on complex tasks can be challenging even for humans (Chen et al. 2021; Nakano et al. 2021; Li et al. 2022; Menick et al. 2022). The concept of scalable oversight (Amodei et al. 2016) suggests using AI to assist in evaluation. Saunders et al. (2022) explore using LLM-generated critiques to help humans identify flaws in model outputs, demonstrating that this form of assistance improves human performance. What's more, Pozdniakov et al. (2024) focus on designing conversational user interfaces, which helps educators use LLMs to evaluate assignments of students.

5.3 Pros and Cons

Human–LLM collaborative evaluation offers a compelling balance between the accuracy of human judgment and the efficiency of automated methods. Key advantages include: (1) Efficiency and Cost-Effectiveness: LLMs can significantly reduce the time and resources required for evaluation. (2) Complementary Strengths: Humans excel at nuanced judgment and critical thinking, while LLMs excel at processing large amounts of data and generating diverse outputs. (3) Improved Accuracy: Combining human and LLM strengths can lead to more accurate and reliable evaluations than either approach alone.

However, challenges remain: (1) Prompt Sensitivity: LLM evaluation results can be sensitive to the phrasing of prompts, requiring careful prompt engineering (Li et al. 2023b; Rastogi et al. 2023). (2) Confidence Calibration: LLMs' ability to accurately assess their own confidence is still limited, making it difficult to know when to trust their judgments. (3) Need for Human Oversight: Although reduced, human supervision is still necessary, limiting the potential for full automation. (4) Explainability: Ensuring the collaborative process is transparent and understandable can be challenging.

6. Conclusions and Future Trends

6.1 Comparison with Traditional Evaluation Metrics

Traditional evaluation metrics are criticized for their poor correlation with human judgments (Stent, Marge, and Singhai 2005), uninterpretable evaluation results (Zhang, Vogel, and Waibel 2004), and inability to adapt to specific evaluation criteria (Wiseman, Shieber, and Rush 2017), which are being greatly mitigated by LLM-based evaluation. However, the higher cost, the requirements for computing resources, and the issues of reproducibility may be the downside.

6.2 Comparison Between Different Types of LLM-based NLG Evaluation

We compare different types of LLM-based evaluation according to flexibility and reproducibility due to the difficulty of comparing the effectiveness of different types of methods in various scenarios.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Flexibility. Human–LLM Collaborative Evaluation > Prompting LLMs > Fine-tuning LLMs > LLM-derived Metrics. Human–LLM Collaborative Evaluation involves human annotators, which provides the highest flexibility. LLM-derived Metrics are typically designed to evaluate specific aspects, such as text similarity, and do not fully allow evaluation criteria to be expressed in natural language, making them the least flexible. When comparing Prompting LLMs and Fine-tuning LLMs, the former, which uses proprietary models, generally performs better at following instructions compared with smaller open-source models.

Reproducibility. LLM-derived Metrics ≈ Prompting LLMs > Fine-tuning LLMs > Human–LLM Collaborative Evaluation. Human–LLM Collaborative Evaluation requires human annotators, and the recruitment and training of these annotators pose greater challenges to reproducibility. LLM-derived Metrics and Prompting LLMs do not modify the existing models, and therefore have better reproducibility than Fine-tuning LLMs. However, they may still become non-reproducible if proprietary models are deprecated.

Performance. Human–LLM Collaborative Evaluation > Fine-tuning LLMs ≈ Prompting LLMs > LLM-derived Metrics. We compare the performance of different LLM-based evaluation approaches on the most commonly used NLG evaluation benchmark on summarization, SummEval (Fabbri et al. 2021), as shown in Table 4. When using the same LLMs, LLM-derived metrics perform worse than directly prompting LLMs for evaluation, and the latter is more convenient. Moreover, among methods of fine-tuning LLMs, only models focused on NLG evaluation scenarios, such as Themis, outperform prompting-based methods, including those using GPT-4. Other studies either use relatively outdated foundation LLMs or lack training on specific evaluation aspects like those in SummEval, leading to relatively weaker performance. Furthermore, Human–LLM Collaborative Evaluation enhances the LLM evaluation by incorporating checklists elaborated with human expert insights and LLM knowledge, resulting in the strongest performance.

Cost. LLM-derived Metrics ≈ Prompting open-source LLMs < Fine-tuning LLMs ≈ Prompting proprietary LLMs < Human–LLM Collaborative Evaluation. When using the same open-source LLM, the inference costs of LLM-derived metrics, prompting LLM, and fine-tuning LLM methods are the same, while fine-tuning LLM incurs additional training costs. When prompting proprietary LLMs, the cost is high and mainly concentrated in API calls during evaluation, making it difficult to directly compare with the training cost required for fine-tuning LLM. Moreover, human–LLM collaborative evaluation requires the involvement of human experts for each task, making it the most expensive approach.

6.3 Future Directions

Unified Benchmarks for LLM-based NLG Evaluation Approaches. As mentioned above, each of the studies that fine-tuned LLMs to construct specialized evaluation models uses different settings and data during testing, making them incomparable. In the research on prompting LLMs for NLG evaluation, there are some publicly available human judgments on the same NLG task, such as SummEval for summarization. However, the existing human judgments have many problems. First, most of the existing data only involve one type of NLG task and a single human evaluation method (e.g.,

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Table 4
Performance of different types of LLM-based NLG evaluation approaches on SummEval, where some results are from Fu et al. (2023a), Hu et al. (2024), and Chu, Kim, and Yi (2025).

Method	Parameter	Coherence	Consistency	Fluency	Relevance	Overall
Traditional Metrics						
BERTScore	355M	0.285	0.151	0.186	0.302	0.231
BARTScore	400M	0.474	0.266	0.258	0.318	0.329
LLM-derived Metrics						
GPTScore (FT5)	11B	0.456	0.438	0.424	0.343	0.415
GPTScore (OPT)	66B	0.359	0.453	0.380	0.337	0.382
GPTScore (GPT-3)	175B	0.434	0.449	0.403	0.381	0.417
GPTScore (Phi-4)	14B	0.319	0.436	0.386	0.154	0.324
GPTScore (LLaMa-3.1)	70B	0.415	0.478	0.437	0.288	0.405
GPTScore (Qwen-2.5)	72B	0.447	0.486	0.437	0.376	0.436
Prompting LLMs						
G-Eval (GPT-3.5)	–	0.440	0.386	0.424	0.385	0.409
G-Eval (GPT-4)	–	0.582	0.507	0.455	0.548	0.523
Phi-4	14B	0.479	0.454	0.421	0.452	0.451
LLaMa-3.1	70B	0.510	0.387	0.317	0.494	0.427
Qwen-2.5	72B	0.515	0.509	0.435	0.528	0.497
Fine-tuning LLMs						
INSTRUCTSCORE	7B	0.328	0.232	0.260	0.211	0.258
Prometheus 2	7B	0.403	0.318	0.269	0.356	0.336
Themis	8B	0.566	0.600	0.571	0.474	0.553
TIGERScore	13B	0.381	0.427	0.363	0.366	0.384
CompassJudger-1	32B	0.494	0.424	0.318	0.410	0.411
Human–LLM Collaborative Evaluation						
InteractEval (GPT-3.5 1st)	–	0.583	0.630	0.734	0.614	0.640
InteractEval (GPT-3.5 2nd)	–	0.590	0.614	0.726	0.623	0.638
InteractEval (GPT-4 1st)	–	0.649	0.799	0.783	0.626	0.714
InteractEval (GPT-4 2nd)	–	0.660	0.781	0.816	0.642	0.725

Corrected Proof

scoring), making it difficult to evaluate LLMs' performance on different tasks, as well as using different evaluation methods on the same task. Second, many of the texts in these human judgments are generated by outdated models (such as Pointer Network) and do not include texts generated by more advanced LLMs. Lastly, many human evaluation datasets are too small in scale. There is an urgent need for large-scale, high-quality human evaluation data covering various NLG tasks and evaluation methods as a benchmark.

NLG Evaluation for Low-resource Languages and New Task Scenarios. Almost all existing research focuses on English data. However, it is doubtful whether LLMs have similar levels of NLG evaluation capability for texts in other languages, especially low-resource languages. As Zhang et al. (2023a) point out, we should be more cautious about using LLMs to evaluate texts in non-Latin languages. We believe that the lack of evaluation capability of LLM-based evaluators on low-resource languages may be due to the

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

insufficient presence of these languages in the pretraining corpus. Therefore, further fine-tuning on certain low-resource languages may be a potential strategy to address this issue, and Hada et al. (2024) have already shown promising preliminary results. Additionally, existing research mainly focuses on more traditional NLG tasks such as translation, summarization, and dialogue. However, there are many new scenarios in reality with different requirements and evaluation criteria. For example, using LLMs to automatically evaluate scientific reviews could be valuable in identifying and flagging content that is unfaithful or unclear, alerting reviewers to potential issues. Research on low-resource languages and new task scenarios will provide a more comprehensive understanding of LLMs' evaluation capabilities.

Diverse Forms of Human–LLM Collaborative NLG Evaluation. According to the literature reviewed above, there is little research on collaborative evaluation between humans and LLMs. Neither humans nor LLMs are perfect, and each has its strengths. Because the ultimate goal of NLG research is to evaluate text quality more accurately and efficiently, we believe that collaboration between humans and LLMs can achieve better results than pure human evaluation or automatic evaluation. In the collaboration between humans and LLMs, technologies in the field of human–computer interaction may bring new implementation methods to the collaboration. In addition, what roles humans and LLMs should play in the evaluation and how they can better complement each other are still worth researching.

Acknowledgments

This work was supported by Beijing Science and Technology Program (Z231100007423011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *CoRR*, abs/1606.06565.

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 technical report. *Computing Research Repository*, arXiv:2305.10403.

Ashktorab, Zahra, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and LLM judgments: Insights from EvalAssist on task-specific evaluations and AI-assisted assessment strategy preferences. *arXiv preprint arXiv: 2410.00873*.

Bai, Yushi, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. *CoRR*, abs/2306.04181.

Bavaresco, Anna, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of ACL*, pages 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*.

Cao, Maosong, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. CompassJudger-1: All-in-one judge model helps model

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

evaluation and evolution. *arXiv preprint arXiv:2410.16256*.

Cegin, Jan, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905. https://doi.org/10.18653/v1/2023.emnlp-main.117

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

Chan, Chi-Min, 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. *CoRR*, abs/2308.07201.

Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *CoRR*, abs/2310.00785.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):39:1–39:45. https://doi.org/10.1145/3641289

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Chiang, David Cheng-Han and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of ACL (1)*, pages 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

Chiang, David Cheng-Han and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Proceedings of EMNLP (Findings)*, pages 8928–8942. https://doi.org/10.18653/v1/2023.findings-emnlp.599

Chu, SeongYeub, JongWoo Kim, and MunYong Yi. 2025. Think together and work better: Combining humans' and LLMs' think-aloud outcomes for effective text evaluation. In *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Article No.: 1089*, pages 1–23. https://doi.org/10.1145/3706598.3713181

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Cohen, Roi, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. In *Proceedings of EMNLP*, pages 12621–12640. https://doi.org/10.18653/v1/2023.emnlp-main.778

Deriu, Jan, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Reviews*, 54(1):755–810. https://doi.org/10.1007/s10462-020-09866-x, PubMed: 33505103

Durmus, Esin, He He, and Mona T. Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of ACL*, pages 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

Eddine, Moussa Kamal, Guokan Shang, Antoine J.-P. Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In *Proceedings of ACL (1)*, pages 1305–1318. https://doi.org/10.18653/v1/2022.acl-long.93

Es, Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *CoRR*, abs/2309.15217.

Fabbri, Alexander R., Wojciech Kryscinski, 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. https://doi.org/10.1162/tacl_a_00373

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George F. Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU - neural metrics are better and more robust. In *Proceedings of WMT*, pages 46–68.

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

Uncorrected Proof

Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023a. GPTScore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Fu, Xue-Yong, Md. Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan T. N. 2023b. Are large language models reliable judges? A study on the factuality evaluation capabilities of LLMs. *CoRR*, abs/2311.00681.

Gao, Mingqi, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with ChatGPT. *CoRR*, abs/2304.02554.

Gehrmann, Sebastian, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672. https://doi.org/10.18653/v1/2021.gem-1.10

Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166. https://doi.org/10.1613/jair.1.13715

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056. https://doi.org/10.1073/pnas.2305016120, PubMed: 37463210

Gong, Peiyuan and Jiaxin Mao. 2023. CoAScore: Chain-of-aspects prompting for NLG evaluation. *CoRR*, abs/2312.10355.

Guan, Jian, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *CoRR*, abs/2310.14564. https://doi.org/10.18653/v1/2024.naacl-long.62

Hada, Rishav, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. METAL: Towards multilingual meta-evaluation. In *Proceedings of NAACL-HLT (Findings)*, pages 2280–2298. https://doi.org/10.18653/v1/2024.findings-naacl.148

Hada, Rishav, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *CoRR*, abs/2309.07462.

Hämäläinen, Mika and Khalid Al-Najjar. 2021. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. *CoRR*, abs/2108.00308. https://doi.org/10.18653/v1/2021.gem-1.9

Hasanbeig, Hosein, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. 2023. ALLURE: Auditing and improving LLM-based evaluation of text using iterative in-context-learning. *CoRR*, abs/2309.13701.

He, Hangfeng, Hongming Zhang, and Dan Roth. 2023. SocREval: Large language models with the Socratic method for reference-free reasoning evaluation. *CoRR*, abs/2310.00074.

He, Tianxing, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097. https://doi.org/10.18653/v1/2023.acl-long.674

Hu, Xinyu, 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. https://doi.org/10.18653/v1/2024.emnlp-main.891

Hu, Yebowen, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. DecipherPref: Analyzing influential factors in human preference judgments via GPT-4. In *Proceedings of EMNLP*, pages 8344–8357. https://doi.org/10.18653/v1/2023.emnlp-main.519

Jain, Sameer, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Proceedings of ACL (Findings)*, pages 8487–8495. https://doi.org/10.18653/v1/2023.findings-acl.537

Ji, Yunjie, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

ChatGPT's ability to rank content: A preliminary study on consistency with human preferences. *CoRR*, abs/2303.07610.

Jia, Qi, Siyu Ren, Yizhu Liu, and Kenny Q. Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. In *Proceedings of EMNLP*, pages 11017–11031. https://doi.org/10.18653/v1/2023.emnlp-main.679

Jiang, Dongfu, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. TIGERScore: Towards building explainable metric for all text generation tasks. *CoRR*, abs/2310.00752.

Jones, Erik and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Proceedings of NeurIPS*.

Jung, Jaehun, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic Prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/2022.emnlp-main.82

Ke, Pei, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. CritiqueLLM: Scaling LLM-as-critic for effective and explainable evaluation of large language model generation. *CoRR*, abs/2311.18702.

Kim, Joonghoon, Saeran Park, Kiyoon Jeong, Sangmin Lee, Seung Hun Han, Jiyoon Lee, and Pilsung Kang. 2023a. Which is better? Exploring prompting strategy for LLM-based metrics. *CoRR*, abs/2311.03754. https://doi.org/10.18653/v1/2023.eval4nlp-1.14

Kim, Seungone, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023b. Prometheus: Inducing fine-grained evaluation capability in language models. *CoRR*, abs/2310.08491.

Kim, Seungone, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Choi, Sheikh Shafayat, Jinheon Baek, et al. 2024a. The BiGGen Bench: A principled benchmark for fine-grained evaluation of language models with language models. *CoRR*, abs/2406.05761.

Kim, Seungone, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353. https://doi.org/10.18653/v1/2024.emnlp-main.248

Kim, Tae Soo, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023c. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. *CoRR*, abs/2309.13633. https://doi.org/10.1145/3613904.3642216

Kocmi, Tom and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of WMT*, pages 768–775. https://doi.org/10.18653/v1/2023.wmt-1.64

Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of EAMT*, pages 193–203.

Kotonya, Neema, Saran Krishnasamy, Joel R. Tetreault, and Alejandro Jaimes. 2023. Little giants: Exploring the potential of small LLMs as evaluation metrics in summarization in the Eval4NLP 2023 shared task. *CoRR*, abs/2311.00686. https://doi.org/10.18653/v1/2023.eval4nlp-1.17

Kwon, Woosuk, 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 SOSP*, pages 611–626. https://doi.org/10.1145/3600006.3613165

Leiter, Christoph and Steffen Eger. 2024. PrExMe! Large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In *Proceedings of EMNLP*, pages 11481–11506. https://doi.org/10.18653/v1/2024.emnlp-main.641

Leiter, Christoph, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. *CoRR*, abs/2310.19792. https://doi.org/10.18653/v1/2023.eval4nlp-1.10

Li, Junlong, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu.

Uncorrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

2023a. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470.

Li, Qintong, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023b. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *CoRR*, abs/2310.19740.

Li, Ruosen, Teerth Patel, and Xinya Du. 2023. PRD: Peer rank and discussion improve large language model based evaluations. *CoRR*, abs/2307.02762.

Li, Yujia, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097. https://doi.org/10.1126/science.abq1158, PubMed: 36480631

Li, Zongjie, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023c. Split and merge: Aligning position biases in large language model based evaluators. *CoRR*, abs/2310.01432.

Lin, Yen-Ting and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *NLP4ConvAI 2023*. arXiv:2305.13711. https://doi.org/10.18653/v1/2023.nlp4convai-1.5

Liu, Yang, 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 EMNLP*, pages 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

Liu, Yixin, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023b. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *CoRR*, abs/2311.09184. https://doi.org/10.18653/v1/2024.findings-naacl.280

Liu, Yixin, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023c. On learning to summarize with large language models as references. *CoRR*, abs/2305.14239.

Liu, Yongkang, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023d. Evaluate what you can't evaluate: Unassessable generated responses quality. *CoRR*, abs/2305.14658.

Liu, Yuxuan, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023e. Calibrating LLM-based evaluator. *CoRR*, abs/2309.13308.

Liusie, Adian, Potsawee Manakul, and Mark J. F. Gales. 2023. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. *Computing Research Repository*, arxiv:2307.07889.

Lu, Qingyu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT. *CoRR*, abs/2303.13809. https://doi.org/10.20944/preprints202303.0255.v1

Luo, Zheheng, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a factual inconsistency evaluator for abstractive text summarization. *CoRR*, abs/2303.15621.

Manakul, Potsawee, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, pages 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

Mendonça, John, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. In *Proceedings of the Eleventh Dialog System Technology Challenge*. *arXiv.2308.16797*.

Menick, Jacob, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.

Murugadoss, Bhuvanashree, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. *arXiv preprint arXiv:2408.08781*. https://doi.org/10.1609/aaai.v39i18.34157

Naismith, Ben, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of BEA@ACL*, pages 394–403.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

https://doi.org/10.18653/v1/2023.bea-1.32
Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.
Ni'mah, Iftitahu, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of ACL (1)*, pages 1240–1266. https://doi.org/10.18653/v1/2023.acl-long.69
Ostyakova, Lidiia, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254. https://doi.org/10.18653/v1/2023.sigdial-1.23
Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*. arXiv:2406.16897.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. https://doi.org/10.3115/1073083.1073135
Pozdniakov, Stanislav, Jonathan Brazil, Solmaz Abdi, Aneesha Bakharia, Shazia Sadiq, Dragan Gašević, Paul Denny, and Hassan Khosravi. 2024. Large language models meet user interfaces: The case of provisioning feedback. *Computers and Education: Artificial Intelligence*, 7:100289. https://doi.org/10.1016/j.caeai.2024.100289
Rastogi, Charvi, Marco Túlio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-AI collaboration in auditing LLMs with LLMs. In *Proceedings of AIES*, pages 913–926. https://doi.org/10.1145/3600211.3604712
Ribeiro, Marco Túlio and Scott M. Lundberg. 2022. Adaptive testing and debugging of NLP models. In *Proceedings of ACL (1)*, pages 3253–3267. https://doi.org/10.18653/v1/2022.acl-long.230
Saha, Swarnadeep, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *CoRR*, abs/2310.15123.
Sai, Ananya B., Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation checklists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234. https://doi.org/10.18653/v1/2021.emnlp-main.575
Sai, Ananya B., Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2):26:1–26:39. https://doi.org/10.1145/3485766
Saunders, William, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *CoRR*, abs/2206.05802.
Shankar, Shreya, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, pages 1–14. https://doi.org/10.1145/3654777.3676450
Shen, Chenhui, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Proceedings of EMNLP (Findings)*, pages 4215–4233. https://doi.org/10.18653/v1/2023.findings-emnlp.278
Sheng, Shuqian, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, and Chenghu Zhou. 2024. RepEval: Effective text evaluation with LLM representation. In *EMNLP*, pages 7019–7033. https://doi.org/10.18653/v1/2024.emnlp-main.398
Shu, Lei, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, and Lei Meng. 2023. Fusion-Eval: Integrating

Uncorrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

evaluators with LLMs. *CoRR*, abs/2311.09204.

Stent, Amanda, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *CICLing*, volume 3406 of *Lecture Notes in Computer Science*, pages 341–351. https://doi.org/10.1007/978-3-540-30586-6_38

Sun, Tianxiang, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *EMNLP*. https://doi.org/10.18653/v1/2022.emnlp-main.245

Törnberg, Petter. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *CoRR*, abs/2304.06588.

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

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151. https://doi.org/10.1016/j.csl.2020.101151

Varshney, Neeraj, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. *CoRR*, abs/2307.03987.

Wang, Jiaan, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? A preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, arXiv:2303.04048. https://doi.org/10.18653/v1/2023.newsum-1.1

Wang, Peiyi, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

Wang, Tianlu, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-Taught evaluators. *arXiv preprint arXiv:2408.02666*.

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

Wang, Yaqing, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023d. Automated evaluation of personalized text generation using large language models. *CoRR*, abs/2310.11593.

Wang, Yidong, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023e. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087.

Wang, Zifan, Kotaro Funakoshi, and Manabu Okumura. 2023. Automatic answerability evaluation for question generation. *CoRR*, abs/2309.12546.

Wiseman, Sam, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *EMNLP*, pages 2253–2263. https://doi.org/10.18653/v1/D17-1239

Wu, Minghao and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025.

Wu, Ning, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In *Proceedings of NLPCC (1)*, pages 695–707. https://doi.org/10.1007/978-3-031-44693-1_54

Xiao, Ziang, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *EMNLP*, pages 10967–10982. https://doi.org/10.18653/v1/2023.emnlp-main.676

Xie, Zhuohan, Miao Li, Trevor Cohn, and Jey Han Lau. 2023. DeltaScore: Fine-grained story evaluation with perturbations. In *Proceedings of EMNLP (Findings)*, pages 5317–5331. https://doi.org/10.18653/v1/2023.findings-emnlp.353

Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag,

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025

William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of EMNLP*, pages 5967–5994. https://doi.org/10.18653/v1/2023.emnlp-main.365

Ye, Seonghyeon, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023a. FLASK: Fine-grained language model evaluation based on alignment skill sets. *CoRR*, abs/2307.10928.

Ye, Xi, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Proceedings of ACL (Findings)*, pages 4469–4484. https://doi.org/10.18653/v1/2023.findings-acl.273

Yin, Kayo and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of EMNLP*, pages 184–198. https://doi.org/10.18653/v1/2022.emnlp-main.14

Yuan, Peiwen, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, and Kan Li. 2024. BatchEval: Towards human-like text evaluation. *CoRR*, abs/2401.00437. https://doi.org/10.18653/v1/2024.acl-long.846

Yuan, Weizhe, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Proceedings of NeurIPS*, pages 27263–27277.

Zhang, Chen, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2023a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *CoRR*, abs/2312.15407.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Zhang, Xinghua, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023b. Wider and deeper LLM networks are fairer LLM evaluators. *CoRR*, abs/2308.01862.

Zhang, Yangjun, Pengjie Ren, and Maarten de Rijke. 2021. A human-machine collaborative framework for evaluating malevolence in dialogues. In *Proceedings of ACL/IJCNLP (1)*, pages 5612–5623. https://doi.org/10.18653/v1/2021.acl-long.436

Zhang, Ying, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC*.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *CoRR*, abs/2306.05685.

Zhou, Kaitlyn, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of NAACL-HLT*, pages 314–324. https://doi.org/10.18653/v1/2022.naacl-main.24

Zhou, Yongxin, Fabien Ringeval, and François Portet. 2023. A survey of evaluation methods of generated medical textual reports. In *ClinicalNLP@ACL*, pages 447–459. https://doi.org/10.18653/v1/2023.clinicalnlp-1.48

Zhu, Lianghui, Xinggang Wang, and Xinlong Wang. 2023. JudgeLM: Fine-tuned large language models are scalable judges. *CoRR*, abs/2310.17631.

Zhuang, Ziyu, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. *CoRR*, abs/2308.07902.

Zouhar, Vilém, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. *arXiv preprint arXiv: 2406.12419*.

Corrected Proof

Downloaded from http://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00561/2513169/coli_a_00561.pdf by guest on 21 June 2025