Leveraging Large Language Models for NLG Evaluation: Advances and Challenges

Anonymous ACL submission

Abstract

In the rapidly evolving domain of Natural Language Generation (NLG) evaluation, introducing Large Language Models (LLMs) has opened new avenues for assessing generated content quality, e.g., coherence, creativity, and context relevance. This paper aims to provide a thorough overview of leveraging LLMs for NLG evaluation, a burgeoning area that lacks a systematic analysis. We propose a coherent taxonomy for organizing existing LLM-based evaluation metrics, offering a structured framework to understand and compare these methods. Our detailed exploration includes critically assessing various LLM-based methodologies, as well as comparing their strengths and limitations in evaluating NLG outputs. By discussing unresolved challenges, including bias, robustness, domain-specificity, and unified evaluation, this paper seeks to offer insights to researchers and advocate for fairer and more advanced NLG evaluation techniques.

1 Introduction

Natural Language Generation (NLG) stands at the forefront of AI-driven communication, with advancements in LLMs (Ouyang et al., 2022; OpenAI, 2023). These models demonstrate exceptional text generation proficiency, highlighting the need for robust evaluation. Traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) mainly focus on surface differences, inadequately capturing semantic quality (Freitag et al., 2020). Embedding-based methods (Liu et al., 2016; Sellam et al., 2020; Zhang et al., 2020) suffer from limited scope (Freitag et al., 2021a), low alignment with human judgment (Liu et al., 2023c), and lack of interpretability (Xu et al., 2023). These underscores the urgent need for more effective and flexible evaluation techniques in NLG.

The emergent capabilities of LLMs, such as Chain-of-Thought (CoT) (Wei et al., 2022) and

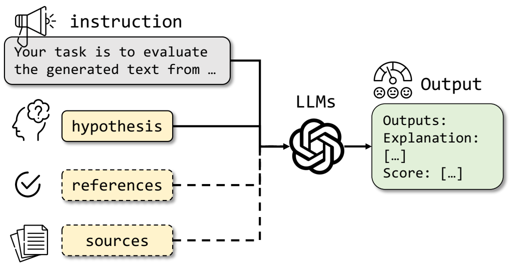

Figure 1: Illustration of LLMs for NLG evaluation. The dashed line means that the references and sources are optional based on the scenarios.

better alignment with human preferences (Ouyang et al., 2022), position them as effective tools for NLG evaluation, offering sophisticated and human-aligned assessments beyond traditional methods (Liu et al., 2023c; Kocmi and Federmann, 2023; Fu et al., 2023). For example, LLMs can provide explanations for scores (Xu et al., 2023), and reinforcement learning with human feedback (RLHF) further aligns LLMs with human judgment (Ouyang et al., 2022; Zheng et al., 2023). As illustrated in Figure 1, the key strategy involves prompting LLMs to evaluate texts from various aspects, with or without references or sources.

Given the burgeoning body of work on LLMs for NLG evaluation, there is an urgent need for a synthesized summary to navigate the advanced and varied works in this area. This paper aims to offer a comprehensive overview with a coherent taxonomy for categorizing existing research. We carefully outline the existing methods, and engage in an analytical discussion on their unique features and limitations. Additionally, we navigate through the unresolved challenges and open questions, highlighting potential directions for future research.

Organization of this paper: We start by setting up a formal framework for NLG evaluation and introduce a taxonomy to organize relevant research (§2). We then provide detailed discussions on these works (§3). Furthermore, we provide a thorough comparison of LLM-based evaluators

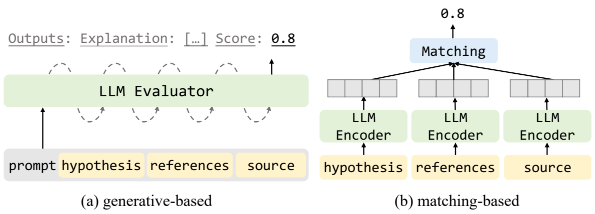

Figure 2: Illustration of NLG evaluation functions: (a) generative-based and (b) matching-based methods.

with traditional evaluators in terms of performance, efficiency and qualitative qualitative analysis (Section 4). Acknowledging the field's swift progress, we highlight and explore potential open problems for future investigation (§5).

2 Formalization and Taxonomy

Formalization The goal of LLM-based NLG evaluation is to evaluate model-generated text across various dimensions, such as fluency, consistency, etc. To maintain generality, the LLM-based NLG evaluation framework for task t is defined as:

$$E = f_t(h, s, r), \quad (1)$$

where f represents the evaluation function executed by LLMs, h is the hypothesis text (i.e. the candidate generation) under evaluation, s stands for the source of the generation, which could include source text or supporting documents, and r denotes the ground truth references.

Taxonomy We classify works along three primary dimensions according to Eq. 1: *evaluation task*, *evaluation references* and *evaluation function*.

Evaluation Task t: NLG encompasses a diverse range of tasks, such as Machine Translation (MT) (Farhad et al., 2021; Bapna et al., 2019), Text Summarization (TS) (Liu and Liu, 2021; Zhang et al., 2023a), Dialogue Generation (DG) (Wang et al., 2022; Kann et al., 2022), Story Generation (SG) (Yang et al., 2022; Fan et al., 2018), etc, each with its unique evaluation requirements. The specific nature of each task determines the target evaluation aspects and scenarios.

Evaluation References r: Evaluation scenarios are categorized into *reference-based* and *reference-free* based on the availability of references. In *reference-based* evaluation, the generated text h is assessed against ground truth references r, focusing on accuracy, relevance, coherence, and similarity to the references. Conversely, the *reference-free* method evaluates h without external references, concentrating on its intrinsic qualities or alignment with the source context s. It is suitable for evaluating fluency, originality, context relevance, etc.

Evaluation Function f: The evaluation function can be categorized as either *matching-based* or *generative-based*, depending on how LLMs are utilized. As depicted in Figure 2, *matching-based* methods assess the semantic similarity between the hypothesis and the reference or source text. These methods include token-level matching in representation space (Zhang et al., 2020; Zhao et al., 2019) or in discrete string space (Lin, 2004; Papineni et al., 2002), and sequence-level evaluation (Sellam et al., 2020; Rei et al., 2020; Peyrard et al., 2017). On the other hand, *generative-based* methods use LLMs to produce textual evaluations directly, tapping into the generative powers of LLMs to design instructions for assessing text quality.

Scope of this paper *Matching-based* methods are typically based on encoder-based language models to calculate a score-specific aspect of evaluation. Most of them often face challenges such as limited interpretability, lower correlation with human judgments, and restricted aspects (Xu et al., 2023; Fu et al., 2023). In contrast, generative LLMs tend to have huge size with powerful emergent abilities. These abilities include improved interpretability through CoT, higher customization via instruction-following capabilities, and better alignment with human evaluations through RLHF (Xu et al., 2023; Zheng et al., 2023). *Given the abundance of recent surveys primarily focusing on matching-based evaluation methods (refer to (Celikyilmaz et al., 2020; Sai et al., 2022; Goyal et al., 2023) for comprehensive summaries), our paper is dedicated to exploring more burgeoning generative-based methods (Figure 3).*

3 Generative Evaluation

Amidst the rapid evolution of LLMs, a burgeoning body of research has directed its focus toward leveraging LLMs as NLG evaluators, which we refer to as generative evaluation. This category, broadly classified into ***prompt-based*** and ***tuning-based evaluation***, hinges on whether the parameters of LLM evaluators require fine-tuning. The former typically involves directly prompting LLMs to assess generated text through prompt engineering, while the latter relies on open-source LLMs that are specifically calibrated for NLG evaluation. Both approaches cater to diverse evaluation protocols for measuring the quality of generated texts.

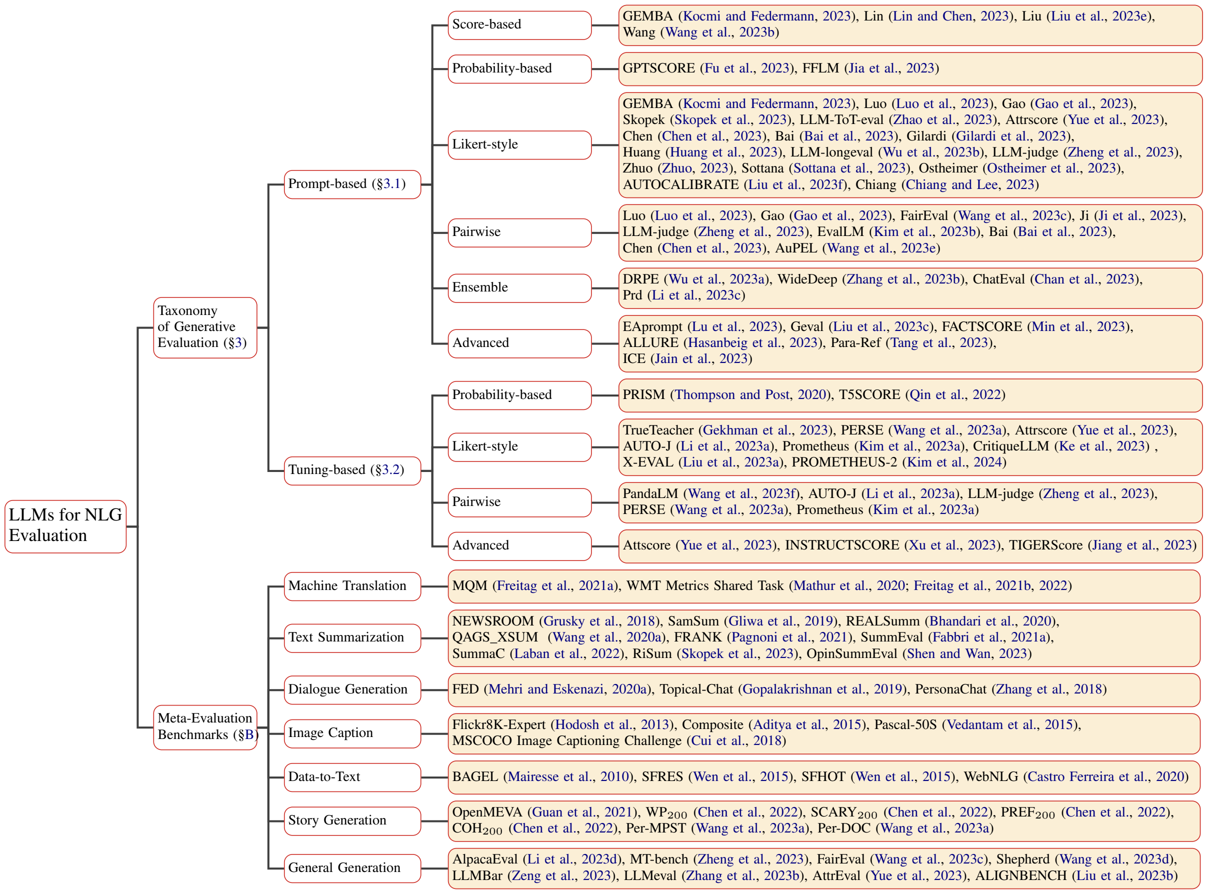

Figure 3: Taxonomy of research in NLG evaluation with large language models.

Some endeavors deploy LLM evaluators to yield continuous scalar quality scores for generated texts—termed as ❶ *score-based evaluation*. Others calculate the generation probability of generated texts based on prompts, sources or reference texts (optional) as the evaluation metric, denoted as ❷ *probability-based evaluation*. Certain works assess the quality of generated text by assigning it to a specific quality level using quality labels or likert scales—referred to as ❸ *likert-style evaluation*. Meanwhile, ❹ *pairwise comparison methods* involve using LLM evaluators to compare quality of pairs of generated texts. Additionally, ❺ *ensemble evaluation methods* utilize multiple LLM evaluators, orchestrating communication among evaluators to yield final evaluation results. Finally, some recent studies explore ❻ *advanced evaluation methods* that consider fine-grained criteria or combine the capabilities of chain-of-thought or in-context leaning. Table 1 provides a comprehensive overview of current representative prompt-based and tuning-based evaluation methods. This section delves into a detailed exploration of these two overarching categories, each accompanied by their respective evaluation protocols.

3.1 Prompt-based Evaluation

Prompt-based text evaluation stands at the forefront of advancements in NLG, particularly leveraging the capabilities of LLMs. In this method, the evaluation process is intricately woven into the crafting of prompts – specialized cues designed to guide LLMs in assessing the quality of generated text. More recently, the Eval4NLP workshop held a shared task on prompting LLMs as explainable metrics (Leiter et al., 2023). By harnessing the prowess of LLMs, prompt-based evaluation not only provides a comprehensive understanding of NLG system performance but also offers a nuanced approach to extracting valuable insights.

Score Evaluation. An intuitive and widely employed protocol for text evaluation involves prompting LLM evaluators to generate a continuous quality score. A concrete example is illustrated in the first row of Table 4 in the appendix. Pioneering this method, GEMBA (Kocmi and Federmann, 2023) proposed to utilize LLM evaluators to assign translation quality scores ranging from 0 to 100 with or without reference. Building on this foundation, Lin and Chen (2023) and Liu et al. (2023e) extended score evaluation methods to open-domain and closed-end conversations evaluation. Further-

Metric	MT	TS	DG	IC	D2T	SG	GE	REF	LLMs	Protocol	Aspects
									Prompt-based Evaluation		
BARTScore (Yuan et al., 2021)	✓	✓	*	*	✓	*	*	✓	BART	Prob	CON/COH/REL/FLU/ INF/COV/ADE
GPTScore (Fu et al., 2023)	✓	✓	✓		✓	*	*		GPT3	Prob	CON/COH/REL/FLU/COV/ACC MQM/INF/FAC/INT/ENG/NAT
G-EVAL (Liu et al., 2023c)	*	✓	✓		*	*	*		ChatGPT/GPT-4	Advanced	CON/COH/REL/FLU /NAT/ENG/GRO
ICE (Jain et al., 2023)	*	✓	*		*	*	*		GPT-3	Score	CON/COH/REL/FLU
GEMBA (Kocmi and Federmann, 2023)	✓	*	*		*	*	*		ChatGPT	Score/Likert	NONE
LLM_eval (Chiang and Lee, 2023)	*	*	*		*	✓	*		ChatGPT	Likert	GRAM/COH/REL/LIK
FairEval (Wang et al., 2023c)	*	*	*		*	*	✓		ChatGPT/GPT-4	Pairwise	NONE
AuPEL (Wang et al., 2023e)	*	*	*		*	*	✓		PaLM-2	Pairwise	PER/QUA/REL
DRPE (Wu et al., 2023a)	*	✓	*	*	*	*	*	✓	GPT-3	Ensemble	CON/COH/REL/FLU/INT/USE
ChatEval (Chan et al., 2023)	*	*	✓		*	*	✓		ChatGPT/GPT-4	Ensemble	NAT/COH/ENG/GRO
WideDeep (Zhang et al., 2023b)	*	*	*		*	*	✓		ChatGPT	Ensemble	COH/REL/HARM/ACC
PRD (Li et al., 2023c)	*	*	*		*	*	✓		GPT-4/GPT-3.5 Vicuna/Claude/Bard	Ensemble	INF/COH
FACTSCORE (Min et al., 2023)		*					✓		ChatGPT	Advanced	FAC
EAprompt (Lu et al., 2023)	✓	*	*		*	*	*		ChatGPT/text-davinci-003	Advanced	NONE
AUTOCALIBRATE (Liu et al., 2023f)	*	✓	*		*	*	*		GPT-4	Likert	CON/COH/REL/FLU/INF/NAT
ALLURE (Hasanbeig et al., 2023)	*	✓	*		*	*	✓		GPT-4	Advanced	CON/COH/FLU/REL
									Tuning-based Evaluation		
PRISM (Thompson and Post, 2020)	✓	*	*	*	*	*	*	✓	Transformer	Prob	NONE
T5Score (Qin et al., 2022)	✓	✓	*	*	*	*	*	✓	T5	Prob	NONE
TrueTeacher (Gekhman et al., 2023)	*	✓	*		*	*	*		T5	Likert	CON
X-EVAL (Liu et al., 2023a)	*	✓	✓		✓	*	*		FLAN-T5-large	Likert	DEP/LIK/UND/FLE/INF/INQ INT/SPE/COR/SEM/COH/ENG NAT/GRO/CON/REL/FLU
AUTO-J (Li et al., 2023a)	*	*	*		*	*	*		LLaMA	Likert/Pairwise	ACC/CLA/FEA/CRE/THO STR/LAY/COM/INF
PERSE (Wang et al., 2023a)	*	*	*	*	*	✓	*	✓	LLaMA	Likert/Pairwise	INT/ADA/SUR/CHA/END
PandaLM (Wang et al., 2023f)	*	*	*		*	*	✓		LLaMA	Pairwise	CLA/COM/FOR/ADH
Attscore (Yue et al., 2023)	*	*	*		*	*	✓		Roberta/T5/GPT2 LLaMA/Vicuna	Advanced	CON
TIGERScore (Jiang et al., 2023)	✓	✓	*		✓	✓	✓		LLaMA	Advanced	COH/INF/ACC/COM
INSTRUCTSCORE (Xu et al., 2023)	✓	*	*	*	*	*	*	✓	LLaMA	Advanced	NONE
Prometheus (Kim et al., 2023a)	*	*	*		*	*	✓		LLaMA-2	Likert/Pairwise	NONE
Prometheus-2 (Kim et al., 2023a)	*	*	*		*	*	✓		Mistral 7B	Likert/Pairwise	NONE
CritiqueLLM (Ke et al., 2023)	*	*	*		*	*	✓		ChatGLM	Likert	NONE

Table 1: Automatic metrics proposed (✓) and adopted (*) for various NLG tasks. **REF** indicate the method is source context-free. **MT**: Machine Translation, **TS**: Text Summarization, **DG**: Dialogue Generation, **IC**: Image Captioning, **D2T**: Data-to-Text, **SG**: Story Generation, **GE**: General Generation. We adopted the evaluation aspects for different tasks from Fu et al. (2023). Specifically, for each evaluation aspect, *CON*: consistency, *COH*: coherence, *REL*: relevance, *FLU*: fluency, *INF*: informativeness, *COV*: semantic coverage, *ADE*: adequacy, *NAT*: naturalness, *ENG*: engagement, *GRO*: groundness, *GRAM*: grammaticality, *LIK*: likability, *PER*: personalization, *QUA*: quality, *INT*: interest, *USE*: usefulness, *HARM*: harmlessness, *ACC*: accuracy, *FAC*: factuality, *ADA*: adaptability, *SUR*: surprise, *CHA*: character, *END*: ending, *FEA*: feasibility, *CRE*: creativity, *THO*: thoroughness, *STR*: structure, *LAY*: layout, *CLA*: clarity, *COM*: comprehensiveness, *FPR*: formality, *ADH*: adherence, *DEP*: topic depth, *UND*: understandability, *FLE*: flexibility, *INQ*: inquisitiveness, *SPE*: specificity, *COR*: correctness, *SEM*: semantic appropriateness. *NONE* means that the method does not specify any aspects and gives an overall evaluation. The detailed explanation of most evaluation aspect can be found in Fu et al. (2023).

more, Wang et al. (2023b) prompted LLM to generate quality scores for generated texts across various tasks, both with and without reference.

Probability-based Evaluation. Recognizing that the quality of the generated text is often correlated with the ease of generation by LLMs based on source or reference text, some studies frame the evaluation task as a conditional generation task. In this context, the generative likelihood of the produced text is calculated, serving as the score indicative of text quality, as illustrated in the second row of Table 4. Yuan et al. (2021) first leveraged BART (Lewis et al., 2019) as an evaluator to compute the probability of the generated text based on source or reference text in machine translation, text summarization, and data-to-text tasks. Fu et al. (2023) prompt LLM evaluator to calculate the generation probability of generated text with definitions of evaluation tasks and aspects. Unlike conventional use of generation probability as a quality score, Jia et al. (2023) calculated three probability changes to evaluate the faithfulness of the generated summary including changes with prior and conditional probability.

Likert-Style Evaluation. Inspired by the human annotation process, many studies employ LLM evaluators to assess the quality levels of generated texts based on a likert-style scale (Bai et al., 2023; Gao et al., 2023; Ostheimer et al., 2023; Gilardi et al., 2023; Huang et al., 2023; Zhao et al., 2023;

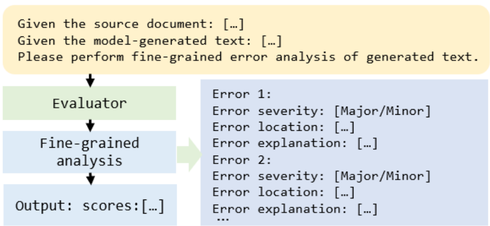

Figure 4: A example of fine-grained evaluation inspired by Jiang et al. (2023).

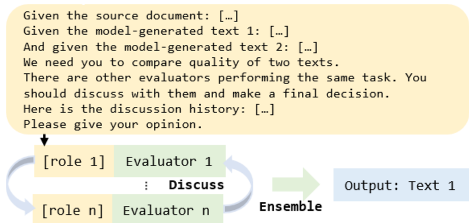

Figure 5: A example of ensemble evaluation inspired by Li et al. (2023c).

Wu et al., 2023b; Luo et al., 2023; Zheng et al., 2023; Zhuo, 2023; Sottana et al., 2023; Skopek et al., 2023). A representative likert-style prompt is depicted in the third line of Table 4. Chiang and Lee (2023) provided LLM evaluators with the same evaluation instructions as human annotators, prompting them to rate the quality of generated texts using a 5-point likert scale. Meanwhile, Gao et al. (2023) instructed ChatGPT to rate model-generated summarizations across multiple evaluation aspects, using a scale ranging from 1 (worst) to 5 (best) based on the provided source document. Ostheimer et al. (2023) designed multiple prompts, each guiding the LLM evaluator to assess a specific evaluation aspect of text style transfer task with a discrete scale. Liu et al. (2023f) utilized LLMs to draft, filter, and refine comprehensive evaluation criteria with a likert scale as score instructions.

Pairwise Evaluation. Compared with utilizing LLM evaluators to individually evaluate the quality of generated texts, another way is explicitly comparing with other generated text and decide which one is superior (Bai et al., 2023; Ji et al., 2023). A representative prompt is shown in the last row of Table 4. Wang et al. (2023c) employed LLM to assess a pair of model-generated responses, integrating a methodology involving multifaceted evidence and calibrated positioning, and leveraging human annotators if necessary to mitigate the influence of response pair order. Wang et al. (2023e) introduced a personalized evaluation framework prompting LLM to perform pairwise comparisons on three aspects: personalization, quality, and relevance.

Ensemble Evaluation. Since the evaluation process typically entails collaboration among multiple human annotators, some studies employ diverse LLM evaluators with varying base models or prompts, enabling assessments of text quality from different perspectives, as illustrated in Figure 5. Wu et al. (2023a) set multiple roles for the LLM to evaluate the quality of the generated summary by comparing it with the reference one on both subjective and objective dimensions. Li et al. (2023c) employed multiple LLM evaluators to conduct pairwise evaluations of model-generated responses which iteratively discuss comparison results. Besides, Chan et al. (2023) designed diverse communication strategies with various role prompts during collaborative discussions.

Advanced Evaluation. Some recent works investigate advanced evaluation to achieve comprehensive assessment outcomes by leveraging chain-of-thought, in-context learning capabilities, fine-grained analysis, etc (Jain et al., 2023; Min et al., 2023; Hasanbeig et al., 2023; Tang et al., 2023). A representative fine-grained evaluation method is shown in Figure 4. Liu et al. (2023c) utilized LLMs with chain-of-thought to evaluate the quality of generated texts across various NLG tasks and evaluation aspects. Lu et al. (2023) combined CoT to prompt the LLM evaluator to analyze different types of pre-defined errors in the generated translation, and then measured the quality of a generated translation. To enhance and improve the robustness of LLM-based evaluators, Hasanbeig et al. (2023) proposed ALLURE, a systematic protocol for auditing and improving LLM-based evaluation of text using iterative in-context-learning. Tang et al. (2023) leveraged LLMs to paraphrase a single reference into multiple high-quality ones in diverse expressions, which enhances evaluation methods on several NLG tasks. Liu et al. (2023f) mined and calibrated rubrics utilizing in-context learning to automatically align the LLM evaluator.

3.2 Tuning-based Evaluation

Researchers are also increasingly turn their attention towards fine-tuning open-source LLMs (e.g., LLaMA). In contrast to closed-based models demanding expensive API calls, the fine-tuning of smaller LLMs provides a cost-effective alternative. Additionally, the process of prompting LLMs for NLG evaluation requires meticulous crafting of prompts, with variations potentially resulting in significant differences in outcomes. Furthermore, the consideration of domain adaptability underscores the evolving landscape of NLG evaluation. Fine-tuning open-source LLMs affords researchers the flexibility to tailor models to diverse domains and tasks, transcending the limitations imposed by closed-based models confined to specific niches.

Likert-Style Evaluation. Some works tune LLMs to provide quality level or label for generated texts (Li et al., 2023a; Gekhman et al., 2023; Yue et al., 2023; Wang et al., 2023a; Kim et al., 2023a). Gekhman et al. (2023) employed FLAN-PaLM 540B (Chung et al., 2022) to annotate the quality of real model-generated summaries and utilized these annotated data as training data to tune a light-weight LLM (e.g., T5-11B) as a factual consistency summary evaluator. Li et al. (2023a) created a dataset containing multiple scenarios and used GPT-4 (OpenAI, 2023) to generate evaluation judgments for each scenario as supervision signals to tune LLaMA as a generative evaluator. Wang et al. (2023a) repurposed existing datasets with new personalized labels to tune LLaMA2 (Touvron et al., 2023) as a personalized story evaluation model which outputs a grade in [1, 10] and detailed reviews. Ke et al. (2023) collected referenced and reference-free data with dialogue-based prompting by instructing GPT-4, utilized which to tune LLMs for evaluating generated texts with explanations. Liu et al. (2023a) constructed a reference-free instruction-tuning dataset tailored for multi-aspect evaluation across various tasks, and tuned evaluator with auxiliary aspects additionally.

Probability-based Evaluation. Some works train generative LLMs to calculate the generation probability of generated texts to evaluate text quality. Thompson and Post (2020) trained a transformer as a multilingual reference-to-candidate paraphraser to obtain the generated probability of generated translation based on reference. Qin et al. (2022) tuned the T5 model in the generative and discriminative fashion, used which to calculate generative probability of generated text.

Pairwise Evaluation. There are also some works tuning LLMs for comparison between generated text pairs. Wang et al. (2023f) collected response pairs from LLMs and asked GPT-3.5 to generate output judgments, utilized which to tune LLaMA-7B to evaluate a pair of model-generated responses with the given query, accompanied by a concise description of the evaluation procedure. Zheng et al. (2023) performed fine-tuning on Vicuna using a human votes dataset from Chatbot Arena to pairwise evaluate two answers with the given query.

Advanced Evaluation. Nearly all tuning-based evaluators are trained to emulate evaluation behavior produced by strong closed models (e.g., GPT-4 or ChatGPT). Most studies gravitate towards ***holistic evaluation*** (Li et al., 2023a; Wang et al., 2023f,a; Kim et al., 2023a), which takes into account a diverse range of aspects to offer a holistic understanding of the quality of the hypothesis text. Besides, some studies explore ***error-oriented evaluation*** which focused on examining and explaining the specific errors in the hypothesis text, offering insights into why a particular score is derived. For instance, Yue et al. (2023) first defined different types of attribution errors, and then explored prompting LLMs or fine-tuning smaller LLMs on simulated and repurposed data from related tasks such as QA, NLI, and summarization. Xu et al. (2023) utilized GPT-4 to construct fine-grained analysis data to tune LLaMA as error-oriented evaluator, after which this work utilized real model-generated response-reference pairs to refine and self-train evaluator. Furthermore, Jiang et al. (2023) sampled data from diverse text generation datasets with real system output and GPT-4 synthesis, and tuned LLaMA using error analysis generated by GPT4 for fine-grained evaluation.

4 Comparing Traditional Evaluators

Qualitative Comparison Traditional evaluation metrics (e.g., BLEU (Papineni et al., 2002) and ROUGE) focus on exacting n-gram matches, which penalizes semantically correct but lexically different hypotheses. These methods are simple and fast but not robust to paraphrasing. BERTScore (Zhang et al., 2020) measures quality through semantic similarity based on BERT contextual embeddings, effectively handling paraphrases and synonyms. However, such matching-based evaluators depend on the quality of the pre-trained embeddings, may struggle with very fine-grained semantic distinctions, and neglect the overall semantics of the hypotheses and references. Additionally, neither metric accounts for fluency or readability during evaluation and both still rely on reference texts.

In contrast, LLMs have a strong capability for language understanding and generation, which supports evaluating quality without needing references. They can adapt to various domains and languages, making them suitable for a wide range of NLG tasks without requiring task-specific feature engi-

Metrics	Sup	SummEval					Topical-Chat					WMT22		
		COH	CON	FLU	REL	Avg	NAT	COH	ENG	GRO	Avg	En-De	En-Ru	Zh-Eu
Traditional Metrics (Word Overlap)														
ROUGE-1		0.167	0.160	0.115	0.326	0.192	0.158	0.206	0.319	0.264	0.233	-	-	-
ROUGE-2		0.184	0.187	0.159	0.290	0.205	0.168	0.247	0.337	0.311	0.266	-	-	-
ROUGE-L		0.128	0.115	0.105	0.311	0.165	0.145	0.205	0.306	0.293	0.237	-	-	-
BLEU		-	-	-	-	-	0.175	0.235	0.316	0.310	0.259	0.169	0.140	0.145
BERT-based Metrics														
BERTScore		0.284	0.110	0.193	0.312	0.225	0.209	0.233	0.335	0.317	0.273	0.232	0.192	0.316
BLEURT	✓	-	-	-	-	-	-	-	-	-	-	0.344	0.359	0.361
BARTScore	✓	0.448	0.382	0.356	0.356	0.385	-0.053	-0.079	-0.084	-0.197	-0.103	-	-	0.220
UniEval	✓	0.575	0.446	0.449	0.426	0.474	0.450	0.616	0.615	0.590	0.568	-	-	-
LLM-based Metrics														
GPTScore		0.434	0.449	0.403	0.381	0.417	-	-	-	-	-	-	-	0.187
CHATGPT(DA)		0.451	0.432	0.380	0.439	0.425	0.474	0.527	0.599	0.576	0.544	0.306	0.332	0.371
G-Eval		0.582	0.507	0.455	0.547	0.514	0.607	0.590	0.605	0.536	0.590	-	-	-
Embed Llama		-	-	-	-	-	-	-	-	-	-	0.400	0.227	0.217
X-Eval	✓	0.530	0.428	0.461	0.500	0.480	0.478	0.622	0.593	0.728	0.605	-	-	-

Table 2: Performance of traditional and LLM-based metrics on Summarizing (SummEval), Dialogue (Topical-Chat) and MT (WMT22) tasks. We demonstrate the sample-level Spearman correlations on SummEval and Topical-Chat benchmarks and the segment-level Kendall-Tau correlations on WMT22 benchmarks respectively. **Sup** indicates the metric is supervised. The specific meaning of the evaluation aspects is shown in Table 1.

neering. LLMs also provide more nuanced evaluation criteria beyond traditional metrics, such as semantic coherence, fluency and possible explanations. However, LLM-based methods are computationally more intensive due to their vast architectures. Additionally, prompting LLMs for NLG evaluation requires careful crafting of prompts. Variations in these prompts can lead to substantial differences in evaluation outcomes, as indicated in (Gao et al., 2023). Section 5 summarizes more open problems of LLM-based metrics.

Performance Comparison Table 2 summarizes the performance of both traditional word-overlap metrics, BERT-based metrics and recent LLM-based metrics on representative benchmarks such as SummEval, WMT, and Topical-Chat. We can easy to observe that the latter two metrics generally perform better than word-overlap metrics. Despite not being fine-tuned, the most competitive LLM-based methods (e.g., G-Eval for summarization and CHATGPT(DA) for machine translation) generally achieve a higher correlation with all traditional metrics, whether for unsupervised or fine-tuned methods. These results reveal the strong capability of LLMs in language understanding, contextual analysis, coherence checking, and fluency assessment of generated text. Among the three tasks, the performance gap between LLM-based evaluators and traditional evaluators is not significant in the machine translation task. This phenomenon might be due to the limitations of LLM-based models in cross-lingual understanding. Additionally, according to the results of last row in the table, we can observe that the performance of different LLM-based metrics varies significantly, which implies their sensitivity to prompt crafting. In contrast, traditional unsupervised methods like ROUGE, BLEU, and BERTScore are more robust, although their overall performance is relatively worse.

Efficiency Comparison Table 3 presents the average number of texts evaluated per second for different metrics in the SummEval (TS task) and Topical-chat (DG task) benchmarks. This comparison highlights the efficiency differences between traditional metrics and LLM-based metrics. Our tests were conducted on an NVIDIA A40 GPU. The results show that efficiency generally correlates with model size and traditional word-overlap metrics (e.g., BLEU and ROUGE) are significantly faster than other metrics. Specifically, LLM-based evaluators are about 200 to 400 times slower than traditional word-overlap metrics. However, their efficiency can be improved with advanced LM inference tools such as vLLM. While LLM-based evaluators are suitable for offline evaluation, they may not be feasible for online evaluation.

5 Open Problems

Despite significant efforts and achievements in various benchmarks, several challenges persist for LLM-based evaluators.

Bias of LLM-based Evaluators. The use of LLMs as evaluators inherently cast the text evaluation as a generation task. Consequently, when

Methods	Backbone	TS	DG
BLEU	-	977.31	2344.16
ROUGE	-	446.36	2379.24
BERTScore	BERT	37.64	42.37
ChatGPT(DA)	ChatGPT	1.94	1.87
G-Eval	GPT-4	1.51	1.40
TIGERScore	Llama	2.67	3.72

Table 3: The average number of texts evaluated per second for different metrics.

LLMs are employed in this evaluator role, they may carry over biases intrinsic to their function as generators. These biases may include social biases, such as stereotypes related to specific demographic identities (e.g., race, gender, religion, culture, and ideology) (Sheng et al., 2021). In addition to these general biases, LLMs-as-evaluators are subject to specific biases unique to their evaluative role. These include order bias, where preference is given to options based on their sequence (Zheng et al., 2023; Wang et al., 2023c); egocentric bias, where a tendency exists to favor texts generated by the same LLM (Liu et al., 2023d; Koo et al., 2023); and length bias, which leads to a preference for longer or shorter texts (Zheng et al., 2023).

Robustness of LLM-based Evaluators. Most LLMs-based evaluation methods rely heavily on prompt engineering. However, the process of prompting LLMs for NLG evaluation demands careful crafting of prompts. The variations in these prompts can potentially lead to substantial differences in the outcomes of the evaluation process. As demonstrated in Liu et al. (2023e) and Koo et al. (2023), LLMs exhibit limited robustness when subjected to the adversarial dataset containing incorrect facts, irrelevant information, or fabricated statistics. The robustness of LLM-based evaluators emerges as a critical area of exploration, underscoring the need for further research to enhance their robustness in the face of challenging or misleading inputs.

Which came first, the chicken or the egg? If the evaluator possesses capabilities comparable to the model being evaluated, e.g. using GPT-4 to evaluate GPT-4 itself, there may exist egocentric issue of favoring their own generated responses (Bai et al., 2023). This scenario mirrors the chicken-and-egg dilemma: an LLM-based evaluator relies on a more powerful LLM, yet the development of a more powerful LLM depends on having a robust evaluator. To address this dilemma, a broader spectrum of evaluation method is necessary, involving various benchmark (Srivastava et al., 2022; Liang et al., 2022), evaluation criteria (Sellam et al., 2020), and human feedback (Xu et al., 2023; Ouyang et al., 2022) to ensure more comprehensive assessments.

Domain-Specific Evaluation. Most LLM-based evaluators are general-purpose and not tailored to specific domains. The domain-specific evaluation poses significant challenges of checking domain factuality and designing specific evaluation prompts. For example, while evaluating legal documents, aspects such as legal accuracy and adherence to the judicial system are crucial (Cui et al., 2023). Therefore, to enhance the efficacy of LLMs as evaluators in specialized domains, there's a pressing need to develop models that are not only domain-aware but also equipped with the capability to evaluate based on domain-specific criteria.

Unified Evaluation. As LLMs become increasingly versatile, there is a need for more comprehensive and flexible assessment methods. However, most current LLM-based evaluators are limited to constrained tasks and aspects (cf. Table 1). Some promising attempts have been made in this direction. For instance, MT-Bench (Zheng et al., 2023) uses GPT-4 as an evaluator across multiple domains for multi-turn questions. Another model, Auto-J (Li et al., 2023b), accommodates diverse evaluation protocols and has been validated in 58 different scenarios. In light of increasingly diverse user queries, developing a more unified evaluation protocol is a promising direction. Additionally, constructing high-quality, comprehensive datasets to train unified models holds great potential.

6 Conclusion

In this paper, we have comprehensively surveyed the role of LLMs in the evaluation of NLG. Our comprehensive taxonomy classifies works along three primary dimensions: evaluation function, evaluation references and evaluation task. Additionally, we summarize holistic LLM-based approaches and prevalent meta-evaluation benchmarks for NLG evaluation. Through our paper, we highlight unresolved issues, including bias, robustness, and the need for domain-specific and unified evaluation within LLM-based evaluators. We anticipate that addressing these challenges will pave the way for more reliable, general, and effective LLM-based NLG evaluation techniques.

7 Limitations

In this paper, we propose an overview of leveraging large language models for NLG evaluation. This paper provides a comprehensive overview about the usage of LLM evaluators in evaluation of NLG tasks. Nevertheless, due to space restrictions, we are unable to provide further details on LLM evaluators and meta-evaluation benchmarks in this survey. Additionally, we do not compare the performance of various LLM evaluators in the paper. Furthermore, as LLM-based NLG evaluation field is rapidly evolving, our paper may not include the latest LLM evaluators which are emerged shortly before or after its completion. In the future, we plan to demonstrate more detailed information for each LLM evaluators and track the latest progress through updating periodically GitHub repository.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

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

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

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

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

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

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

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

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George 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 the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Z. Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *ArXiv*, abs/2308.11995.

Rupali Goyal, Parteek Kumar, and VP Singh. 2023. A systematic survey on automated text generation tools and techniques: application, evaluation, and challenges. *Multimedia Tools and Applications*, pages 1–56.

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

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

Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. 2023. Allure: A systematic protocol for auditing and improving llm-based evaluation of text using iterative in-context-learning. *arXiv preprint arXiv:2309.13701*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Sameer Jain, 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. *arXiv preprint arXiv:2306.01200*.

Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences. *arXiv preprint arXiv:2303.07610*.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Q Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. *arXiv preprint arXiv:2310.11648*.

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

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165.

Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023b. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. *arXiv preprint arXiv:2309.13633*.

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

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Christoph Leiter, 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. *arXiv preprint arXiv:2310.19792*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470.

Ruosen Li, Teerth Patel, and Xinya Du. 2023c. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023d. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2023a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. *arXiv preprint arXiv:2311.08788*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023b. Alignbench:

Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

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

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023d. Llms as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

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

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

Qingyu Lu, 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. *arXiv preprint arXiv:2303.13809*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

OpenAI. 2023. Gpt-4 technical report.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization. *arXiv preprint arXiv:2310.18122*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293.

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, and Victor Carbune. 2023. Towards better evaluation of instruction-following: A case-study in summarization. *arXiv preprint arXiv:2310.08394*.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing. *arXiv preprint arXiv:2305.15067*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

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

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

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2023a. Learning personalized story evaluation. *arXiv preprint arXiv:2310.03304*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

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

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

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2698–2703.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023e. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

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

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

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

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023a. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023b. Less is more for long document summary evaluation by llms. *arXiv preprint arXiv:2309.07382*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. *arXiv preprint arXiv:2305.14282*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

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

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xinghua Zhang, 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. *arXiv preprint arXiv:2308.01862*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios.

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

Terry Yue Zhuo. 2023. Large language models are state-of-the-art evaluators of code generation. *arXiv preprint arXiv:2304.14317*.

A Appendix

B Benchmarks and Tasks

Numerous meta-evaluation benchmarks serve the purpose of validating the efficacy of NLG evaluators. These benchmarks incorporate human annotations gauging the quality of generated text, and evaluating the degree of concurrence between automatic evaluators and human preferences. Categorized based on the tasks involved, these benchmarks can be classified into single-scenario examples, such as summarization, as well as multi-scenario benchmarks. This section will provide an overview of these NLG tasks and their associated meta-evaluation benchmarks.

Machine Translation (MT). MT task is centered around converting a sentence or document from a source language into a target language while preserving the same semantic meaning. The Annual WMT Metrics Shared tasks (Freitag et al., 2021b, 2022) annually introduce a set of benchmarks encompassing model-generated translations, source text, reference text, and human judgment across multiple languages. Simultaneously, Freitag et al. (2021a) curated and annotated outputs from 10 translated systems for translation pairs in the WMT 2020 news translation task (Barrault et al., 2020). They used professionals and crowd workers to rate translations on a 7-point scale using multi-dimensional metrics.

Text Summarizing (TS). TS involves generating a summary of a given text while capturing its essential meaning. There are many meta-evaluation benchmarks proposed (Grusky et al., 2018; Gliwa et al., 2019; Bhandari et al., 2020; Wang et al., 2020b; Pagnoni et al., 2021; Laban et al., 2022; Skopek et al., 2023; Shen and Wan, 2023). One of the widely used benchmarks is SummEval (Fabbri et al., 2021b) which includes summaries generated by 16 models from 100 source news articles. Each summary underwent annotation by crowd-sourced workers and experts on four dimensions: coherence, consistency, fluency and relevance. In addition, Shen and Wan (2023) presented a meta-evaluation benchmark for opinion summarization tasks, including human judgments and outputs from 14 models over four dimensions.

Dialogue Generation (DG). DG task aims to generate human-like responses in the context of a conversation which should be natural and consistent. Mehri and Eskenazi (2020b) performed human annotations across two open-domain dialog corpora Topical-Chat (Gopalakrishnan et al., 2019) and PersonaChat (Zhang et al., 2018), where each response is scored from 6 dimensions including naturalness, coherence, engagingness, groundedness, understandability and overall quality. Similaritily, Mehri and Eskenazi (2020a) sampled and annotated a subset from a set of conversations across eighteen dialog quality dimensions.

Image Caption (IC). The task involves generating textual descriptions or captions for images. Meta-evaluation benchmarks of IC contain human annotations for image-textual pairs or hypothesis-reference caption pairs (Aditya et al., 2015; Vedantam et al., 2015; Cui et al., 2018). For example, the commonly used Flickr 8k dataset (Hodosh et al., 2013) collected human annotations from both expert and CrowdFlower for each image-caption pair. Cui et al. (2018) collected human judgments for twelve submission entries with reference captions from the 2015 COCO Captioning Challenge on the COCO validation set (Vinyals et al., 2016).

Data-to-Text (D2T). D2T task involves generating fluent and factual human-readable text from structured data. Mairesse et al. (2010) proposed BAGEL, which contains 202 structured information samples about restaurants in Cambridge. Wen et al. (2015) further proposed SFRES and SFHOT, which contain 581 samples of restaurants and 398 samples of hotels in San Francisco, respectively.

Story Generation (SG). The task involves creating relevant narratives or stories with the given beginning of a story or writing requirement. Most meta-evaluation benchmarks of story generation always contain stories and corresponding manually annotated judgment scores (Guan et al., 2021; Chen et al., 2022). Besides, Wang et al. (2023a) created two personalized story evaluation benchmarks denoted as Per-MPST and Per-DOC. This work repurposed existing datasets (Kar et al., 2018; Yang et al., 2023) through anonymizing and summarizing. Both them provide personalized human judgements for each generated story.

General Generation (GE). As LLMs have been increasingly used in general NLG tasks, LLM evaluators have been proposed to effectively evaluate the generated texts across multiple scenario (Kim et al., 2023a; Ke et al., 2023). Accordingly, there are many multi-scenario meta-evaluation benchmarks (Wang et al., 2023c; Zheng et al., 2023; Wang et al., 2023d; Yue et al., 2023; Liu et al.,

Prompt Type	Prompt	Output

Score-based	`Given the source document: [. . . ]`
	`Given the model-generated text: [. . . ]`
	`Please score the quality of the generated text from 1 (worst) to 5 (best)`
		`Scores: 2`

Likert-style	`Given the source document: [. . . ]`
	`Given the model-generated text: [. . . ]`
	`Is the generated text consistent with the source document? (Answer Yes or No)`
		`Yes`

Pairwise	`Given the source document: [. . . ]`
	`Given the model-generated text 1: [. . . ]`
	`And given the model-generated text 2: [. . . ]`
	`Please answer which text is better-generated and more consistent.`
		`Text 1`

Table 4: Illustration of different types of prompts.

2023b; Zeng et al., 2023). Typically, Zhang et al. (2023b) sampled 2,553 evaluation samples, including instructions and generated responses with corresponding human-annotated labels from multiple tasks. Additionally, Zeng et al. (2023) introduced a benchmark divided into NATURAL and ADVERSARIAL sets. The former set comprises instances from human-preference benchmarks, ensuring objective preferences. The latter set contains instances created by authors to challenge evaluators, deviating from instructions but maintaining superficial quality.