Large Language Models for NLP Evaluation: A Survey

Harshvivek Kashid and **Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
{harshvivek, pb}@cse.iitb.ac.in

Abstract

Evaluating natural language generation (NLG) remains a critical yet challenging task in artificial intelligence. Traditional metrics, which largely rely on surface-level measures such as n-gram overlap between system outputs and references, often fail to capture deeper aspects of quality. In contrast, large language models (LLMs), such as ChatGPT, have recently demonstrated strong potential as evaluators of NLG outputs. A growing body of LLM-based evaluation approaches has emerged, including LLM-derived metrics, prompting strategies, and fine-tuning on labelled evaluation data. This survey presents a taxonomy of these methods, analyzing their strengths and limitations, and explores the potential for human-LLM collaboration in evaluation. We also review key open challenges and promising directions for future research. The survey covers the application of LLM-based evaluation across diverse NLP tasks—including summarization, machine translation, dialogue, and question-answering—and spans both reference-based (e.g., parallel corpora) and reference-free evaluation paradigms.

1 Introduction

The automatic evaluation of generated text is a long-standing challenge in The field of natural language generation (NLG) has seen rapid progress in recent years, driven by advances in neural architectures and pre-trained language models. Tasks such as abstractive summarization, open-domain dialogue, machine translation, and question answering are now dominated by large-scale models with remarkable fluency and versatility. However, the evaluation of NLG systems remains an open challenge. Traditional automatic metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and even more recent embedding-based metrics like BERTScore (Zhang et al., 2019), provide limited insight into key dimensions of generation quality, such as factual accuracy, coherence, faithfulness, and overall helpfulness. These shortcomings are particularly acute in tasks with diverse valid outputs or lacking high-quality references.

Large language models (LLMs), including ChatGPT, GPT-4, and open models such as LLaMA and Mistral, have recently emerged as promising tools for evaluating NLG outputs (Chiang and Lee, 2023a; Siledar et al., 2024; Kobayashi et al., 2024). LLMs can provide nuanced, context-sensitive judgments without requiring large annotated datasets or expensive human evaluation. Consequently, there is growing interest in leveraging LLMs as universal evaluators, capable of assessing generation quality across tasks and domains (Li et al., 2024).

In this survey, we present a structured overview of the current landscape of LLM-based evaluation. We describe a taxonomy of methods—including prompting-based evaluation, LLM-derived metrics, and fine-tuned evaluators—and analyze their effectiveness across key NLG tasks. We also discuss the trade-offs between reference-based and reference-free evaluation, the design of evaluation prompts, and the role of human-LLM collaboration. Finally, we identify open research questions and outline promising directions for future work.

2 Motivation

Existing automatic metrics have well-known shortcomings. Reference-based metrics depend on one or a few gold outputs, which may not capture all valid variations and often reward surface overlap rather than true meaning. Reference-free metrics (e.g. embedding-based or entailment-based) partially address this, but often require task-specific training or heuristics. Meanwhile, state-of-the-art LLMs exhibit strong linguistic competence and world knowledge. This raises the hypothesis that they could serve as effective proxies for human evaluators, assigning higher scores to higher-quality outputs and vice versa.

Early experiments support this: LLM-based metrics have shown improved correlation with human ratings on several tasks. Moreover, LLMs allow the evaluation criteria to be specified flexibly in natural language. By framing evaluation as a prompting task, one can instruct the model on what dimensions to check (e.g. grammar, fluency, factuality) and obtain multi-dimensional feedback. Such flexibility is a key advantage of LLM evaluation, motivating the exploration of LLMs as general-purpose evaluators.

3 Problem Statement

We formalize the task of LLM-based evaluation as follows: Given an input (e.g. a source text or dialogue history) and one or more candidate outputs (e.g. system-generated summaries, translations, or responses), an LLM evaluator should assign a quality score or ranking to each candidate. Quality may be judged along dimensions such as fluency, relevance, coherence, or factual accuracy.

We consider two main paradigms:

Reference-based evaluation: Here human-written references are available. The evaluator can directly compare each candidate against the reference. For example, in machine translation evaluation, one might ask the LLM to judge adequacy or error types by comparing the translation to the reference. Traditional metrics like BLEU/ROUGE fall into this category.

Reference-free evaluation: Here no gold reference is given; the evaluator judges each output on its own merits (possibly considering the input). This mimics human preference judgment or rating. Implementation often involves prompting the LLM to critique or rate the output directly.

Additional considerations include whether evaluation is done at the sentence-level or document-level, and whether the LLM produces a single scalar score or multiple aspect-specific scores. Crucially, we distinguish between *prompt-based* evaluation (using the LLM with an engineered prompt) and *fine-tuned* evaluation (adapting the LLM via training on annotated data).

4 Background: Traditional Metrics and LLMs based evaluation

Classic automatic metrics include BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) for text overlap, as well as newer metrics like METEOR (Banerjee and Lavie, 2005). Embedding-based measures such as BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), and BARTScore (Yuan et al., 2021) compare semantics with references. These metrics are easy to compute, but many studies have shown they often fail to capture true quality. For example, they may give high scores to trivial paraphrases or fluent but irrelevant text. Reiter and Belz (2009) and others have demonstrated poor validity of these metrics for open-ended generation tasks.

Reference-free evaluation has a longer history as well. QA-based approaches generate questions about the source and check if the candidate provides consistent answers. Notable examples include QAGS (Wang and Lewis, 2020), which generates questions from the source and checks consistency of the summary, and FEQA, a similar QA-based method. For dialogue, metrics like USR (Mehri and Eskenazi, 2020) use learned predictors of coherence, consistency, and engagement without references. However, these methods often rely on additional NLP systems or training data and can still miss subtle issues.

As LLMs continue to advance, one of the biggest challenges in utilizing these models is finding effective ways to evaluate them. With numerous powerful models available, each capable of tackling a range of complex and open-ended tasks, it becomes challenging to distinguish their performance differences. Human feedback is the most dependable means of assessing LLMs, but gathering this feedback is often noisy, time-consuming, and expensive. While human evaluation provides essential insights into model capabilities, relying solely on it can hinder rapid iteration during model development. Therefore, we require an evaluation metric that is efficient, cost-effective, and straightforward, yet still closely aligns with human evaluation outcomes.

Gillick and Liu (2010) and Guzmán et al. (2015) discuss the significance of human evaluation in understanding the performance of NLP models. However, human evaluation poses challenges, such as difficulties in reproducibility and unstable quality (Clark et al., 2021). Additionally, obtaining reference-based datasets at a large scale can be an expensive process. Recent studies highlight the effectiveness of LLMs as reference-free evaluators for NLG outputs (Wang et al., 2023; Liu et al., 2023a). For instance, G-EVAL by Liu et al. (2023a) leverages LLMs with chain-of-thought (CoT) rea-

soning and a form-filling paradigm, achieving high correlation with human judgments on summarization tasks. In a related effort, Kocmi and Federmann (2023a) utilized GPT models to assess various machine learning tasks, demonstrating the versatility of these models in different evaluation contexts (Wei et al., 2022b) and assigns weights to a predetermined set of integer scores based on their generation probabilities derived from either GPT-3 or GPT-4, allowing for a nuanced assessment of text quality.

Chen et al. (2023) were among the first to investigate reference-free evaluation techniques for NLG using LLMs, concluding that obtaining an explicit score generated by ChatGPT is the most effective and stable approach for evaluating text quality without relying on reference outputs. Fu et al. (2024) introduced GPTScore, which is founded on the concept that generative pre-training models, such as GPT-3, are more likely to assign higher probabilities to the creation of high-quality text that aligns with the instructions and context provided. This suggests that these models can effectively assess the quality of generated text based on its adherence to specified guidelines. In their pioneering work, Chiang and Lee (2023a) were the first to delve into the use of large language models (LLMs) for evaluation tasks, opening up new avenues for assessing model outputs. Following this, Chiang and Lee (2023b) presented detailed guidelines aimed at enhancing the correlation between the evaluations performed by ChatGPT and those made by human judges, thereby improving the reliability of automated assessments. The experiments conducted by Zheng et al. (2023) demonstrate that powerful LLMs like GPT-4 can reach agreement levels that are comparable to those of human evaluators. This finding underscores the potential of these models to approximate human preferences accurately, making them valuable assets in the field of natural language processing. While LLMs have shown promise as reference-free metrics for NLG evaluation, their potential remains unexplored for opinion summary evaluation. Siledar et al. (2024) gives a detailed analysis, comparing an open-source LLM against a closed-source LLM acting as evaluators for automatic evaluation of opinion summaries on seven dimensions: *fluency, coherence, relevance, faithfulness, aspect coverage, sentiment consistency, and specificity*.

5 Different methods used for evaluation

5.1 Prompting and Inference

The main use of LLMs as evaluators is via prompting. **Zero-shot prompting** involves giving the model an instruction like "Rate this summary on a scale from 1 to 5." In one example, Fu et al. (2023) proposed GPTScore, which uses an LLM's conditional probability of a special token as a quality score. **Few-shot prompting** includes example (input, output, score) tuples in the prompt to guide the model; this can improve consistency but is limited by context length and example selection.

Another approach is to leverage the LLM's generative nature. One can ask the model to critique the output in free text or to answer specific quality questions about it. A popular technique is **Chain-of-Thought (CoT) prompting**, where the LLM is prompted to generate intermediate reasoning steps before giving a final score. Wei et al. (2022a) showed that CoT can improve reasoning; in evaluation settings, Lu et al. (2023) use CoT to break down translation errors before scoring.

Some methods use **likelihood scoring**: measuring the probability of the model generating the output itself. GPTScore (Fu et al., 2023) is analogous to using perplexity: it interprets higher output likelihood as higher quality. This avoids phrasing bias from question prompts.

5.2 Fine-Tuning and Adaptation

Beyond prompting, researchers have fine-tuned LLMs on evaluation tasks. For example, Zhong et al. (2022) trained UniEval, a BERT model, on multi-dimensional human ratings. More recent work fine-tunes large LMs on annotated evaluator data. Parameter-efficient tuning (e.g. LoRA) makes this feasible. Kartáč et al. (2025) fine-tuned LLaMA 3.1B with LoRA on synthetic preference pairs to mimic human judgments. Reinforcement learning with human feedback (RLHF) techniques have also been used to align LLM outputs to human evaluation criteria.

Table 2 summarizes common configurations and strategies encountered in LLM-based evaluation studies.

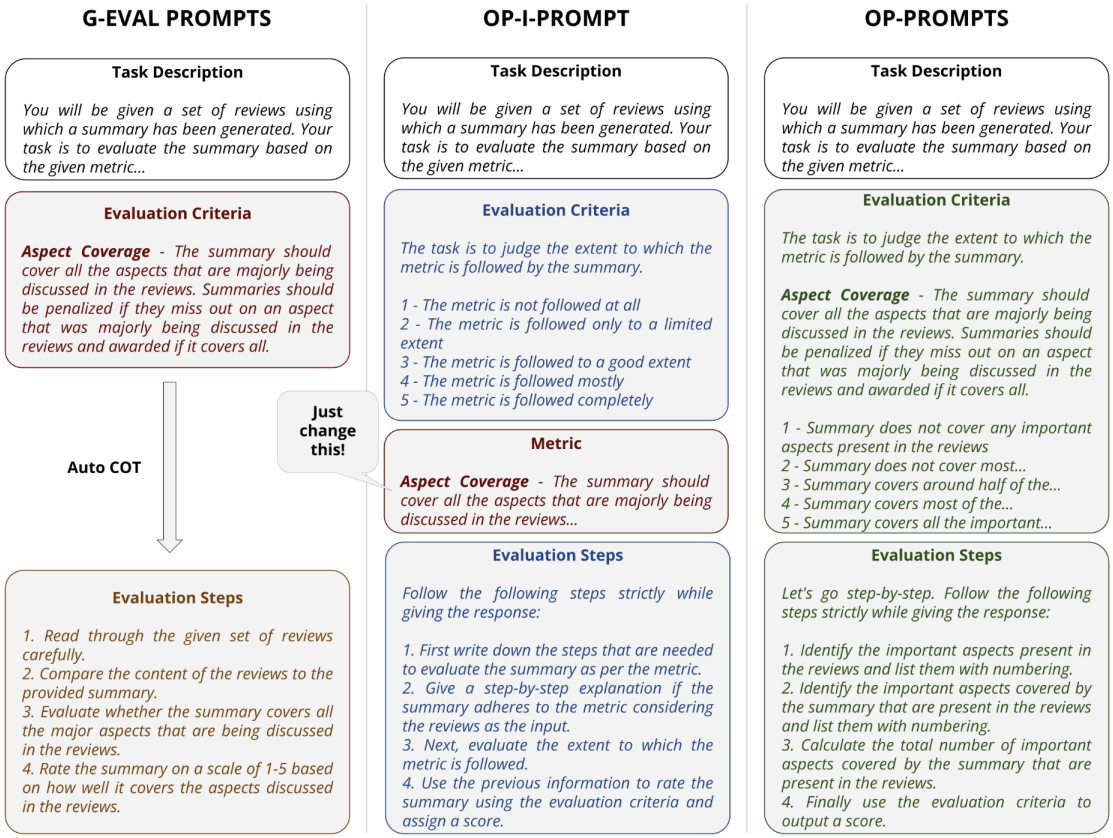

Figure 1: Comparison of Prompt Approaches. G-EVAL PROMPTS first generates the Evaluation Steps using Task Description and Evaluation Criteria in Chain-of-Thought fashion. Finally the full prompt is used to evaluate the opinion summaries. In contrast, OP-I-PROMPT is simpler and has Task Description, Evaluation Criteria, and Evaluation Steps fixed for a dimension/metric independent evaluation. Here, only the Metric part needs to be changed for evaluating any dimension/metric. Finally OP-PROMPTS are dimension/metric dependent prompts that needs to be specifically crafted for each dimension/metric.

6 Case Study

6.1 LLM based evaluation of *GrahakNyay* chatbot

Consumer grievance redressal remains a complex process due to procedural complexity, legal jargon, and barriers like jurisdiction and limitation periods, despite various initiatives aimed at simplifying it. ***Grahak-Nyay*** (Justice-to-Consumers) is a chatbot designed to simplify the consumer grievance redressal process for Indian consumers. Powered by open-source LLMs with Retrieval-Augmented Generation (RAG), *Grahak-Nyay* is supported by a concise and up-to-date Knowledge Base.

6.1.1 Evaluation of GrahakNyay chatbot

We strongly believe that any user-facing chatbot should help the user address the query, should be accurate while doing so, and should keep the user engaged. We assess the quality of chatbot conversations using **HAB metrics**: **H**elpfulness, **A**ccuracy, and **B**revity. HAB metrics allow us to assess not only how effectively the chatbot addresses user issues and provides accurate information but also how concisely it communicates these responses.

The HAB metrics are defined as follows:

- **Helpfulness**: This metric assesses how helpful the chatbot was in resolving the user's issue or query. It evaluates chatbot's ability to understand the user's problem and provide actionable, relevant, and clear resolution.
- **Accuracy**: This metric evaluates the correctness of the information provided by the chatbot in response to user queries, ensuring that the responses are factually accurate and reliable.
- **Brevity**: This metric measures the conciseness of the chatbot's responses, ensuring efficient communication without unnecessary

Table 1: Representative datasets and benchmarks for evaluating LLM-based metrics.

Task	Dataset / Benchmark	Notes / References
Summarization	CNN/DailyMail, XSum, Multi-News	Reference summaries of news articles; standard for summarization evaluation
	SummEval (Fabbri et al., 2021)	Human judgments of coherence, relevance, etc. on CNN/DailyMail summaries
Machine Translation	WMT Metrics Tasks	Test sets with MQM human ratings on translation outputs.
Dialogue	TopicalChat (Gopalakrishnan et al., 2023)	Multi-turn knowledge-grounded dialogues; annotated for response coherence and engagement
	USR (Mehri and Eskenazi, 2020)	Reference-free response quality metrics for dialogue
QA / Consistency	QAGS (Wang and Lewis, 2020)	QA-based factual consistency test for summaries
	FEQA (Durmus et al., 2020)	QA-based evaluation for summary factuality

Table 2: Common LLM evaluation settings and methodologies.

Setting	Description	Example / Citation
Zero-shot prompting	Prompt the LLM to judge outputs directly with instructions, without examples.	GPTScore (Fu et al., 2023) uses GPT probabilities for quality; GEMBA (Kocmi and Federmann, 2023b) prompts GPT-3.5 on WMT translations.
Few-shot prompting	Include a few annotated examples in the prompt for in-context learning.	Used in tutorials and chat-based evaluations (e.g. few-shot ChatGPT demos). Selection of examples affects bias.
Chain-of-Thought (CoT)	Prompt the LLM to generate intermediate reasoning steps or criteria before the final judgment.	EAPrompt (Lu et al., 2024) first generates an error analysis then a score; G-EVAL (Liu et al., 2023b) instructs detailed evaluation steps.
Likelihood scoring	Use the LLM's log-likelihood of the output (or a special token) as a quality score.	GPTScore (Fu et al., 2023) interprets generation probability as quality, similar to perplexity-based metrics.
Fine-tuning / LoRA	Adapt the LLM by training on labeled examples of high/low quality or human scores. Often uses low-rank adapters (LoRA).	UniEval and similar models fine-tuned on human ratings; OPENLGAUGE (Kartáč et al., 2025) proposed an explainable metric for NLG evaluation with open sourced LLMs.

elaboration. It ensures efficient communication by focusing on delivering essential information straight to the point, while avoiding excessive questioning or verbosity.

Through this comprehensive evaluation framework, we aim to enhance the effectiveness of chatbots in addressing consumer grievances and improving overall user satisfaction.

The table 4 presents detailed results from the evaluation of 65 chats conducted by the Grahak-Nyay chatbot, categorized into two groups: Reference-based and Reference-free evaluations. For these 65 chats, reference responses annotated by the legal experts were available, enabling the application of Reference-based metrics. Additionally, for the Reference-free evaluation, we utilized HAB metrics to assess the chatbot's performance in providing relevant and concise responses. We used the best-performing model, Llama-3.1-70B model, which demonstrated the highest correlation with human evaluations, for the assessment of the HAB metrics.

To reduce human effort in evaluating the chatbot according to HAB metrics, we employ LLM-based automatic evaluation. The LLM evaluators are instructed to assign scores on a 5-point Likert scale and provide detailed explanations for their assigned scores using the structured prompt (Fig. 2, 3, and 4). The prompt includes task description, scoring instructions based on the HAB metrics, as well as the conversation which is to be evaluated and the context[1].

We evaluated 75 conversations for which we have human-evaluated data available in binary form (Yes, if the metric is followed, No if not), on the HAB metrics, using different LLMs sourced from HuggingFace[2] and Groq[3]. The table 3 summarises

[1]Context is passed only for the *Accuracy* metric.
[2]https://huggingface.co
[3]https://groq.com/

Models	Helpfulness		Accuracy		Brevity	
	r_{pb}	ρ	r_{pb}	ρ	r_{pb}	ρ
Gemma-2-9B	0.256	0.242	0.113	0.102	0.183	0.182
Mixtral-8x7B	0.557	0.490	0.205	0.207	0.159	0.141
Llama-3.3-70B	<u>0.689</u>	<u>0.627</u>	<u>0.565</u>	<u>0.530</u>	<u>0.433</u>	<u>0.418</u>
gpt-4o-mini	**0.719**	**0.687**	**0.571**	**0.540**	**0.675**	**0.595**

Table 3: Performance metrics for various models based on Helpfulness, Accuracy, and Brevity metrics. Each metric includes point biserial correlation (r_{pb}) and Spearman's rank correlation coefficient (ρ) scores for each model. The best scores are bolded, and the second-best scores are underlined.

Reference-based evaluation						Reference-free evaluation		
ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	BLEU	Helpfulness	Accuracy	Brevity
66.9	41.1	33.2	90.9	41.9	37.4	4.65	3.61	3.58

Table 4: Performance of *Grahak-Nyay chatbot* on Reference-based and Reference-free evaluation. We evaluated the Grahak-Nyay chatbot on 65 conversations for which reference was available. We performed LLM-based automatic evaluation on HAB metrics on the 5-point Likert scale using the gpt-4o-mini model.

the performance of LLM-based evaluators for HAB metrics. We applied point biserial correlation to assess the relationship between the available binary human evaluation and the ordinal LLM scores from the 5-point Likert scale. This correlation is particularly useful in determining how well the LLM evaluations align with the binary outcomes. Additionally, we used Spearman correlation to evaluate the rank order of scores, providing further insights into the agreement between human and LLM evaluations. The `gpt-4o-mini` model consistently outperformed others across all three metrics, achieving the highest point biserial correlation and Spearman's correlation coefficients with *p-value<0.05*, indicating its superior effectiveness.

6.2 LLM based evaluation of *Nyay-Darpan* Summarization tool

Judicial assistance and case verdict prediction through AI have always been a challenge. Although much work has been presented on the front of similar case prediction in criminal and civil cases, and NLP tools being created for them, the domain of consumer laws, specifically in India, remains almost untouched. In this paper, we propose a two-in-one methodology for summarizing case files and extracting similar case files, aiming at assisting the decision-making process in redressal of consumer cases not only in India but throughout the world and also a novel methodology for predicting the efficacy of the summarization. We name our tool as ***Nyay-Darpan***, which reflects the contents of a case file by acting as a "mirror of justice". The term "*Nyay*" translates to "justice", and "*Darpan*" means "mirror", symbolizing our tool's ability to reflect the essence of consumer case files through comprehensive summarization and intelligent case matching. Through this reflection, we managed to achieve more than 75 percent accuracy on similar case prediction and around 70 percent accuracy across all material summary evaluation parameters.

6.2.1 Summary Structure

The input consists of the complaint document and the written statement. This includes all textual data related to the complaint, the parties involved, claims, and fragments of evidence provided by the complainant and the opposite party. The system prompt, combined with the case document, is inputted into the LLM. The output is a structured summary that includes the following:

- **Overview:** This section presents a factual summary of the consumer case, detailing the product or service that is the subject of the dispute. It explains the nature of the grievance, such as defects, deficiencies, service failure, and the damage or inconvenience caused. It also highlights any grievance mechanisms the consumer has availed of, such as complaints to the retailer, manufacturer, or service provider, and whether any resolutions were offered. Additionally, it outlines the claims of the opposite party, whether they accept or deny responsibility. The core legal issue in the dispute is also briefly stated.

- **Sector:** This section classifies the complaint under a specific consumer protection sector based on a predefined list with sector codes. The sector identification ensures that the grievance falls under the relevant regulatory framework for appropriate adjudication. Each sector is also associated with its corresponding sector code. For example, the banking and financial services sector has a sector code of 101.

- **Issues:** This section lists the factual claims made by the complainant and the counterarguments of the opposite party. It identifies key questions, such as whether the product/service was defective, whether consumer rights were violated, and whether compensation is justified.

- **Evidence:** This section is divided into two sections and categorizes the evidence presented by both parties. The complainant may provide receipts, contracts, images, videos, or communication records, while the opposite party may submit warranty details, service reports, or policy documents.

- **Reliefs:** This section enumerates the specific remedies sought by the complainant, such as refunds, replacements, compensation for damages, or legal costs.

6.2.2 Evaluation Metrics

We conduct reference-based, reference-free and human evaluation of the generated summary. To assess the quality and correctness of the generated summary, we use eight evaluation metrics, which were suggested by legal domain experts. We use a 5-point Likert scale evaluation of metrics: Overview Accuracy, Oversimplification, Overview Retrieval and Issues Accuracy; and binary evaluation of metrics: Evidence Accuracy, Issue Formatting, Sector Relevance and Relief Accuracy.

The metrics used for the evaluation are:

1. **Overview Accuracy**: Measures how accurately the generated summary reflects the factual details of the original case, including dates, amounts, parties involved, and key facts. A higher score indicates greater fidelity to the original material.

2. **Overview Oversimplification**: Assesses whether key elements of the legal case, such as the product/service, issue, damages, grievance mechanisms, claims, and involved parties, are retained in the summary. A lower score indicates that important details have been omitted or excessively simplified.

3. **Overview Retrieval**: Evaluates how well the summary retrieves relevant facts from the original case. A high score signifies that all critical details are included, while a lower score suggests missing or inaccurately represented information.

4. **Sector Relevance**: Determines whether the sector name and sector code in the generated summary match those in the human-annotated material summary. The evaluation is binary (Yes/No), with "Yes" indicating a correct match and "No" indicating discrepancies.

5. **Issues (Formatting)**: Checks if the issues in the generated summary are presented in a structured manner, such as a numbered list, and whether they clearly distinguish the factual claims made by different parties. The evaluation is binary (Yes/No).

6. **Issues (Accuracy)**: Measures the correctness of the issues outlined in the summary, ensuring they align with the factual matrix of the case and logically derive from the original material. A higher score indicates more accurate and relevant issue framing.

7. **Evidence Accuracy**: Verifies that the evidence listed in the generated summary matches the evidence presented in the original case, ensuring no hallucinated or missing evidence. The evaluation is binary (Yes/No).

8. **Relief Accuracy**: Ensures that the reliefs (e.g., compensation, actions) mentioned in the summary accurately reflect those in the original case. The evaluation is binary (Yes/No).

In human evaluation, we achieve an average score of more than 4 (out of 5) on the overview accuracy, oversimplification, overview retrieval and issue accuracy metrics, and a score of more than 0.60 out of 1 on the Evidence Accuracy, Issue Formatting, Sector Relevance, and Relief Accuracy metrics, demonstrating the general effectiveness of using CoT with the `Llama-3.1-8B-Instruct`

model (Table 8 & 9). We use `gpt-4o-mini` model for the LLM-based evaluation (Table 6 & 7). The correlation result of LLM-based evaluation with human evaluation is in Table 5. Appendix 5 presents the prompts used for LLM-based evaluation. The same prompts are also meant as instructions for annotators to facilitate the evaluation process.

Metric	Spearman Correlation
Overview Accuracy	0.5105
Oversimplification	0.5181
Overview Retrieval	0.4804
Issues Accuracy	0.4282
Evidence Accuracy	0.7134
Issue Formatting	0.7886
Sector Relevance	0.8551
Relief Accuracy	0.6986

Table 5: Spearman's rank correlation coefficient of human evaluation with LLM-based evaluation using `gpt-4o-mini` model as an evaluator of the generated summaries.

7 Conclusion

LLM-based evaluators have emerged as a versatile, high-performance paradigm. They work in both reference-based and reference-free modes, and consistently outperform traditional metrics in correlation with human scores. For summarization, results show that GPT-4's assessments of coherence, relevance, and factuality align well with human judgments. In machine translation, LLM-based metrics rival or exceed the best existing metrics on WMT benchmarks. Dialogue evaluation is still nascent, but early work indicates LLMs can also judge conversational relevance and coherence effectively. We have surveyed recent advances in using large language models as automatic evaluators for NLP generation tasks. The evidence indicates that LLMs (particularly GPT-4 and similar) offer a powerful new evaluation paradigm: they can replace or augment traditional metrics and even approximate human judgments in many dimensions. Our review covered core tasks (summarization, translation, dialogue, QA), described LLM prompting and adaptation methods, and synthesized empirical results. While LLM evaluators excel in many cases, they do not replace the need for human oversight in critical applications.

Several important directions remain open. **Bias and calibration:** Ensuring that LLM evaluators are fair and unbiased across different systems and styles is critical. Future work should develop calibration techniques or benchmarks to detect and mitigate biases (e.g. content or model biases) in LLM scoring. **Robustness:** We need standardized meta-evaluation datasets (analogous to robustness benchmarks in other areas) to systematically evaluate evaluator reliability. Another important direction is **efficiency and accessibility:** developing lightweight open LLM evaluators (e.g. fine-tuned smaller models) to democratize this capability across the research community. Finally, extending evaluation methods to diverse languages and domains will ensure broad applicability.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation*.

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

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

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

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

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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

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

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

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do humans evaluate machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Kartáč, Mateusz Lango, and Ondřej Dušek. 2025. Openlgauge: An explainable metric for nlg evaluation with open-weights llms.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. 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.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization*.

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

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

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of ACL 2023*.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference-free evaluation metric for dialog generation. In *Proceedings of ACL 2020*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating nlg systems. *Computational Linguistics*, 35(4):529–558.

Tejpalsingh Siledar, Swaroop Nath, Sankara Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: LLMs for opinion summary evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12119–12134, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of ACL 2020*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS) 35*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Shijie Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems (NeurIPS) 34*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *Proceedings of EMNLP 2019*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluation with contextualized embeddings and earth mover's distance. In *Proceedings of EMNLP-IJCNLP 2019*.

Lianmin Zheng, 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.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Prompts for automatic evaluation on HAB metrics

Model Name	Overview Acc.	Oversimplification	Overview Retrieval	Issues Acc.
Llama 3.1 8B (Single Prompt)	2.65	2.29	2.02	2.06
Llama 3.1 8B + Partwise	3.53	**4.17**	2.90	3.53
Llama 3.1 8B + Partwise + CoT	**4.20**	4.03	**3.03**	**3.83**
DeepSeek 8B + Partwise + CoT	3.23	3.03	2.37	3.67
Ministral 8B + Partwise + CoT	3.57	3.87	2.70	3.67
Qwen 2.5 7B + Partwise + CoT	4.07	3.63	2.60	3.63

Table 6: Performance of summary generated by different models on a 5-point Likert scale for the metrics: Overview Accuracy, Oversimplification, Overview Retrieval and Issues Accuracy. LLM-based evaluation of generated summaries is done using `gpt-4o-mini` model.

Model Name	Evidence Acc.	Issue Formatting	Sector Relevance	Relief Acc.
Llama 3.1 8B (Single Prompt)	0.14	0.61	0.28	0.33
Llama 3.1 8B + Partwise	0.33	0.50	0.63	0.60
Llama 3.1 8B + Partwise + CoT	**0.37**	**0.67**	0.60	**0.70**
DeepSeek 8B + Partwise + CoT	0.33	0.57	**0.90**	0.27
Ministral 8B + Partwise + CoT	0.37	0.50	0.67	0.10
Qwen 2.5 7B + Partwise + CoT	0.13	0.33	0.77	0.63

Table 7: Performance of summaries generated by different models on the binary metrics: Evidence Accuracy, Issue Formatting, Sector Relevance and Relief Accuracy. LLM-based evaluation of generated summaries is done using `gpt-4o-mini` model.

Model Name	Overview Acc.	Oversimplification	Overview Retrieval	Issues Acc.
Llama 3.1 8B (Single Prompt)	3.11	3.23	2.94	2.76
Llama 3.1 8B + Partwise	**4.35**	4.05	**3.25**	3.00
Llama 3.1 8B + Partwise + CoT	4.25	**4.19**	3.14	**3.50**
Deepseek 8B + Partwise + CoT	3.30	3.35	2.71	3.43
Ministral 8B + Partwise + CoT	4.15	3.95	3.05	3.25
Qwen 2.5 7B + Partwise + CoT	4.25	4.10	3.00	3.20

Table 8: Human evaluation of summaries generated by different models with Chain-of-Thought (CoT) prompting. Evaluated on a 5-point Likert scale for Overview Accuracy, Oversimplification, Overview Retrieval, and Issues Accuracy.

Model Name	Evidence Acc.	Issue Formatting	Sector Relevance	Relief Acc.
Llama 3.1 8B (Single Prompt)	0.16	0.75	0.33	0.33
Llama 3.1 8B + Partwise	0.50	**0.80**	0.55	0.73
Llama 3.1 8B + Partwise + CoT	**0.67**	0.67	0.60	0.75
DeepSeek 8B + Partwise + CoT	**0.67**	0.71	**0.95**	0.48
Ministral 8B + Partwise + CoT	0.50	0.70	0.70	0.10
Qwen 2.5 7B + Partwise + CoT	0.40	0.75	0.70	**0.85**

Table 9: Human evaluation of summaries generated by different models with Chain-of-Thought (CoT) prompting. Evaluated on the binary metrics: Evidence Accuracy, Issue Formatting, Sector Relevance, and Relief Accuracy.

```
Task Description: You will evaluate a conversation between a user and a Consumer
Grievance Chatbot. Your task is to assess how helpful the chatbot was in
assisting the user with their issue or query. Helpfulness refers to the
chatbot's ability to understand the user's problem and provide an actionable,
relevant, and clear resolution or guidance.

Evaluation Criteria:
The task is to judge the extent to which the metric is followed by the
conversation.
Following are the scores and the evaluation criteria according to which scores
must be assigned.
<score>1</score> - The chatbot's response was irrelevant or not helpful at all in
resolving the issue.
<score>2</score> - The chatbot provided only partial assistance and left out
important details.
<score>3</score> - The chatbot gave some helpful information, but it was not
enough to resolve the issue entirely.
<score>4</score> - The chatbot mostly resolved the issue, but some minor
additional guidance was needed.
<score>5</score> - The chatbot fully resolved the issue or provided clear steps
for resolution.

Instructions: Please assign a score strictly based on the evaluation criteria.
Provide a detailed explanation justifying the score. The score must be
presented within <score></score> tags only.

Example of response format:
1. Detailed explanation of evaluation.
2. Final score: Score- <score>[1-5]</score>

{conversation}
```

Figure 2: Prompt used for the evaluation of the *Helpfulness* metric.

```
Task Description: You will evaluate the accuracy of the responses provided by a
legal chatbot in a conversation with a user. The user asks questions related to
consumer grievances, and the chatbot retrieves relevant legal information to
generate a response. Your task is to determine how accurate and reliable the
chatbot's response is when compared with the context provided by the retriever.
Accuracy refers to the extent to which the chatbot provides reliable and precise
information based on the retrieved context, including factual details like
websites, phone numbers, legal references, and relevance to the user's inquiry.

Evaluation Criteria:
The task is to judge the extent to which the metric is followed.
Following are the scores and the evaluation criteria according to which scores
must be assigned.
<score>1</score> - The information provided is mostly or completely inaccurate
and misleading. The response does not align with the retrieved context.
<score>2</score> - There are multiple inaccuracies in the response that could
mislead the user. The response poorly reflects the context.
<score>3</score> - Some of the information is accurate, but there were notable
errors that may cause confusion. The response only partially reflects the
context.
<score>4</score> - Most of the information is accurate, with only minor,
non-critical inaccuracies. The response largely reflects the context.
<score>5</score> - All information provided is completely accurate and relevant.
The response aligns perfectly with the retrieved context.

Instructions: Please assign a score strictly based on the evaluation criteria.
Provide a detailed explanation justifying the score. The score must be
presented within <score></score> tags only.

Example of response format:
1. Detailed explanation of the evaluation.
2. Final score: Score- <score>[1-5]</score>.

{conversation}
{context}
```

Figure 3: Prompt used for the evaluation of the *Accuracy* metric. We provide the conversation and context to the LLM for the evaluation.

```
Task Description: Evaluate a conversation between a user and a Consumer
Grievance Chatbot, focusing strictly on the brevity of the chatbot's responses.
Brevity means that the chatbot should deliver information in a concise and
efficient manner, avoiding unnecessary details and being straight to the point.
Give low score if the bot asks too many questions.

Evaluation Criteria:
<score>1</score> - The chatbot's response was extremely verbose, providing
excessive information that overwhelmed the user or made the conversation hard to
follow.
<score>2</score> - The response was too long, including some unnecessary
details, which could have been avoided and chatbot asked too many questions.
<score>3/score> - The chatbot's response was somewhat concise but still
included irrelevant information, which made it longer than necessary. The
chatbot asked many questions before giving the resolution.
<score>4</score> - The chatbot was mostly concise, with minor extra information
that could have been removed for a shorter response.
<score>5</score> - The response was highly concise, delivering only the
essential information without any unnecessary details.

Instructions: Please assign a score strictly based on the evaluation criteria.
Provide a detailed explanation justifying the score. The score must be
presented within <score></score> tags only.

Example of response format:
1. Detailed explanation of the evaluation.
2. Final score: Score- <score>[1-5]</score>.

{conversation}
{context}
```

Figure 4: Prompt used for the evaluation of the *Brevity* metric.

Task Description: Evaluate the accuracy of the issues presented in the generated summary by comparing it is with the ground truth of the legal case summary. Ensure that the issues align with the scope and factual details provided in the ground truth. The issues must be logically derived from the factual matrix and the claims made in the case. Inaccuracies, omissions, or misalignments should result in a lower score based on the evaluation criteria.
Ground truth summary: original
Generated Summary: generated
Evaluation Criteria: Rate the accuracy of the issues on a scale from 1 to 5:
<score>5</score>: The issues are perfectly accurate, comprehensive, and logically derived from the facts and claims. <score>4</score>: The issues are mostly accurate, with minor inconsistencies or omissions. <score>3</score>: The issues are somewhat accurate but include some significant inconsistencies or omissions. <score>2</score>: The issues are largely inaccurate or fail to align with the factual details. <score>1</score>: The issues are completely inaccurate, irrelevant, or not derived from the factual matrix. Instructions:
Instructions: 1. Assign a score strictly based on the evaluation criteria. 2. Include the score within '<score></score>' tags at the end of your response.
Response Format: Final score: Present the score in this format: '<score>[1-5]</score>'.

Figure 5: Prompt for LLM-based evaluation of Issues Accuracy metric if NyayDarpan