The Need for a Leaderboard: A Survey of LLM as a Judge in NLP

Anonymous ACL submission

Abstract

Recently, the use of large language model (LLM) as a judge gains popularity in Natural Language Processing (NLP) research. This paper reviews recent studies on LLM-as-a-judge, revealing significant efforts in developing various methods for LLM-based assessment. However, there is a lack of a common standard for meta-evaluations, and several potential risks associated with LLMs need to be acknowledged. Therefore, we recommend creating a leaderboard and offer a draft proposal to support the development and adoption of LLM-as-a-judge.

1 Introduction

Human evaluation is typically regarded as the gold standard for assessing automatically generated text, but it is both expensive and time-consuming. Therefore, automatic metrics (Papineni et al., 2002; Lin, 2004; Sellam et al., 2020) are used as proxies for human judges. Although these metrics have shown some correlation with human evaluations, they have proven to be insufficient for reliable assessment (Belz and Reiter, 2006; Novikova et al., 2017; Bubeck et al., 2023). Recently, using large language model (LLM) as a judge is gaining popularity in NLP research (Zheng et al., 2023), due to their emergent capabilities (Brown et al., 2020; Wei et al., 2022a). LLM-as-a-judge has shown promising performance; for example, GPT-4 has been found to evaluate machine translation outputs more effectively than previous metrics (Kocmi and Federmann, 2023b). However, it is crucial to conduct thorough validation to ensure its correlation with human evaluations and to recognize potential risks associated with its application.

In this paper, we survey 42 papers on LLM-as-a-judge. Our findings reveal that numerous methods have been developed to obtain assessments from LLMs and LLM-based evaluators show a strong correlation with human evaluations across most tasks. However, we identify a lack of a common standard for meta-evaluations and several potential risks associated with LLMs. Thus, we recommend establishing a leaderboard to provide a common platform for developers of LLM-as-a-judge and inform users about best practices and limitations.

We summarize our contribution as follows:

- We provide a review of different approaches on using LLMs for assessment, categorizing them into four types as shown in Figure 1.

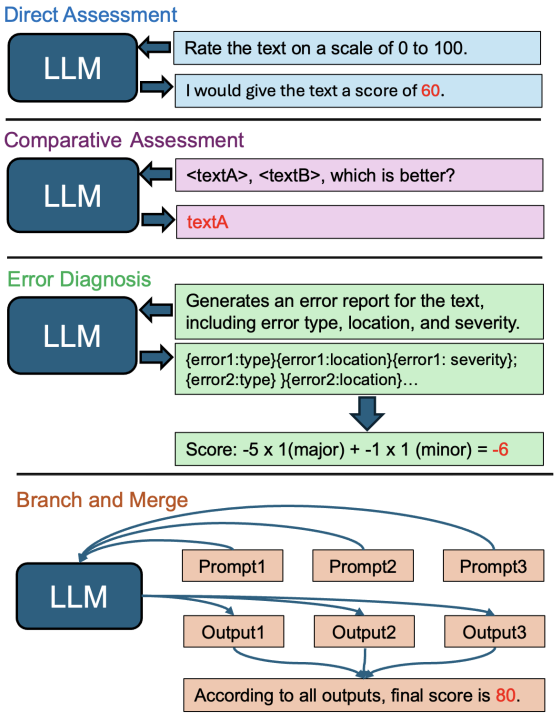

Figure 1: An illustration of four types of methods on using LLMs for assessment. Direct assessment involves asking the LLMs directly for a score. Comparative assessment requests LLMs to rank a pair of texts. Error diagnosis seeks an error analysis report from the LLMs and calculate the score based on a predefined scheme. Branch-and-merge strategies involve generating multiple prompts and then combining the outputs to determine a final score.

automatic metrics LLM	automatic metrics ChatGPT	LLM evaluator
automatic evaluation LLM	automatic metrics GPT-4	GPT Evaluator

Table 1: Keywords for identifying papers in the ACL Anthology.

- We discuss the meta-evaluations performed on LLM-as-a-judge and the potential risks associated with its use.
- We present a draft outlining the creation of a leaderboard for LLM-as-a-judge.

2 Method

Our survey includes a total of 42 papers. To identify these papers, we initially searched the ACL Anthology [1] for all relevant publications using keywords listed in Table 1, available before early June 2024. We selected papers that included meta-evaluation on LLM-as-a-judge and excluded those that solely utilized LLM-as-a-judge without meta-evaluation. Additionally, we explored the citation graph of our initial set of papers, adding any relevant papers that met our criteria. Out of the 42 papers [2], 33 are indexed by the ACL Anthology, while the rest originate from NeurIPS, ICLR, or arXiv. Once identified, we proceeded to investigate how LLMs are used for assessment, how meta-evaluations are conducted and the findings on LLM-as-a-Judge.

3 Using LLMs for Assessment

3.1 Direct Assessment

As shown in Table 2, direct assessment (DA) is the most common approach, where LLMs are prompted for a score. These prompts typically include guidelines, criteria, and few-shot examples (Chiang and Lee, 2023a,b). In addition to hand-crafted criteria, some researchers use LLMs to draft and refine the criteria (Liu et al., 2024), or to generate chain-of-thoughts (Wei et al., 2022b) as guidelines (Liu et al., 2023). Furthermore, multi-dimensional DA (Lin and Chen, 2023; Zhou et al., 2024) requires several scores for different aspects, such as grammar, and fluency.

3.2 Comparative Assessment

Comparative assessment (CA) involves comparing pairs of texts (Liusie et al., 2024; Zheng et al., 2023). It is often observed that humans find it more intuitive to compare two options rather than score each one independently, though this approach has not been extensively studied for LLM-as-a-judge.

Method	Papers
Direct assessment	36
Comparative assessment	3
Error diagnosis	5
Branch and merge	4

Table 2: The methods covered by 42 papers (some papers cover multiple methods).

3.3 Error Diagnosis

Inspired by human evaluation methodologies like Multidimensional Quality Metrics (MQM), the error diagnosis approach (Fernandes et al., 2023; Kocmi and Federmann, 2023a; Xu et al., 2023) uses LLMs to identify and label error spans by their category, location, and severity (major or minor). The overall score will then be calculated by counting the number of major and minor errors based on a predefined scheme.

3.4 Branch and Merge

To improve output consistency, Leiter et al. (2023) discussed combining outputs from multiple prompts through a majority vote. Whereas, Saha et al. (2024) employed LLMs to merge all outputs. Additionally, Chan et al. (2024) suggested having multiple LLMs debate (i.e., add responses from other LLMs in the prompt) before taking a majority vote. Zhang et al. (2023b) introduced a multi-layer LLM network where the final result is merged either by averaging or majority voting. Despite their differences, these methods fundamentally operate on a branch-and-merge principle. Besides, the prompts can be either DA or CA.

4 Meta Evaluation

4.1 Tasks

A wide variety of tasks have been explored, with a majority centered on conventional text generation tasks such as dialogue (Mendonça et al., 2023),

[1] https://www.aclweb.org/anthology/

[2] A list of all 42 papers are provided in Appendix A

Datasets and Benchmarks	Description	Papers
SummEval (Fabbri et al., 2021)	A dataset containing human annotations on generated text from 12 abstractive systems.	6
Eval4NLP 2023 Shared Task (Leiter et al., 2023)	A shared task on prompting LLMs as explainable metrics.	6
WMT (Ma et al., 2019; Freitag et al., 2022)	Human annotations on machine translations released by the Conference on Machine Translation (WMT).	5
Topical-Chat (Gopalakrishnan et al., 2019; Mehri and Eskenazi, 2020)	A dataset evaluating response quality based on dialogue history and related knowledge.	4
MT-Bench (Zheng et al., 2023)	A benchmark consisting of LLM's responses in multi-turn conversations.	2
NewsRoom (Grusky et al., 2018)	A dataset for machine summarization.	2
QAGS (Wang et al., 2020)	A benchmark for evaluating hallucinations in summarization.	2
WebNLG (Gardent et al., 2017; Castro Ferreira et al., 2020)	A benchmark for data-to-text evaluation methods.	2

Table 3: Datasets and benchmarks used by multiple papers.

summarization(Liu et al., 2023), and machine translation (Fernandes et al., 2023). However, there are also instances that use LLMs as reviewers for text written by human, such as evaluating test-taker written responses (Naismith et al., 2023) and performing paper reviewing tasks (Zhou et al., 2024).

4.2 Datasets and Benchmarks

There is considerable variation in the datasets and benchmarks employed, with only a minority of papers utilizing the same ones. Table 3 illustrates the datasets and benchmarks shared by multiple papers. Among the 42 surveyed papers, a maximum of 6 papers use any single dataset, while approximately 20 papers utilize datasets that are unique to their studies and not used elsewhere.

4.3 Correlations

To assess the correlation between LLMs' assessments and human judgments, commonly used methods include Pearson (r), Spearman (ρ), and Kendall-Tau (τ) correlations for direct assessment, and accuracy (the frequency with which the rankings match) for comparative assessment. Some studies employ alternative approaches; for instance, in one study (Huang et al., 2024), it is treated as a classification task, where assessments are categorized into tiers based on scores, and the performance of LLMs in classification is measured.

4.4 Results

Most studies report that LLM-as-a-judge achieves strong correlations with human assessments and surpasses state-of-the-art methods (Liu et al., 2023; Ferron et al., 2023). However, there are cases where no significant correlation is found, such as factuality evaluation (Fu et al., 2023) or grading math questions (Zheng et al., 2023). In paper reviewing task, it has been shown that LLM-based evaluators struggle with processing long papers and frequently make mistakes(Zhou et al., 2024). Additionally, LLMs have difficulty comparing candidates with similar performance and become less reliable when evaluating higher-quality summaries in summarization tasks (Shen et al., 2023).

4.5 Interpretability

Interpretability is recognized as a advantage of LLM-as-a-judge, as it enables the request for explanations (Zheng et al., 2023). Several studies have examined explanations for assessments. For instance, Naismith et al. (2023) discovered that LLMs can provide coherent rationales, whereas Zhou et al. (2024) suggested caution is needed as mistakes are frequently found. Moreover, the method of deriving assessments through error diagnosis also emphasizes interpretability by requesting error reports instead of scores.

5 Potential Risks

5.1 Biases

Zheng et al. (2023) investigated three types of biases, which we outline below along with other studies that support their findings.

- **Verbosity bias**: Zheng et al. (2023) discovered that LLM-based evaluators are prone to verbose text. Furthermore, Wu and Aji (2023) revealed that texts with factual errors receive higher ratings than those that are too brief.
- **Self-enhancement bias**: LLMs recognize and favor their own generations (Xu et al., 2024; Panickssery et al., 2024).
- **Position bias**: When doing comparative assessment, the judgement of LLMs can be skewed by manipulating the order of the candidate responses (Wang et al., 2023b).

5.2 Replicability

The majority (33 out of 42) of the papers use GPT-3.5/4 as the backbone, which raises concerns about replicability, as GPTs might be constantly updated.

6 Towards Building a Leaderboard for LLM-as-a-judge

As detailed in Section 4, the meta-evaluations conducted vary across the papers. This could lead to a worrisome scenario where different developers can claim a new state-of-the-art on specific datasets and settings. Thus we recommend building a leaderboard as a common ground.

6.1 Correlations

To assess correlation, we suggest using methods from the recent WMT metrics shared tasks (Freitag et al., 2023), like pairwise ranking accuracy with tie calibration (Deutsch et al., 2023) and Pearson's r. These methods have been validated through extensive testing in previous shared tasks or have been well-supported by recent studies.

6.2 Core Tasks

Chatbot Arena is an open platform facilitates anonymous comparisons between models. Users can engage with two anonymous models simultaneously by asking them the same question and voting for their preferred response. Instead of predefined questions, this approach allows for diverse use cases and gathers votes reflecting users' varied interests. Additionally, Chiang et al. (2024) have released more than 100k pairwise votes collected from this platform, enabling large-scale comparative assessments.

WMT is a key event in machine translation research. Annually, WMT releases annotations by professional translators evaluating outputs from various translation systems. These high-quality human annotations, coupled with their annual release schedule, make this task valuable and up-to-date. Furthermore, the data is extensive; for instance, WMT23 comprises more than 4,000 segments translated by over 10 systems.

SummEval includes both expert and crowd-sourced human evaluations for outputs from 16 models across 100 articles, evaluated on 4 different criteria. Humans rate the summaries based on various aspects, including coherence and consistency. It is the most utilized dataset in previous studies.

MT-Bench is a benchmark comprising 80 high-quality multi-turn questions, encompassing typical use cases and emphasizing challenging inquiries designed to distinguish between models. It includes 8 standard categories of user prompts, such as writing, math, and coding.

6.3 Challenge Sets

The challenge sets are intended to stress test the LLM-based evaluators, with a primary focus on identifying and testing biases. For the biases discussed in Section 5.1, Zheng et al. (2023) have explored applicable methods. For instance, to test position bias, they interchange the positions of two responses and measure how frequently LLMs change their judgment. Besides, it is crucial to update the challenge sets accordingly if new biases are revealed.

7 Conclusion

By surveying 42 papers on LLM-as-a-judge in NLP, we found that various methods are proposed to use LLMs for assessment; LLM-based evaluators have surpassed previous state-of-the-art methods, while still have limitations on certain tasks; and the meta-evaluations vary across the papers. To support the development and adoption of LLM-as-a-judge, we recommend creating a leaderboard and offer a draft proposal.

Limitations

It is possible that our survey missed some existing publications. Additionally, some of the papers we reviewed have not gone through peer review.

Our recommendations for creating a leaderboard are not comprehensive and further discussion is needed.

References

Pavan Baswani, Ananya Mukherjee, and Manish Shrivastava. 2023. LTRC_IIITH's 2023 submission for prompting large language models as explainable metrics task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 156–163, Bali, Indonesia. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

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.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

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

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom

Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. 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.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan Tn. 2023. Are large language models reliable judges? a study on the factuality evaluation capabilities of LLMs. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 310–316, Singapore. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019*.

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.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. 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.

Neema Kotonya, Saran Krishnasamy, Joel Tetreault, and Alejandro Jaimes. 2023. Little giants: Exploring the potential of small LLMs as evaluation metrics in summarization in the Eval4NLP 2023 shared task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 202–218, Bali, Indonesia. Association for Computational Linguistics.

Daniil Larionov, Vasiliy Viskov, George Kokush, Alexander Panchenko, and Steffen Eger. 2023. Team NLLG submission for Eval4NLP 2023 shared task: Retrieval-augmented in-context learning for NLG evaluation. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 228–234, Bali, Indonesia. Association for Computational Linguistics.

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. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

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. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. 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.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2638–2656, Torino, Italia. ELRA and ICCL.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

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

John Mendonça, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 133–143, Prague, Czech Republic. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

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.

Abhishek Pradhan and Ketan Todi. 2023. Understanding large language model based metrics for text summarization. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 149–155, Bali, Indonesia. Association for Computational Linguistics.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. *Preprint*, arXiv:2310.15123.

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.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Characterizing the confidence of large language model-based automatic evaluation metrics. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 76–89, St. Julian's, Malta. Association for Computational Linguistics.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating online human evaluation of social chatbots with prompting. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 268–281, Prague, Czechia. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. 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.

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

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

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

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

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

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Perils of self-feedback: Self-bias amplifies in large language models. *Preprint*, arXiv:2402.11436.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Rui Zhang, Fuhai Song, Hui Huang, Jinghao Yuan, Muyun Yang, and Tiejun Zhao. 2023a. HIT-MI&T lab's submission to Eval4NLP 2023 shared task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 139–148, Bali, Indonesia. Association for Computational Linguistics.

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. *Preprint*, arXiv:2308.01862.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Terry Yue Zhuo. 2024. ICE-score: Instructing large language models to evaluate code. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2232–2242, St. Julian's, Malta. Association for Computational Linguistics.

A Surveyed Papers

A.1 Indexed by ACL Anthology

Kotonya et al. (2023); Larionov et al. (2023); Zhang et al. (2023a); Liu et al. (2024); Chiang and Lee (2023b); Liusie et al. (2024); Leiter et al. (2023); Liu et al. (2023); Kocmi and Federmann (2023a); Xu et al. (2023); Ferron et al. (2023); Fernandes et al. (2023); Hada et al. (2024); Baswani et al. (2023); Yue et al. (2023); Zhuo (2024); Stureborg et al. (2024); Svikhnushina and Pu (2023); Mendonça et al. (2023); Naismith et al. (2023); Pradhan and Todi (2023); Fu et al. (2023); Kocmi and Federmann (2023b); Chiang and Lee (2023a); Chan et al. (2023); Lin and Chen (2023); Huang et al. (2024); Wang et al. (2023a); Shen et al. (2023); Zhou et al. (2024); Sottana et al. (2023); Stammbach et al. (2023); Freitag et al. (2023)

A.2 Others

Saha et al. (2024); Zhang et al. (2023b); Chan et al. (2024) (ICLR); Wu and Aji (2023); Zheng et al. (2023) (NeurIPS); Bubeck et al. (2023); Xu et al. (2024); Panickssery et al. (2024); Wang et al. (2023b)