Expert Versus Metric-Based Evaluation: Testing the Reliability of Evaluation Metrics in Large Language Models Assessment

Bartłomiej Balsamski
Krakow University of Economics
Krakow, Poland *balsamb@uek.krakow.pl*

Jakub Kanclerz
Krakow University of Economics
Krakow, Poland *kanclerj@uek.krakow.pl*

Dariusz Put
Krakow University of Economics
Krakow, Poland *putd@uek.krakow.pl*

Janusz Stal
Krakow University of Economics
Krakow, Poland *stalj@uek.krakow.pl*

Abstract

This study examines the reliability of automatic evaluation metrics in assessing responses generated by large language models (LLMs) in the context of university recruitment. A total of 113 domain-specific questions were used to prompt five prominent LLMs, each in three configurations: basic, document-context, and internet-context. The generated responses were evaluated using three categories of metrics: lexical, semantic, and LLM-as-a-Judge. These metric-based assessments were subsequently compared with expert evaluations conducted using a 5-point Likert scale. The findings indicate that although automatic metrics offer considerable efficiency, their consistency with expert judgments varies substantially. Moreover, the results suggest that both the model configuration and its underlying architecture significantly affect evaluation outcomes. Among the metric categories, LLM-as-a-Judge appears to yield the highest alignment with expert assessments, suggesting greater reliability in this approach.

Keywords: large language model, generative artificial intelligence, Bert, Rouge, LLMaj.

1. Introduction

In recent years, the rapid advancement of generative artificial intelligence (GenAI) and large language models (LLMs) has spurred extensive research into their application across various societal domains. One area with significant potential is the higher education sector, particularly in improving access to information during student recruitment. GenAI can support prospective students in obtaining detailed information about study programs, schedules, tuition fees, and required documents. However, broader adoption of AI at universities is hindered by institutional resistance, with decision-makers frequently blocking such initiatives. As a result, GenAI-based implementations in academia remain limited [15]. Another key barrier is the challenge of reliably evaluating the effectiveness of LLMs in generating accurate information.

This study addresses this gap by analyzing LLM-generated responses using automated evaluation metrics. Recruitment-related queries—reflecting the information needs of prospective students—serve as prompts for GenAI systems to produce informative answers. The study investigates how effectively these metrics assess the quality of such responses. The central research questions are:

1. How reliably do selected automatic evaluation metrics reflect expert assessments of LLM-generated responses in an educational domain?
2. What directions for improvement can be identified to enhance the consistency and applicability of evaluation methods in practice?

To address these questions, the study examines the use of GenAI in the recruitment process at the Krakow University of Economics (KUE) in Poland. The following sections present the research background, describe the methodology, and summarize the initial findings. A research plan is then outlined to evaluate the accuracy of the selected metrics through expert-based validation, followed by a discussion of the study's implications for both theory and practice.

2. Research Background

Evaluating the performance of LLMs requires the use of robust and reliable metrics that can accurately assess the correctness of the generated responses. A variety of metrics have been developed for this purpose, which can be broadly classified into two main categories: reference-based and LLM-based metrics (see Fig. 1). Reference-based metrics evaluate the similarity between a model-generated response and one or more gold-standard references—typically reliable, human-authored answers that serve as the benchmark. These metrics are commonly based on n-gram overlap techniques [14] (e.g., BLEU, ROUGE, Jensen-Shannon Divergence) or embedding-based semantic similarity measures [21] (e.g., BERTScore, MoverScore, Sentence Mover Similarity, Cosine Similarity). In contrast, LLM-based metrics involve using another (or the same) language model to assess the quality of a given response [10]. This category includes:

- Scoring models, which assign scores based on predefined criteria such as relevance, correctness, fluency, and helpfulness;
- Rubric-based evaluations, which utilize a structured set of evaluation criteria (a rubric) to guide the model in scoring a response with respect to dimensions such as accuracy, coherence, and relevance;
- Pairwise comparisons, in which an LLM evaluates two alternative responses to the same prompt and selects the one deemed superior based on overall quality or specific evaluation dimensions [10].

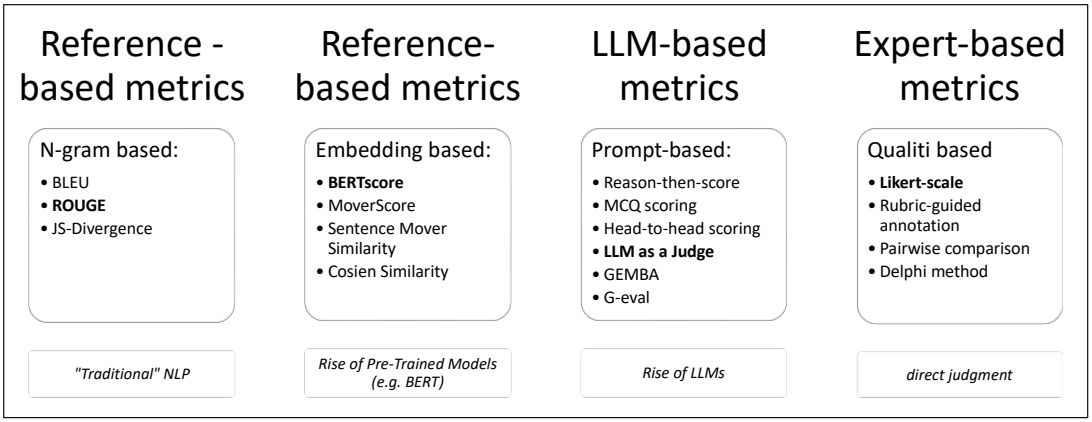

Fig. 1. Reference-based and LLM-based metrics.

Among the most widely used reference-based evaluation metrics are the n-gram-based ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [11] and the embedding-based BERTScore (BERT) [25]. ROUGE measures textual similarity by calculating the overlap of n-grams between the generated response and a reference text, providing a score that reflects lexical correspondence [5]. In contrast, BERT was introduced to address the limitations of n-gram-based metrics by leveraging contextual embeddings to evaluate

semantic similarity. It utilizes transformer-based models (such as BERT) to capture deeper linguistic meaning in the comparison between generated and reference responses [25]. Studies have shown that BERT often achieves better alignment with human judgment than traditional n-gram-based methods. However, its performance can vary depending on the specific BERT variant employed and the nature of the task. Moreover, it may still struggle in cases that involve nuanced language use or subjective interpretation [9], [26]. To further narrow the gap between automatic evaluation and human judgment, the LLM-as-a-Judge (LLMAJ) approach has recently emerged. This method involves using advanced LLMs themselves to evaluate the outputs of other models, treating the LLM as a proxy for human judgment [26].

Schroeder and Wood-Doughty [18] emphasize the importance of a nuanced understanding of the reliability of LLMs and caution against the risks of over-reliance on single-shot evaluations. Their work contributes to the development of more trustworthy and robust LLM-based systems and applications by advocating for evaluation practices grounded in internal consistency across multiple runs with varying random seeds. Empirical studies have shown that even state-of-the-art LLM-based evaluators can exhibit significant instability in response to minimal input perturbations [1]. Additionally, personalization scenarios present unique challenges to the reliability of LLM-as-a-Judge (LLMAJ) approaches [3]. Nonetheless, meta-evaluations of current evaluation methodologies underscore the need for multi-perspective assessment frameworks that combine the scalability and efficiency of automatic metrics with the interpretability, flexibility, and contextual sensitivity offered by human or LLM-based evaluators [6].

The expert-based evaluation method, which involves comparing the outputs of LLMs to answers provided by domain experts, has emerged as a valuable approach for assessing the quality and reliability of LLM-generated content. Tan et al. [22] conducted a comprehensive analysis comparing responses from ChatGPT to those produced by traditional knowledge-based question answering (KBQA) systems and expert-authored answers. Their findings illustrate both the strengths and limitations of ChatGPT across a variety of question types. Chen et al. [2] investigated the use of LLMs in reference-free evaluation settings by benchmarking ChatGPT's judgments against expert assessments. Their results highlight the increasing potential of LLMs to serve as evaluators of text quality, although some variability in consistency remains. Similarly, Chiang and Lee [3] examined the alignment between LLM-generated evaluations and expert ratings in the context of language quality assessment. They observed that, while LLMs demonstrate encouraging performance, multiple factors can influence the degree of agreement with expert judgments. These findings underscore the importance of careful calibration and methodological rigor when applying expert-based evaluation approaches to LLMs.

3. Research Methods

In the present study (see Fig. 2), the domain of analysis was defined with a specific focus on the university recruitment process. Relevant documents and data were collected, and in consultation with domain experts, a list of the most frequently asked questions with corresponding reference answers was compiled. Subsequently, appropriate language models, fine-tuning approaches, and evaluation metrics were selected to assess the quality of the model-generated responses. The study was conducted, and the most salient findings were analyzed.

Three principal categories of evaluation metrics were used to assess the correctness of responses generated by LLMs: lexical (n-gram-based), semantic, and LLM-as-a-Judge (LLMAJ). ROUGE represented the lexical category, measuring surface-level overlap between generated and reference texts. BERT was selected for the semantic category, leveraging contextual embeddings to assess similarity. LLMAJ involved an LLM autonomously evaluating generated responses relative to reference answers across dimensions such as fluency, coherence, consistency, and relevance [23].

In the next phase, responses to 113 domain-specific questions were generated using

five large language models: (1) OpenAI GPT-4o, (2) LLaMA (Llama3.1:8b), (3) DeepSeek (DeepSeek-r1:14b), (4) Claude (Claude-3-5-sonnet-20241022), and (5) BIELIK (Bielik-11B-v2.2-Instruct). Model descriptions and configurations are provided in Table 1. All selected models were designed for Natural Language Processing (NLP) tasks, including data and document analysis, enabling them to generate domain-specific responses. Each of the 113 questions had expert-provided reference answers. The correctness of model-generated responses was evaluated using the described metrics, with direct comparison to the reference answers.

Table 1. Description of selected Large Language Models.

Name	Release Date	Publisher	License Type	Description
OpenAI (GPT-4o)	May 2024	OpenAI	Proprietary	large-scale, based on deep learning architecture, utilizing transformers for processing and generating text with prediction the most likely subsequent token in a sequence based on contextual embeddings, highly effective in applications of conversational artificial intelligence and adaptive reasoning [8], [12]
Llama (Llama3.1: 8b)	February 2023	Meta	Meta Llama 3.1 Community License (commercial)	based on a multi-layered transformer architecture with self-attention mechanisms, using subword tokenization with the *tiktoken* Byte Pair Encoding (BPE) tokenizer, trained on a diverse and extensive corpus of texts sourced from various domains, resulting in a broad understanding of languages and their contextual usage [17], [19], used across a range of fields, including the social sciences
DeepSeek (DeepSeek-r1:14b)	January 2025	DeepSeek	MIT	for reasoning tasks, including mathematics and coding, fine-tuned using additional reinforcement learning strategies to enhance logical consistency and structured problem solving [8]
Claude (Claude-3-5-sonnet-20241022)	January 2025	Anthropic	Proprietary	for structured reasoning and alignment with human values through reinforcement learning with human feedback (RLHF), designed to optimize context retention, coherence in multi-turn dialogues, and domain-specific adaptability [12], by providing structured and precise responses particularly significant in the social sciences
BIELIK (Bielik-11B-v2.2-Instruct)	April 2024	SpeakLeash & ACK Cyfronet AGH	Apache 2.0	based on the Transformer architecture, built upon the Mistral 7B v0.1 model serving as its extension, training process conducted primarily using test datasets in the Polish language enabling optimization specifically for this linguistic context [13]

In the final stage, aimed at validating the effectiveness of the evaluation metrics, an expert-based methodology was employed. Commonly used in social sciences and psychology—especially where reliable empirical data are lacking or conventional methods fall short—this approach is regarded as highly suitable for analyzing complex or nuanced information [7]. It allows for the elicitation of expert judgments on the phenomenon under study and supports the development of evidence-based recommendations through statistical analysis of expert responses [16].

An essential task within this methodology may also include the aggregation and weighting of diverse information sources [24]. In this study, a panel of four domain experts independently evaluated the LLM-generated responses by comparing them to reference answers. Using a 5-point Likert scale - where 5 indicated the highest agreement - they assigned ratings, which formed the basis for further analysis.

4. Research Results

Krakow University of Economics (KUE), a leading public university in Poland with a focus on business education, aimed to improve communication with prospective students by implementing an AI-powered chatbot. The initial solution was a rule-based system operating on a fixed set of predefined responses [20].

Table 2. Similarity values of LLMs responses with reference answers.

Model	ROUGE mean	BERT mean	LLMAJ mean	ROUGE std	BERT std	LLMAJ std
Bielik (CAA)	0,09	0,65	0,48	0,03	0,02	0,10
Bielik (BA)	0,19	0,70	0,60	0,07	0,03	0,09
Bielik (ISCA)	0,31	0,75	0,66	0,22	0,08	0,16
Claude (CAA)	**0,41**	**0,78**	0,66	0,25	0,09	0,23
Claude (BA)	0,28	0,74	**0,76**	0,25	0,07	0,13
Claude (ISCA)	0,30	0,73	0,58	0,26	0,09	0,24
DeepSeek (CAA)	0.06	0.60	0.30	0.08	0.03	0.11
DeepSeek (BA)	0,03	0,57	0,42	0,02	0,02	0,15
DeepSeek (ISCA)	0,07	0,60	0,48	0,07	0,04	0,17
Llama (CAA)	0,11	0,66	0,42	0,04	0,03	0,11
Llama (BA)	0,13	0,67	0,46	0,12	0,05	0,10
Llama (ISCA)	0,21	0,68	0,48	0,20	0,09	0,33
OpenAI (CAA)	**0,47**	**0,81**	**0,70**	0,33	0,12	0,27
OpenAI (BA)	0,13	0,68	0,58	0,07	0,03	0,15
OpenAI (ISCA)	0,31	0,74	0,56	0,35	0,13	0,32

In a related study, researchers evaluated the potential of replacing this rule-based chatbot with LLMs. Five distinct LLMs were tested under three response generation scenarios: (1) plain responses without additional context, (2) responses generated using context from curated university documents, and (3) responses using dynamically retrieved web-based context via information retrieval. The generated responses were assessed using a combination of lexical and semantic evaluation metrics, along with an independent assessment performed by a separate LLM-based evaluator. The evaluation was based on 113 unique questions, provided by the administrator of the original rule-based system. With five models and three variants per model, the experiment produced 1,695 responses. Accuracy and relevance were measured against reference answers using ROUGE (lexical), BERT (semantic), and a dedicated LLM-as-a-Judge (LLMAJ) metric. Together, these formed a comprehensive baseline for the comparative analysis (see Table 2).

To validate the reliability of the semantic evaluation methods - ROUGE, BERT, and LLMAJ - an expert-based assessment was conducted. One representative model was selected from each LLM: Bielik (ISCA), Claude (CAA), DeepSeek (ISCA), LLaMA (ISCA), and OpenAI (CAA), based on the highest average score across all semantic metrics (see Table 2). Notably, no models from the BA group were selected, suggesting they received the lowest scores in the prior evaluation phase.

Each model was then assessed by a panel of three experts who, with access to the reference answers, independently rated each response on a 5-point Likert scale (1 = completely unacceptable, 5 = fully acceptable). Experts agreed on a common set of evaluation criteria to ensure consistency. The median score from the three expert ratings was used as the final value for each response, resulting in a unique rating per question. For comparison across models, total scores were calculated, and the percentage of responses matching the performance of an ideal model (all answers rated 5) was determined. A summary of these results is presented in Table 3.

As the final step of the research, to validate automated evaluation metrics against

human expert judgment, we conducted Wilcoxon signed-rank tests comparing four automated metrics with expert scores. The test results revealed significant differences between all automated metrics and expert evaluations: ROUGE-L showed the largest discrepancy (statistic = 1000.5, p = 0), followed by BERTScore (statistic = 41603.0, p = 0). LLM-score achieved the closest alignment with expert judgment (statistic = 26852.5, p = 0.0109), demonstrating superior performance compared to lexical and embedding-based approaches. However, the persistent statistical significance across all metrics underscores that automated evaluation cannot fully replace human expert assessment in comprehensive model performance evaluation.

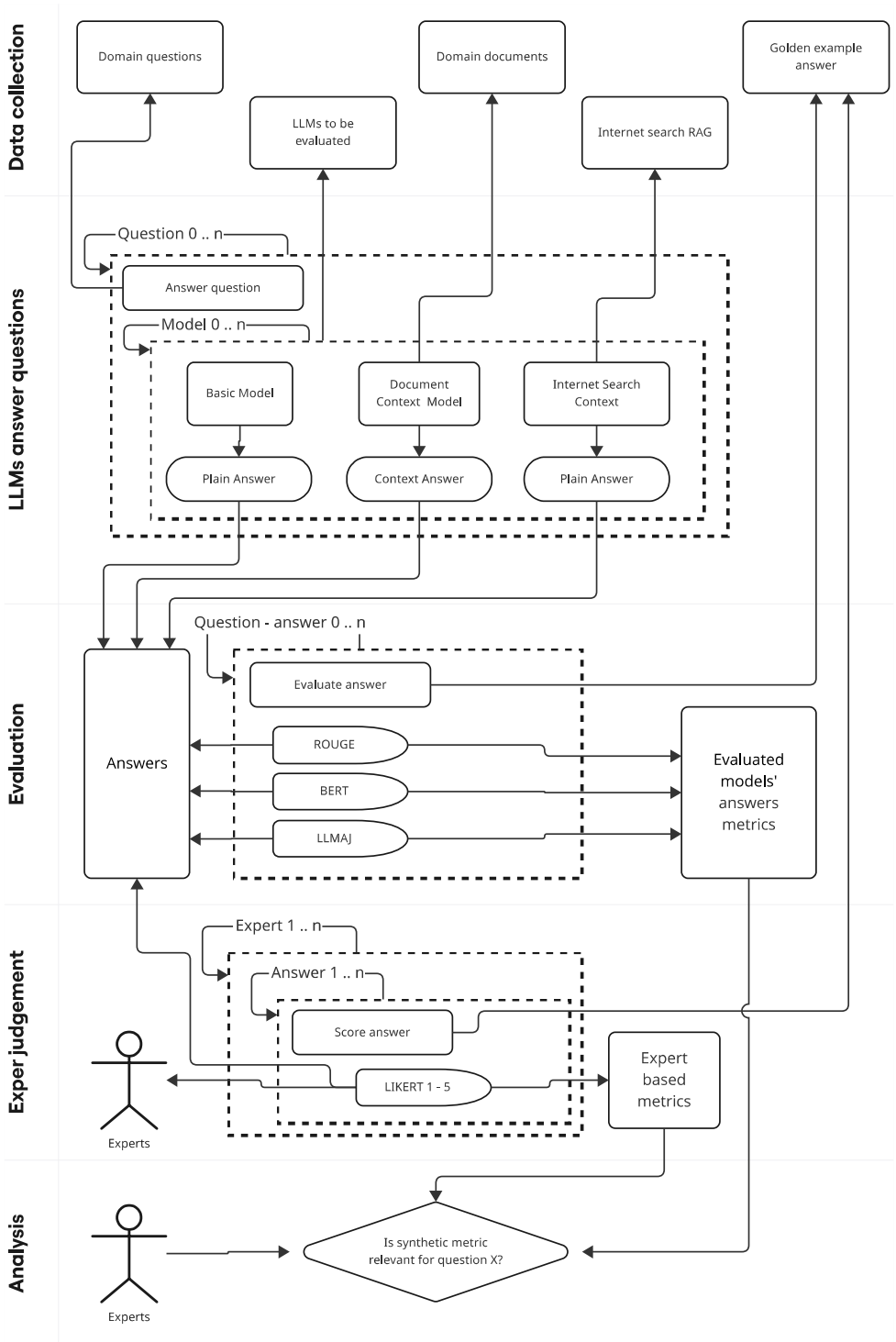

Fig. 2. A structured pipeline for evaluating LLMs.

Table 3. Expert assessment of LLMs selected according to the highest average of semantic metrics.

Model selected for expert evaluation	Outcome of expert evaluation	Semantic method assessment closest to expert evaluation
DeepSeek (ISCA)	0.573	0.60 (BERT mean)
Bielik (ISCA)	0.598	0.66 (LLMAJ mean)
OpenAI (CAA)	0.535	0.47 (ROUGE mean)
Claude (CAA)	0.598	0.66 (LLMAJ mean)
Llama (ISCA)	0.472	0.48 (LLMAJ mean)

5. Discussion and Conclusion

A comparison of the data presented in Table 2 and Table 3 - containing model evaluation scores obtained through semantic metrics and expert assessment, respectively - reveals that expert ratings diverge from those produced by statistical evaluation methods in the majority of cases. The LLMAJ-mean metric yielded scores most closely aligned with expert assessments in three instances: for the Bielik, Claude, and LLaMA models. In contrast, for the OpenAI model, the score obtained using the ROUGE-mean metric was the closest to the expert evaluation, while for DeepSeek, the BERT-mean metric exhibited the highest degree of similarity. In two cases—Bielik and Claude—the expert assessments were identical, and in these same cases, the LLMAJ-mean evaluations were also equal to one another, though slightly higher than the expert scores (0.66 versus 0.598). Based on these findings, three key conclusions can be drawn: (1) The quality assessment of semantic evaluation methods does not yield unequivocal results, (2) The LLMAJ-mean method demonstrated the highest overall alignment with expert evaluations, and (3) Metric variants incorporating standard deviation (std) did not, in any instance, provide the closest approximation to expert judgments.

The OpenAI CAA model presents an interesting case. In 62 out of 113 instances, the model responded with “I don’t know.” According to expert evaluations, the model received a score of 1 in 64 cases and a score of 5 in 45 cases. Intermediate ratings (i.e., scores of 2 to 4) were assigned in only 4 cases. These results suggest a high degree of reliability: the model either provides a fully correct response or explicitly indicates uncertainty when it lacks sufficient information. This behavior implies a relatively low occurrence of hallucinations—incorrect or fabricated content—which is a desirable characteristic for systems intended for high-stakes or factual applications. However, the model's inability to answer in 56.6% of cases also highlights a limitation. While its conservative response strategy reduces the risk of misinformation, it also indicates a need for further tuning and optimization to improve its practical usefulness in real-world applications where informative responses are expected.

As for the impact on theory, the study shows that while automatic metrics such as ROUGE, BERT, and LLMAJ are efficient, their alignment with expert evaluations varies. LLMAJ-mean proved most consistent with expert judgments, highlighting its potential. However, the results point to the need for hybrid evaluation approaches and further validation to improve reliability in different contexts. From a practical perspective, expert input remains crucial in sensitive areas like university recruitment. Some models, especially OpenAI’s, responsibly handled uncertainty, but still require tuning to improve usefulness. Automatic metrics alone are not sufficient for deployment decisions—evaluation methods should be adapted to the specific application and supported by expert oversight. Hence, the subsequent part of the study, aimed at addressing research question 2, will consist of the following stages:

1. Extension of Expert Evaluation and Statistical Validation. This phase will expand the expert assessment to a larger sample of responses and models. Inter-rater reliability (e.g., Kendall’s W, Krippendorff’s alpha) will be calculated to validate the consistency of expert ratings and confirm the robustness of previous findings.
2. Design of a Hybrid Evaluation Approach. A hybrid evaluation method will be developed by combining automatic metrics (e.g., ROUGE, BERT, LLMAJ) with

expert input. The goal is to improve reliability through weighted scoring or adaptive calibration using selected expert-annotated examples.
3. Practical Validation and Usability Testing. The proposed evaluation approach will be tested in a real-world university setting (e.g., GenAI chatbot). The focus will be on assessing its effectiveness, usability, and efficiency in supporting recruitment-related communication.
4. The performed statistical test confirms that using only automated approaches cannot fully replace comprehensive human expert assessment.

Acknowledgements

The article presents the result of the Project no 020/ZIS/2024/POT financed from the subsidy granted to the Krakow University of Economics.

References

1. Chen, G. H., Chen, S., Liu, Z., Jiang, F., Wang, B.: Humans or LLMs as the Judge? A Study on Judgement Bias. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 8301-8327 (2024)
2. Chen, Y., Wang, R., Jiang, H., Shi, S., Xu, R.: Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. Findings of the Association for Computational Linguistics: IJCNLP-AACL, pp. 361-374 (2023)
3. Chiang, C.H., Lee, H.: A Closer Look into Using Large Language Models for Automatic Evaluation. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 8928–8942 (2023)
4. Dong, Y.R., Hu, T.M., Collier, N.: Can LLM be a Personalized Judge? Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 10126-10141 (2024)
5. Fangkai, Y., Pu, Z., Zezhong, W., Lu, W., Jue, Z., Mohit, G., Qingwei, L., Saravan, R., Dongmei, Z.: Empower Large Language Model to Perform Better on Industrial Domain-Specific Question Answering. https://arxiv.org/abs/2305.11541 (2023)
6. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., Guo, J.,: A Survey on LLM-as-a-Judge. https://arxiv.org/abs/2411.15594 (2024)
7. Iriste, A.I., Katane, I.: The Use of Expert Evaluation Method in Social Science Research. Baltic Journal of European Studies, 9(1), pp. 78-97 (2019)
8. Jiang, Q., Gao, Z., Karniadakis, G.E.: DeepSeek vs. ChatGPT vs. Claude: A Comparative Study for Scientific Computing and Scientific Machine Learning Tasks. Theoretical and Applied Mechanics Letters, 15(3), 100583 (2025)
9. Laskar, M.T.R., et al.: A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 13785-13816 (2024)
10. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. https://arxiv.org/abs/2412.05579 (2024)
11. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, Association for Computational Linguistics, pp. 74-81 (2004)
12. Nasirov, R.: The Role of Claude 3.5 Sonet and ChatGPT-4 in Posterior Cervical Fusion Patient Guidance. World Neurosurg, 197:123889 (2025)
13. Ociepa, K., Flis, Ł., Wróbel, K., Gwoździej, A., Kinas, R.: Bielik 7B v0.1: A Polish Language Model – Development, Insights, and Evaluation. https://arxiv.org/abs/2410.18565 (2024)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318 (2002)
15. Pit, P., Linden, T., Mendoza, A.: Generative Artificial Intelligence in Higher Education:

One Year Later. Americas Conference on Information Systems (AMCIS), 11 https://aisel.aisnet.org/amcis2024/is_education/is_education/11 (2024)
16. Poliakova, Y., Novosad, Z.: Application of the Expert Evaluation Method in the Analysis of Trends and Priorities of the Educational Process. Zeszyty Naukowe Wyższej Szkoły Bankowej w Poznaniu, 95(4), pp. 73-81 (2021)
17. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: LLMs in e-commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation. Natural Language Processing Journal, 6(1):100056 (2024)
18. Schroeder, K., Wood-Doughty, Z.: Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. https://arxiv.org/abs/2412.12509v2 (2024)
19. Sharma, B., Ghawaly, J., McCleary, K., Webb, A.M., Baggili, I.: ForensicLLM: A Local Large Language Model for Digital Forensics. Forensic Science International: Digital Investigation, 52:301872 (2025)
20. Singh, J., Joesph, M.H., Jabbar, K.A.: Rule-based Chatbot for Student Enquiries. Journal of Physics: Conference Series, 1228(1):012060. IOP Publishing (2019)
21. Slimani, T.: Description and Evaluation of Semantic Similarity Measures Approaches, International Journal of Computer Applications, 80(10), pp. 25-33 (2013)
22. Tan, Y., Min, D., Li, Y., Li, W., Xue, N., Chen, Y., Qi, G.: Can ChatGPT Replace Traditional KBQA Models? An In-depth Analysis of the Question Answering Performance of the GPT LLM Family. The Semantic Web – ISWC 2023, pp. 348-367 (2023)
23. van Schaik, T.A.: A List of Metrics for Evaluating LLM-generated Content. https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics (2024)
24. von Soest, C.: Why Do We Speak to Experts? Reviving the Strength of the Expert Interview Method. Perspectives on Politics 21(1), pp. 1-11 (2022)
25. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, T.: BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations (ICLR) (2020)
26. Zheng, L., et al.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. 37th Conference on Neural Information Processing Systems (NeurIPS) (2023)