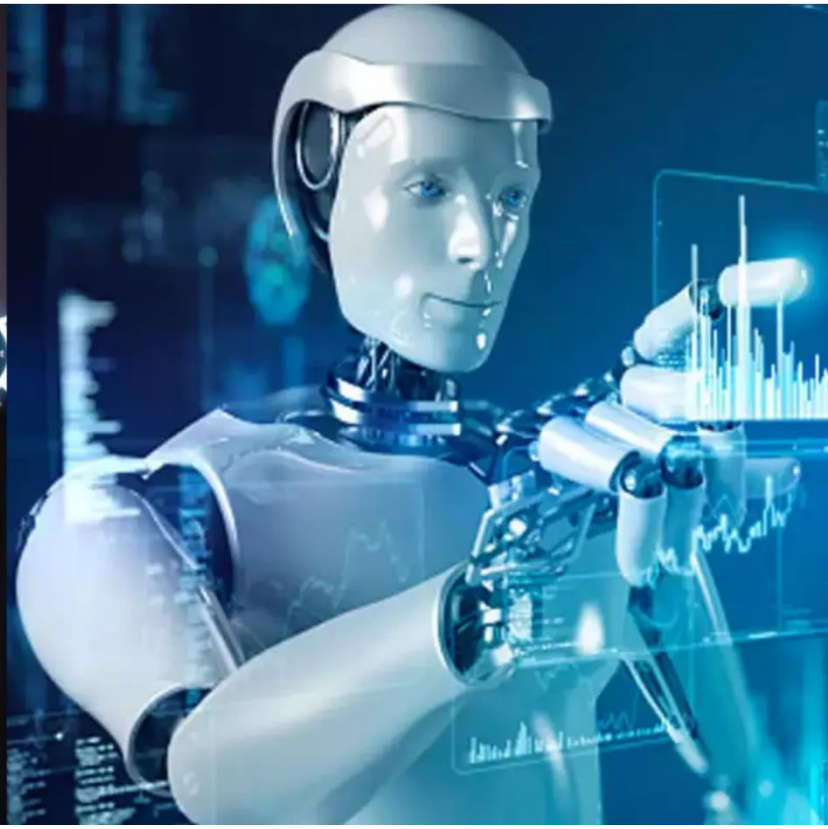
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406011|

LLM Evaluations: A Survey of Programmatic, Human, and LLM-as-Judge Approaches

Sarath Babu Poovassery Krishnan

Senior Solutions Architect, Amazon Web Services, New York, USA

ABSTRACT: The rapid progress of large language models (LLMs) has created an urgent need for robust, multi-faceted evaluation methods that address both objective performance and subjective qualities. This survey reviews state-of-the-art developments from 2023–2025 in LLM evaluation, focusing on three paradigms: (1) programmatic (automated) evaluation, (2) human evaluation, and (3) LLM-as-a-judge approaches. We analyze the advantages, limitations, and practical trade-offs of each method, and highlight recent innovations in benchmarking, human feedback, and scalable AI-assisted evaluation. Our review concludes that comprehensive LLM assessment demands integrated strategies balancing automation, human judgment, and model-based augmentation, supported by dynamic, transparent, and community-driven practices.

KEYWORDS: Large Language Models (LLMs); Evaluation; Benchmarks; Human Evaluation; Automated Metrics; LLM-as-a-Judge; Model Assessment.

I. INTRODUCTION

Large language models (LLMs) are transforming the landscape of artificial intelligence, enabling applications from conversational agents and creative writing to scientific research and code generation. The surge in their deployment across industry and academia has heightened the need for reliable and holistic evaluation frameworks. Traditional NLP metrics and benchmarks—once sufficient for task-oriented models—are increasingly inadequate for capturing the full spectrum of LLM capabilities, especially as models surpass human baselines on many tasks [1,2]. Furthermore, the open-ended nature of LLM outputs, their potential for hallucinations, bias, and harmful content, and their integration into safety-critical systems present new challenges for both technical and ethical evaluation [2,3]. The community has responded by developing and combining three principal paradigms: (1) programmatic or automated evaluation, (2) human-centered evaluation, and (3) LLM-as-a-judge, in which LLMs are themselves used as evaluators. This survey synthesizes recent progress in these paradigms, identifies their trade-offs, and recommends emerging best practices for robust, trustworthy LLM assessment.

II. LITERATURE SURVEY

2.1 From Classical Metrics to Multidimensional Benchmarks

Early evaluation of language models relied heavily on well-established NLP benchmarks. Datasets such as GLUE, SuperGLUE, and SQuAD enabled progress in language understanding and question answering, with metrics like accuracy, F1, BLEU, and ROUGE serving as standard measures [3,4]. However, as LLMs have advanced, these benchmarks have saturated: models achieve or exceed human performance on many leaderboards, diminishing their discriminatory power for state-of-the-art research [2,5].

This saturation prompted the development of broader and more challenging benchmarks, notably multitask and reasoning-focused datasets such as MMLU (Massive Multitask Language Understanding), BIG-bench, and C-Eval. These encompass a wide array of tasks—from mathematics and logical reasoning to law, medicine, and world knowledge—requiring deeper generalization and understanding [6,7]. Simultaneously, the rise of generative models spotlighted the limitations of reference-based metrics such as BLEU and ROUGE. These metrics, based on n-gram overlap with reference answers, often fail to reward the creative, semantically correct, or contextually relevant outputs that LLMs produce [8,9].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406011|

To address these gaps, new evaluation dimensions have emerged. Embedding-based metrics (e.g., BERTScore) assess semantic similarity rather than surface overlap [10]. Execution-based evaluation has gained popularity for code generation: solutions are judged by passing unit tests or running in sandbox environments [11]. Adversarial and dynamic benchmarks—like Dynabench—continually evolve, challenging models with newly generated or human-crafted hard examples to minimize overfitting [12]. Frameworks like HELM now aggregate multiple metrics (robustness, bias, calibration, toxicity, etc.), supporting holistic model profiles [13].

2.2 Human-Centered and Safety-Focused Evaluation

Automated metrics alone, however, have proven insufficient for evaluating qualities such as safety, factuality, user satisfaction, and ethical alignment. Numerous studies document that automatic NLG metrics correlate only weakly with human judgment, especially for open-ended tasks [8,14]. As a result, human evaluation—whether through expert review, crowd-sourced preference studies, or targeted user testing—has become a necessary complement [2,15]. Standardized rubrics now assess helpfulness, harmlessness, relevance, and honesty, with best practices emphasizing clear guidelines, diverse annotators, and reporting inter-rater agreement [2,15].

Safety and alignment have become central concerns in the wake of LLM deployment in real-world systems. New adversarial and red-teaming datasets probe models for harmful, biased, or manipulative outputs [2,16]. “TruthfulQA” and “SafetyBench” exemplify specialized evaluations focusing on factuality and safety, respectively [16]. The growing literature also highlights the need for ongoing, continuous evaluation as models and data distributions evolve, with recommendations for live monitoring, user feedback integration, and transparency in reporting [5,13].

2.3 LLM-as-a-Judge and AI-Augmented Evaluation

The past two years have seen the emergence of LLM-as-a-judge: using strong language models themselves to evaluate the outputs of other models or systems [1,17]. This paradigm leverages LLMs’ ability to perform complex reasoning and natural language evaluation tasks, enabling scalable pairwise ranking, critique generation, and rubric-based scoring. Prominent leaderboards and automated evaluators use this approach to efficiently compare many model variants at low cost [1,17].

Recent empirical work shows that LLM-based judges, with careful prompt engineering and calibration, can approach or even match the reliability of aggregated human preference ratings (80–90% agreement in some studies) [1,17]. However, the literature notes that LLM judges can exhibit systematic biases—such as preferring longer outputs or mimicking their own stylistic tendencies—and require ongoing validation and bias mitigation [1,18]. Hybrid pipelines, in which LLM judges are used for triage or large-scale screening and humans for final validation, are gaining traction as scalable solutions [17].

III. METHODOLOGY

LLM evaluation now commonly integrates three paradigms:

3.1 Programmatic (Automated) Evaluation

Programmatic evaluation uses datasets and computational metrics to measure specific capabilities. Typical benchmarks include tasks for question answering, reasoning, code generation, summarization, and multilingual understanding [3,6,7,11]. Metrics range from accuracy and F1 for classification or QA, to BLEU/ROUGE for generation, to embedding-based similarity (BERTScore) for semantic fidelity [4,8,10]. Execution- or simulation-based tests are used for code and tool-use tasks [11]. These approaches provide scale, objectivity, and reproducibility, but can miss nuanced, subjective, or safety-related aspects, and are susceptible to overfitting or memorization [2,5].

3.2 Human Evaluation

Human-centered evaluation leverages expert review, crowd-sourced ratings, and user studies to assess qualities such as factuality, relevance, style, and user experience [2,8,15]. Pairwise ranking (A/B testing), rubric-based scoring

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406011|

(helpfulness, harmlessness, honesty, etc.), and targeted red-teaming are prevalent methods. Best practices include recruiting diverse annotators, clear rubrics, multiple ratings per item, and reporting inter-annotator agreement [2,15]. Human evaluation is essential for open-ended and safety-critical assessment, but is resource-intensive and variable.

3.3 LLM-as-a-Judge Evaluation

LLM-as-a-judge evaluation uses a language model as the evaluator: the model is prompted to compare, score, or critique outputs from other models, either for pairwise ranking or rubric-based scoring [1,17]. This paradigm supports efficient large-scale comparison, especially where human annotation is impractical. Studies show high agreement with human judgments, though careful prompt engineering and ongoing validation are required to control for judge bias [1,17,18]. This approach is particularly useful for triage and automated benchmarking, with humans often retained for final or high-stakes validation.

IV. COMPARATIVE ANALYSIS

Table 1. Comparative Analysis of LLM Evaluation Paradigms

Approach	Strengths	Limitations	Best Use Cases
Programmatic	Objective, scalable, reproducible; enables fast benchmarking	Misses nuance, may overfit or become stale, lacks context	Routine benchmarking, regression, QA with ground truth
Human Evaluation	Captures nuance, context, and user values; flexible	Costly, slow, variable, not scalable; risk of subjectivity	User experience, safety, creative/open-ended tasks
LLM-as-a-Judge	Scalable, efficient, high agreement with humans (when tuned)	Judge/model bias, prompt sensitivity, needs human validation	Pairwise ranking, large-scale filtering/triage

Table 1 highlights that each LLM evaluation paradigm offers distinct benefits and trade-offs. Programmatic evaluation excels at speed, consistency, and scale, making it ideal for routine benchmarking and regression testing; however, it is limited in capturing creative, nuanced, or subjective aspects of model outputs. Human evaluation remains essential for tasks requiring judgment, ethical consideration, or open-ended responses, as it reflects real-world user perspectives and values; its downsides are high cost, slower throughput, and potential inconsistency. LLM-as-a-judge evaluation strikes a balance, enabling scalable, automated assessment that can closely mirror human preferences for many tasks, yet still requires careful design to mitigate judge/model bias and should be validated against human ratings. As the table suggests, a holistic evaluation strategy—combining these approaches—yields the most robust and reliable understanding of LLM performance.

V. EMERGING TRENDS

Several key trends are reshaping the landscape of LLM evaluation. First, **dynamic and continuous evaluation** has become essential, with practitioners adopting live benchmarking pipelines and dynamic datasets to counteract overfitting and dataset leakage, and to better represent real-world data distribution shifts [12,13]. Rather than relying on static benchmarks, models are now regularly tested on evolving or adversarially-generated examples, and some evaluation suites even enable users or adversaries to submit challenging prompts in real time [5,12].

Second, there is a strong movement toward **unified, multi-metric evaluation platforms** that aggregate not just traditional accuracy scores, but also measures of robustness, calibration, fairness, toxicity, and bias, yielding a more holistic assessment of model strengths and weaknesses [13]. These platforms allow for side-by-side comparison and enable researchers to detect trade-offs or blind spots that might otherwise go unnoticed.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406011|

Third, **hybrid evaluation pipelines** are gaining popularity: scalable LLM-as-a-judge methods are increasingly used for initial triage, pairwise ranking, or large-scale filtering, while human evaluation is reserved for cases where nuance, creativity, safety, or ethical judgment are required [1,2,17]. Studies have shown that, when properly designed, LLM-based judges can achieve high agreement with aggregated human preferences, accelerating the evaluation process and freeing human reviewers to focus on the most important or ambiguous outputs [1,17].

Additionally, **adversarial testing and bias monitoring** are now standard. Evaluation pipelines regularly probe models with adversarial, edge-case, or red-teaming prompts to reveal failure modes and unwanted behaviors, particularly in areas like misinformation, social bias, or safety-related risks [2,16].

Best practices for LLM-as-a-judge include **prompt engineering**—such as randomizing answer order, calibrating model verbosity, and using explicit rubrics—to control for systematic biases and maximize reliability [1,17,18]. Lastly, there is a sustained emphasis on **transparency and reproducibility** through open-source benchmarks, public leaderboards, and community-driven evaluation frameworks, fostering trust and accelerating collective progress [8,12,13].

VI. CONCLUSION

Taken together, these trends point to a future where robust LLM evaluation depends on integrated, adaptive frameworks that balance programmatic automation, human judgment, and scalable model-assisted review [2,5,13,17]. No single approach is sufficient: automated metrics provide scale and speed, human evaluation grounds result in real-world values and ethics, and LLM-as-a-judge methods offer scalable, efficient triage. Only by combining these methods can researchers and practitioners ensure the safe, ethical, and trustworthy deployment of increasingly powerful language models.

REFERENCES

1. L. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” NeurIPS Datasets and Benchmarks Track, 2023.
2. Z. Guo et al., “Evaluating Large Language Models: A Comprehensive Survey,” arXiv:2310.19736 [cs.CL], 2023.
3. T. Brown et al., “Language Models are Few-Shot Learners,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020.
4. C. Dilmegani, “Large Language Model Evaluation in 2025: 10+ Metrics & Methods,” AI Multiple, May 2025.
5. Y. Chang et al., “A Survey on Evaluation of Large Language Models,” ACM Trans. Intell. Syst. Technol., 2023.
6. A. Hendrycks et al., “Measuring Massive Multitask Language Understanding,” ICLR, 2021.
7. H. Huang et al., “C-Eval: A Comprehensive Evaluation for Chinese LLMs,” arXiv:2305.08322 [cs.CL], 2023.
8. M. Novikova et al., “Why We Need New Evaluation Metrics for NLG,” EMNLP, 2017.
9. R. Papineni et al., “BLEU: a method for automatic evaluation of machine translation,” ACL, 2002.
10. T. Zhang et al., “BERTScore: Evaluating Text Generation with BERT,” ICLR, 2020.
11. M. Chen et al., “Evaluating Large Language Models Trained on Code,” arXiv:2107.03374 [cs.LG], 2021.
12. S. Kiela et al., “Dynabench: Rethinking Benchmarking in NLP,” Proc. of ACL, 2021.
13. K. Liang et al., “Holistic Evaluation of Language Models (HELM),” Stanford CRFM, 2022.
14. E. Clark et al., “Sentence Mover’s Similarity: Automatic Evaluation for Multi-Sentence Texts,” ACL, 2019.
15. A. Tamkin et al., “CrowdED: Crowd-Sourced Human Evaluation Data for NLP,” ACL, 2023.
16. S. Lin et al., “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” NeurIPS, 2021.
17. Tatsu Lab, “AlpacaEval: An Automatic Evaluator for Instruction-following Language Models,” GitHub, 2023.
18. OpenAI, “OpenAI Evals Documentation,” 2024
19. LMSYS, “Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings,” 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details