Comprehensive Survey on Evaluation Methods for Large Language Models

Pouya Faridfar[1,*] - Roshanak ghasemian[2] - Reza Foroughnia[3]
[1] Department of computer engineering, Mashhad branch, Islamic Azad University, Mashhad, Iran.
[2] Independent Researcher, Mashhad, Iran.
[3] Department of computer engineering, Mashhad branch, Islamic Azad University, Mashhad, Iran.

ABSTRACT

Large Language Models (LLMs) have rapidly gained prominence in both research and industry due to their remarkable performance across diverse linguistic and cross-disciplinary tasks. This rapid progress highlights the need for systematic evaluation, not only in terms of technical performance but also regarding ethical, social, and safety concerns. This survey provides a comprehensive overview of evaluation approaches for LLMs along three dimensions: what to evaluate, where to evaluate, and how to conduct evaluation. It reviews tasks such as natural language processing, reasoning, medicine, education, social sciences, and agent-based applications. In addition, it summarizes common datasets, benchmarks, and metrics, while identifying both success cases and limitations of current models. Finally, the paper outlines future challenges, including the design of new evaluation protocols, handling bias, ensuring factual reliability, and improving trustworthiness, aiming to guide the development of more capable, safe, and reliable LLMs.

Keywords: *Large Language Models (LLMs), Evaluation Methods, Benchmarks, Trustworthiness, Bias and Robustness*

I. INTRODUCTION

The pursuit of understanding intelligence and replicating it through machines has long been a central theme in Artificial Intelligence (AI) research. Early milestones such as the Turing Test [1] provided a conceptual framework to measure whether machines could exhibit human-like reasoning and behavior. Over time, a succession of algorithms—from the Perceptron [2] to Support Vector Machines (SVMs) [3] and later deep learning [4]—have been developed, each advancing the state of AI while also exposing important limitations. These developments underline the critical role of evaluation, as it is only through rigorous testing that the strengths, weaknesses, and risks of a system can be revealed [5].

In recent years, Large Language Models (LLMs) have emerged as transformative tools in both academic research and industry [16, 7, 8]. Unlike earlier models restricted to narrow tasks, LLMs demonstrate broad generalization capabilities, supporting applications in natural language understanding, reasoning, healthcare, education, and beyond. Models such as GPT-3 [9], InstructGPT [10], and GPT-4 [11] illustrate the leap in scale and performance, largely enabled by transformer architectures [12] and training techniques such as Reinforcement Learning with Human Feedback (RLHF) [23, 14]. Their remarkable versatility has fueled speculation about LLMs as potential precursors to Artificial General Intelligence (AGI) [15]. However, the very breadth of their applications also amplifies risks, including misinformation, bias, ethical concerns, and lack of robustness in safety-critical domains such as finance and healthcare [16, 17].

The evaluation of LLMs, therefore, has become a multidimensional challenge. Traditional evaluation frameworks—such as static validation sets or standard NLP benchmarks like GLUE [18] and SuperGLUE [19]—are insufficient to fully capture the emergent abilities and complex behaviors of modern LLMs [20].

Researchers now emphasize the need to assess LLMs across diverse axes: not only task performance in NLP, reasoning, or domain-specific contexts, but also robustness against adversarial prompts, factual consistency, bias mitigation, and societal trustworthiness [21, 22, 20]. Moreover, the interactive nature of LLMs requires evaluation protocols that consider human–AI collaboration, user safety, and long-term reliability.

This paper provides a comprehensive survey of LLM evaluation along three guiding questions: what to evaluate, where to evaluate, and how to evaluate. The "what" dimension includes core tasks such as natural language understanding, reasoning, multilingual performance, factuality, and applications in science, medicine, and education. The "where" dimension highlights the benchmarks and datasets used for systematic assessment, while the "how" dimension explores methodologies ranging from classical validation to novel prompt-based strategies and adversarial testing. By synthesizing findings across these dimensions, this survey identifies both the successes and failures of LLMs and outlines pressing future challenges, such as the design of new evaluation protocols, handling ethical risks, and improving transparency and interpretability [16, 15, 22].

II. Background

1. Large Language Models

Language Models (LMs) are computational systems designed to understand and generate human language. Their primary function is to predict the likelihood of word sequences or generate coherent text based on given input. Traditional approaches, such as n-gram models [23], estimated probabilities using local context but struggled with unseen words, overfitting, and limited ability to capture complex linguistic dependencies [24, 25]. To overcome these limitations, more advanced architectures were developed, culminating in Large Language Models (LLMs), which are characterized by massive parameter counts and remarkable generalization abilities [19 , 8].

Modern LLMs—such as GPT-3 [9], InstructGPT [10], and GPT-4 [11]—are built on the Transformer architecture [12], whose self-attention mechanism enables efficient handling of sequential data, long-range dependencies, and parallel computation. Two major innovations underpin the success of LLMs. First, in-context learning allows models to generate responses by conditioning on prompts without explicit fine-tuning [14]. Second, Reinforcement Learning from Human Feedback (RLHF) [23, 14] improves alignment with user expectations by refining model outputs using human-provided preferences. Together, these features enable LLMs to perform well in interactive applications such as dialogue systems, reasoning tasks, and domain-specific assistance.

Formally, in autoregressive LMs like GPT-3 and PaLM [24], the goal is to predict the next token y given a sequence of prior tokens

$$P(y \mid X) = P(y \mid x_1, x_2, \dots, x_{t-1})$$

The conditional probability is modeled as:

$$P(y \mid X) = \prod_{t=1}^{T} P\left(y_t \mid x_1, x_2, \dots, x_{\{t-1\}}\right)$$

where T is the sequence length. This formulation enables sequential generation of coherent text outputs. In practice, techniques such as prompt engineering [26] and question–answering interactions are widely used for guiding LLM behavior.

A comparison of traditional machine learning, deep learning, and LLMs highlights their distinctions. LLMs rely on very large datasets, exhibit extremely high model complexity, and require substantial hardware resources, while offering the highest performance but at the cost of poor interpretability [19, 8].

2. AI Model Evaluation

Evaluation is a cornerstone of AI development, providing systematic insight into model performance, limitations, and potential risks. Traditional protocols include k-fold cross-validation, holdout validation,

leave-one-out cross-validation (LOOCV), bootstrap, and reduced set methods [8]. For example, k-fold cross-validation divides data into k parts and tests iteratively, while holdout validation is simpler but more prone to bias. LOOCV minimizes bias but is computationally expensive, and reduced set evaluation is computationally efficient but less reliable.

In deep learning and LLM contexts, static validation sets are more commonly employed due to the scale of training. For instance, vision models use ImageNet and MS COCO, while NLP models rely on GLUE [18] and SuperGLUE [19]. However, the increasing scale and declining interpretability of LLMs have exposed the insufficiency of conventional evaluation strategies.

The evaluation of LLMs, therefore, requires not only static benchmarks but also task-specific, interactive, and adversarial protocols that better reflect their real-world deployment [20]. As a result, evaluation has evolved into a discipline in itself, crucial for ensuring that LLMs are not only powerful but also trustworthy, safe, and aligned with societal needs.

III. WHAT TO EVALUATE

Evaluating Large Language Models (LLMs) first requires identifying the tasks that best reveal their strengths and limitations. In natural language processing (NLP), key tasks include sentiment analysis, text classification, natural language inference (NLI), and semantic understanding. Studies show that LLMs achieve strong performance in sentiment and news classification, yet they remain weaker in capturing human disagreement and nuanced semantics. Another critical dimension is reasoning, which covers mathematical, logical, and multi-step reasoning. While GPT-4 often outperforms traditional fine-tuned models on logical reasoning tasks, its performance in abstract and counterfactual reasoning remains limited.

Natural language generation (NLG) tasks such as summarization, translation, question answering (QA), and dialogue also form a cornerstone of evaluation. For instance, GPT-4 and ChatGPT outperform conventional machine translation systems in human evaluations [16], but still lag behind in English → low-resource language translation. Likewise, ChatGPT delivers high accuracy on QA benchmarks [9] yet occasionally fails on commonsense challenges due to its cautious refusal strategy. Beyond these, multilingual evaluation highlights that LLMs perform poorly on non-Latin and under-resourced languages [16], and factuality assessment shows persistent issues with hallucinations despite GPT-4's progress [21, 17]. Overall, evaluation must span both classical NLP tasks and broader aspects such as reasoning, factual reliability, robustness, and ethical trustworthiness.

IV. WHERE TO EVALUATE

Choosing datasets and benchmarks is central to systematic evaluation. For general NLP, GLUE [18] and SuperGLUE [19] are the most widely used. To measure adversarial and out-of-distribution (OOD) robustness, researchers use AdvGLUE [22] and ANLI. Factuality is assessed with datasets such as TruthfulQA, which deliberately exposes factual inconsistencies. For dialogue, LMSYS-Chat-1M provides a large-scale conversational benchmark. In medical applications, standardized exams like the USMLE are employed to evaluate LLMs' ability to answer professional questions. For multilingual testing, multiple corpora are used to cover both high-resource and low-resource languages [16]. These benchmarks collectively ensure that LLM evaluation reflects not only task accuracy but also robustness, fairness, and domain-specific applicability.

V. HOW TO EVALUATE

Once tasks and datasets are defined, the next step is selecting appropriate methods, metrics, and protocols. Classical approaches include cross-validation, holdout validation, bootstrap, and LOOCV [8]. For large-scale LLMs, however, evaluation typically relies on static test sets due to computational demands. Widely used metrics include accuracy, F1-score, BLEU, and ROUGE for classification and generation [19]. Newer measures have been introduced for specific aspects: FActScore for factuality, PromptBench for adversarial prompt robustness [20], and DecodingTrust for multi-faceted trustworthiness [22].

Recent trends emphasize prompt-based evaluation and adversarial testing, where model responses are tested under manipulated inputs to probe robustness [16, 201]. Importantly, evaluation is expanding beyond static

correctness to consider human–AI interaction quality, bias mitigation, safety, and long-term reliability. This multidimensional approach acknowledges that LLMs are not only technical systems but also social actors whose outputs affect users and institutions.

VI. Key Findings

The survey highlights several major findings regarding the evaluation of Large Language Models (LLMs). First, LLMs show impressive performance in classical NLP tasks such as sentiment analysis, classification, summarization, and QA, often surpassing smaller models or traditional supervised methods. However, their success is uneven across subtasks: while they excel in broad text classification, they struggle with semantic nuance, natural language inference, and multilingual tasks [16].

Second, in reasoning tasks, LLMs demonstrate progress compared to earlier models. GPT-4, in particular, outperforms GPT-3.5 on many benchmarks, especially in logical reasoning . Yet weaknesses remain in areas such as abstract, counterfactual, and multi-hop reasoning. Performance is also inconsistent in mathematical problem solving, where results drop significantly with increasing difficulty [3].

Third, in robustness and trustworthiness, evaluations reveal clear vulnerabilities. Studies using PromptBench show that current LLMs are highly sensitive to adversarial prompts [20]. Similarly, robustness to out-of-distribution data is limited, raising concerns about reliability in real-world applications. Ethical assessments indicate that LLMs can reproduce toxic or biased content and display cultural or political biases [16, 22]. Although GPT-4 often performs better than GPT-3.5 in trust-oriented evaluations, it is simultaneously more vulnerable to targeted attacks [22].

Fourth, in domain-specific applications, such as medicine, education, and law, LLMs provide valuable assistance but remain insufficient for expert-level deployment. For instance, GPT-4 approaches the passing threshold on the USMLE medical exam without domain-specific fine-tuning, while in law, models still generate inconsistent or hallucinated summaries. These findings emphasize that LLMs are powerful assistants but not substitutes for domain specialists.

Overall, the survey confirms that evaluation must be multidimensional, covering not just accuracy in tasks but also robustness, factuality, ethics, and human–AI interaction quality.

VII. Future Challenges

Despite progress, several future challenges remain central to LLM evaluation.

Emergent Capabilities: As models scale, they exhibit behaviors not easily captured by existing benchmarks, such as in-context learning and reasoning beyond training data [15]. New protocols are needed to measure these emergent properties reliably.

Bias and Fairness: Persistent biases in outputs reflect imbalances in training data. Evaluations must evolve to systematically measure demographic, cultural, and ideological biases, as well as to test fairness across multiple languages and domains.

Factuality and Hallucination: Current metrics are inadequate for reliably detecting hallucinations in long-form outputs. Future evaluations must integrate both automated and human-in-the-loop methods to ensure factual consistency [21].

Interactive and Dynamic Evaluation: Most benchmarks rely on static datasets, yet LLMs are deployed in dynamic contexts such as dialogue and decision-making. Evaluations need to incorporate real-time, user-centered, and adversarial testing [22].

Cross-Domain Generalization: LLMs are increasingly applied in high-stakes fields like healthcare, law, and engineering. Designing cross-domain benchmarks that simulate real-world complexity will be critical to ensuring safe deployment.

Resource Efficiency and Environmental Impact: Large-scale evaluations demand massive computational resources, raising sustainability concerns. Future work must balance comprehensiveness with cost-effectiveness in designing benchmarks.

Together, these challenges emphasize that evaluation must move beyond static accuracy metrics toward a holistic, continuous, and context-aware process.

VIII. Conclusion

This survey underscores that evaluation is not a secondary step but a foundational discipline in the development of LLMs. While existing research has produced valuable insights—demonstrating LLMs' strong performance in NLP, reasoning, and certain specialized domains—it has also revealed critical gaps in robustness, factual reliability, and fairness. The multidimensional nature of evaluation, spanning "what, where, and how," illustrates that task selection, benchmark design, and protocol development must work in concert to capture the full spectrum of model capabilities and risks.

Looking forward, the community must prioritize the design of novel benchmarks and evaluation frameworks that address emergent abilities, societal impacts, and safety-critical applications. Open-source initiatives, such as shared benchmark repositories and collaborative evaluation platforms, will be vital to advancing transparent and reproducible assessments. Furthermore, interdisciplinary collaboration across computer science, social sciences, law, and medicine will ensure that evaluations reflect both technical performance and societal trust.

In conclusion, rigorous and comprehensive evaluation will play a decisive role in guiding the development of LLMs toward systems that are not only powerful but also reliable, safe, and aligned with human values. The findings of this survey highlight both the successes achieved and the urgent challenges that remain, setting a clear research agenda for the next generation of LLM evaluation studies.

REFERENCES

[1] *Alan M. Turing. 2009. Computing Machinery and Intelligence. Springer.*
[2] *Stephen I. Gallant et al. 1990. Perceptron-based learning algorithms. IEEE Transactions on Neural Networks 1, 2 (1990), 179–191.*
[3] *Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine Learning 20 (1995), 273–297.*
[4] *Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. Nature 521, 7553 (2015), 436–444.*
[5] *Jean Khalfa. 1994. What is intelligence? (1994).*
[6] *Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).*
[7] *Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai Hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. Trans. Mach. Learn. Res. 2022 (2022).*
[8] *WayneXinZhao,KunZhou,JunyiLi,TianyiTang,Xiaolei Wang,YupengHou,YingqianMin,BeichenZhang,Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223 (2023).*
[9] *Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines 30 (2020), 681–694.*
[10] *Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong*

Zhang, Sandhini Agarwal,KatarinaSlama,AlexRay,etal.2022.Traininglanguagemodelstofollowinstructionswithhumanfeedback. Advances in Neural Information Processing Systems 35 (2022), 27730–27744.

[11] *OpenAI. 2023. GPT-4 Technical Report. (2023). arXiv:cs.CL/2303.08774*

[12] *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).*

[13] *Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems 30 (2017).*

[14] *Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019).*

[15] *SébastienBubeck,VarunChandrasekaran,RonenEldan,JohannesGehrke,EricHorvitz,EceKamar,PeterLee,YinTat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023).*

[16] *Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucina tion, and interactivity. arXiv preprint arXiv:2302.04023 (2023).*

[17] *Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023. Evaluating open ques tion answering evaluation. arXiv preprint arXiv:2305.12421 (2023).*

[18] *Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018).*

[19] *AlexWang,YadaPruksachatkun,NikitaNangia,AmanpreetSingh,JulianMichael,FelixHill,OmerLevy, andSamuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems 32 (2019).*

[20] *Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhen qiang Gong, Yue Zhang, et al. 2023. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint arXiv:2306.04528 (2023).*

[21] *Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. arXiv preprint arXiv:2204.04991 (2022).*

[22] *Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, SimranArora, MantasMazeika,DanHendrycks,ZinanLin,YuCheng,Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. (2023). arXiv:cs.CL/2306.11698*

[23] *Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. Computational Linguistics 18, 4 (1992), 467–480.*

[24] *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).*

[25] *Jianfeng Gao and Chin-Yew Lin. 2004. Introduction to the special issue on statistical language modeling. (2004), 87–93.*