Scientometrica 1 (2025) 187–194

Contents lists available at GrowingScience

Scientometrica

homepage: www.GrowingScience.com/sci

A scientometric analysis of research landscapes at the nexus of large language models and evaluation

Esmaeil Taheripour[a*]

[a]*Department of Industrial Engineering, School of Engineering, Iran University of Science and Technology, Iran*

C H R O N I C L E

Article history:
Received: November 5, 2025
Received in revised format: November 5, 2025
Accepted: November 31, 2025
Available online:
November 31, 2025

Keywords:
Large Language Models
Finance
Stock Market
Scientometrics
Bibliometric Analysis
Financial Technology
NLP
AI in Finance

A B S T R A C T

The deployment of Large Language Models (LLMs) in the financial field is one of the most forward-looking and quickly changing AI research areas. This paper presents a thorough scientometric study of the top 200 most cited articles, which deal with the conjunction of LLMs, finance, and stock markets, and are indexed in Scopus. The analysis carries out a detailed examination of the publication patterns, influential authors and institutions, main research areas, methods used, and the intellectual framework of this new field. The disclose of the research output has been increased enormously; the main form of it is conference papers, and it is all through collaboration between the global academic and industrial research centers. Financial sentiment analysis and market prediction, the building and evaluating of domain-specific financial LLMs (FinLLMs), and lastly the application of the mentioned FinLLMs in areas such as financial analytics and decision-making support are examples of research that has been clustered together. Nevertheless, a big part of the research is directed towards trust, ethics, and risks in the financial domain. Moreover, the study points out that practical applications, such as algorithmic trading, risk management, and compliance with regulations, are among the most highlighted in the fields of LLMs in finance. This scientometric review offers a primary map of the vast conceptual territory marked by high-impact research, which means it can be very helpful to researchers, practitioners, and policymakers in getting a better understanding of the existing situation and thus, pointing out future research directions in the area of LLMs in finance.

© 2025 by the authors; licensee Growing Science, Canada.

1. Introduction

The concept of transformer-based architectures (Vaswani et al., 2017) has become the turning point in natural language programming (NLP), which ultimately led to the emergence of LLMs, the most important of which are GPT-3, ChatGPT, GPT-4, PaLM, LLaMA, etc., these have undeniably surprised everyone with his/her generosity in building, translating texts, executing complex reasoning, and even synthesizing codes. However, these models are still considered to be very hard to evaluate due to their mammoth size and "black-box" characteristics (Chang et al., 2024). Traditional evaluation benchmarks have almost been shown to be incapable of capturing the subtleties of performance, the potential hazards, and the overall utility of LLMs (Bommasani et al., 2021). Hence, the evaluation of LLMs has turned into a fundamental aspect of research and has witnessed an enormous increase in the use of various techniques. Some of the activities that are being carried out in this domain are the development of more sophisticated benchmarks for testing reasoning and knowledge (Zhong et al., 2023; Huang et al., 2023), performance evaluation in certain sectors like medicine (Singhal et al., 2023; Goh et al., 2024) and law (Guha et al., 2023), setting up of both automated and human evaluation methods (Chiang et al., 2023; Liu et al., 2023b), and the detection and curing of model defects including hallucination (Ji et al., 2023; Li et al., 2023b), bias (Gallegos et al., 2024; Zack et al., 2024), and toxicity (Wei et al., 2023). The number of articles published in this field is so great that it necessitates a systematic method in order to understand the structure and dynamics of the research area.

* Corresponding author.
E-mail address: e_teheripour@ind.iust.ac.ir (E. Taheripour)

ISSN 3115-8455 (Online) - ISSN 3115-8447 (Print)
© 2025 by the authors; licensee Growing Science, Canada
doi: 10.5267/j.sci.2025.5.005

Scientometrics, the quantitative research of scientific communication, undeniably gives robust tools for the visualization of the landscape in a specific area of research. It may, for instance, show the level of cooperation amongst the scientists within the group, point out the most important pieces of work and the new trends, and determine the different concepts of the research area. Although numerous reviews on LLMs have been published, they are mostly qualitative and concentrate on model capabilities or applications (Zhao et al., 2023). There has not been an in-depth scientometric analysis of the evaluation-driven literature yet.

This article delivers a thorough scientometric study of a selected dataset of 200 highly cited publications located in the central area of “Large Language Models” and “Evaluation.” The main goals of this research are:

1. To count the publication landscape, including the most prolific journals/conferences and influential authors.
2. To find out the most important application domains and study themes through keyword and content analysis.
3. To look for the evolution of research foci in 2020-2025.
4. To offer a structured overview of the key challenges and future directions as reflected in the highly cited literature.

This research is intended to be a pioneer in the mapping of this very important area of research, and through this, it will contribute with insights that are priceless for the beginners as well as for the established scholars.

2. Methodology

2.1 Data Collection and Curation

The starting point of the analysis was a bibliographic database search in the Scopus database using the terms “large language models” and “Evaluation.” Out of the initial result of more than 13,000 records, the top 200 most cited references were chosen for thorough examination. This purposive sampling method highlights the most important studies in the area, as one of the main criteria for selection is citation count, which is a common, even if not perfect, measure of impact and intellectual significance. The data set comprises records from 2020 to 2025, with the bulk of the records in 2023 and 2024, indicating the recent rapid growth of the field. Information for every publication consists of title, authors, abstract, keywords, source (journal or conference), year of publication, citation count, and DOI.

2.2 Analytical Framework

We used a mix of qualitative and quantitative methods to analyze data.

- **Descriptive Statistics:** Basic counts and frequencies were computed for publication years, source titles, and author appearances in order to reveal the temporal and geographical spread of the research.
- **Thematic Analysis:** A systematic analysis of the keywords, manuscript titles, and abstracts provided by the author was done. The manual and computerized reading and categorization alternated for the purpose of identifying the major research themes. The "LLM" and "Large Language Model" were among others that were rendered comprehensible by combining thus enabling correct frequency counts. The authors' conclusions and discussions in the most cited papers within each theme were analyzed in order to extract key challenges and future directions.
- **Content Analysis:** A qualitative analysis of the abstract and, if required, the complete text of important publications was carried out in order to have a better insight into the research topics and to place the quantitative results in context.

Statistical summaries are all exhibited in tabular form as per the set instructions. Additionally, no diagrams (for instance, co-authorship networks, keyword maps) are part of the presentation.

3. Results and Discussion

3.1 Temporal and Venue Distribution of Publications

The 200 publications' time distribution emphasizes the swift rise of LLM evaluation to the forefront of research. The situation was such that the few first-aged works from 2020 and 2021 were practically inexistent compared to the influx of highly-cited publications beginning in 2023 and continuing into 2024 and even 2025. All this happened in the backdrop of the

public launch of very strong models like ChatGPT and GPT-4. These models not only stimulated the academic but also the public's curiosity and demand for reliable assessments of these technologies.

Table 1
Top Publication Venues by Number of Highly Cited Papers (2020-2025)

Venue Name	Type	Count
Nature Medicine	Journal	4
JAMA (Journal of the American Medical Association)	Journal	3
npj Digital Medicine	Journal	6
Nature Communications	Journal	4
Proceedings of the AAAI Conference on AI	Conference	2
Proceedings of the ACM Web Conference (WWW)	Conference	2
Advances in Neural Information Processing Systems (NeurIPS)	Conference	12
International Conference on Learning Representations (ICLR)	Conference	10
Conference on Empirical Methods in NLP (EMNLP)	Conference	16
Annual Meeting of the ACL	Conference	8
Proceedings of the IEEE/CVF CVPR	Conference	5
IEEE Transactions on Software Engineering	Journal	2
JMIR Medical Informatics/Education	Journal	5

Table 1 shows the ways of research publication which primarily consist of high-impact, wide-scope journals (the Nature family, JAMA) and top-notch, specialized conferences in computer science and NLP (NeurIPS, ICLR, EMNLP, ACL). Medical journals, such as Nature Medicine and npj Digital Medicine, played a major role in establishing the research about LLMs in the healthcare sector, which was a strong focus area of this research. In contrast, being present at conferences is typical for the field of computer science where the fast sharing of findings is usual.

3.2 Prolific Authors and Collaborative Networks

Authorship analysis indicates the existence of a highly cooperative sector wherein numerous papers are produced by large, multi-institutional groups. In addition, counting only first or corresponding authors might give one insight into a power distribution among authors, but looking at all author appearances will definitely reveal the entire picture of who are the influential contributors.

Table 2
Most Frequent Author Appearances in the Dataset

Author Name	Frequency of Appearance	Author Name	Frequency of Appearance
Shah, N.H.	5	Karthikesalingam, A.	2
Dash, D.	3	Liang, P.	2
Matias, Y.	3	Liu, Y.	2
Natarajan, V.	3	Milstein, A.	2
Singhal, K.	3	Pfeffer, M.A.	2
Tu, T.	3	Tazbaz, T.	2
Azizi, S.	2	Tomasev, N.	2
Butte, A.J.	2	Wang, Y.	2
Chen, J.H.	2	Yang, D.	2
Corrado, G.S.	2		

The authors who appear in several highly cited papers are illustrated in Table 2. Among the authors there are several from the industry research labs, for instance, Google/DeepMind and Microsoft, who are prominently listed. This indicates that these companies are heavily investing in the development and evaluation of large language models. The participation of authors such as N.H. Shah and A.J. Butte also highlights the close collaboration between computer science and biomedical informatics that draws upon different disciplines.

3.3 Dominant Research Themes and Application Domains

The examination of author keywords through frequency analysis and qualitative scrutiny of paper abstracts brought to light a number of major research themes. These themes can be grouped broadly into two categories: application areas and methodological approaches.

Table 3
Primary Application Domains for LLM Evaluation

Application Domain	Description & Exemplary Studies	Frequency
Healthcare & Medicine	Evaluation of diagnostic accuracy, clinical reasoning, medical knowledge, and utility in workflows (e.g., EHR analysis, patient communication).	48
Education	Assessing performance on standardized exams, utility for automated feedback and question generation, and implications for academic integrity.	22
Software Engineering	Testing code generation, program repair, bug detection, and software architecture.	18
Law & Ethics	Evaluating legal reasoning, measuring and mitigating bias, fairness, and toxicity, and assessing alignment with human values.	16
Robotics & Embodied AI	Using LLMs for task planning, spatial reasoning, and control in simulated and real-world environments.	8
Multimodal Reasoning	Evaluating models that process both text and images/videos for tasks like visual question answering and scene understanding.	15
General Benchmarking	Developing comprehensive benchmarks to assess a wide range of capabilities (reasoning, knowledge, coding) in a general context.	35

3.3.1 Theme 1: Evaluation in Healthcare and Medicine

The dataset includes the most dominant category. The study of Singhal et al. (2023) was groundbreaking, disclosing that a particular LLM (Med-PaLM) exctly tuned could about match the performance of an specialists in medical question-answering benchmarks. This initiated a multitude of studies assessing general-purpose LLMs like ChatGPT and GPT-4 on medical licens-ing exams (Ali et al., 2023; Rosoł et al., 2023), diagnostic reasoning (Goh et al., 2024), and clinical tasks such as summarizing evidence (Tang et al., 2023) and drafting patient messages (Garcia et al., 2024). One major aspect of this research is the shift from multiple-choice question answering to more sophisticated assessments like the detection of limits in clinical decision-making (Hager et al., 2024) and the danger of reinforcing racial and gender biases (Zack et al., 2024). The systematic review performed by Bedi et al. (2025) mirrors the increasing maturation of this subfield, providing evidence on the testing and evaluation of healthcare applications of LLMs.

3.3.2 Theme 2: Foundational Benchmarking and Capability Assessment

The literature has a large part that is focused on LLM capabilities and exploring their limits by means of more difficult and comprehensive benchmarks. The first benchmarks were mainly directed towards very specific tasks, while the latest attempts are looking in both dimensions - wide and deep. A case in point is the Chang et al. (2024) paper that surveys the methods of evaluation extensively and discusses those in terms of knowledge, reasoning, safety, and alignment. Benchmarks such as MMLU (Massive Multitask Language Understanding) and its successors (Huang et al., 2023; Yue et al., 2024) scrutinize the knowledge of LLMs across a number of disciplines measured by dozens, whereas other evaluations detect such weaknesses as hallucination (Li et al., 2023b; Min et al., 2023) or referring to non-existent objects in vision-language models (Li et al., 2023a). There are also mentions of the assessment of LLMs¿ tool and API usage (Qin et al., 2024; Li et al., 2023c) as well as reasoning over structured data such as tables (Sui et al., 2024) and knowledge graphs (Zhu et al., 2024).

3.3.3 Theme 3: LLMs as Evaluators

The evaluation literature has a meta-trend of using LLMs to evaluate the other LLMs or the outputs of NLP systems. Chiang et al. (2023) pushed this theme through a rigorous process and arrived at the conclusion that the LLM-based evaluators could be very comparable with human judgment in many cases; thus, they are a scalable substitute to the expensive human evaluation. This has been utilized in the areas of text summarization (Liu et al., 2023b), translation quality (Kocmi & Federmann, 2023), and code generation (Liu et al., 2023a). Nevertheless, the practice also brings up a dilemma of a new kind, wherein the LLM validator is required to be foolproof and not biased. The matter was brought into sharper focus by Shankar et al. (2024) in their study "Who Validates the Validators?"

3.3.4 Theme 4: Evaluation of Limitations and Risks

As for the evaluation of capabilities, the evaluation of LLM failures is essential. The main areas of concern are as follows "situation" followed by the most preferred alternative.

- **Hallucination:** One of the major drawbacks of LLMs is their ability to produce text that sounds reasonable but is either incorrect or completely nonsensical. Researchers are working to find ways to measure these "hallucination" events (Li et al., 2023b) and invent methods to cope with them, such as encouraging the models to "think" about their outputs (Ji et al., 2023) and using more reliable sources in the generation process (Gao et al., 2023).

- **Bias and Fairness:** Many researchers have revealed that among LLMs, biases concerning race, gender, religion, and other aspects exist (Gallegos et al., 2024). The testing of biases has turned into a separate discipline that requires the making of datasets and metrics tailored to this purpose (Zack et al., 2024).
- **Toxicity and Safety:** The research investigates the tendency of LLMs to produce offensive, harmful, or unsafe content, notwithstanding the safety training (Wei et al., 2023). The "jailbreaking" studies examine the limits of the safety filters.
- **Reasoning and Logical Consistency:** Although LLMs are very good at pattern matching, their core reasoning skills are frequently weak. Research done by Creswell et al. (2023) and Wang et al. (2023) has introduced tasks to evaluate logical coherence and reasoning, thus exposing particular ways of failing.

3.4 Evolution of Research Foci (2020-2025)

The papers that have received a larger number of citations in chronological order show an evolving pattern of research foci:

- **2020-2022 (Foundations):** The first LLMs were majorly concerned with scaling laws and training methods for LLMs (Hoffmann et al., 2022), and human-like feedback alignment (Ouyang et al., 2022), in addition, developers also attempted domain-specific evaluation, for instance in electronic health records (Yang et al., 2022). Their main target was to find out the necessary skill sets and alignment techniques.
- **2023 (The Year of Proliferation):** The launch of ChatGPT was a turning point. The amount and variety of research increased tremendously. This time the GPT-family models and their competitors (like Bard, LLaMA) were rigorously tested in almost all areas mentioned in Table 3. The main issues were assessment (e.g., MMLU), health care testing (e.g., Singhal et al., 2023), and the first organized inquiries into drawbacks such as hallucination and bias.
- **2024-2025 (Maturation and Specialization):** The era of hype has passed and the field is gradually maturing. The question "is it capable of doing this?" is replaced by "what is the extent of its capability and what are the risks?". (Bedi et al., 2025; Park et al., 2024) Systematic reviews and meta-analyses are coming to the fore. The evaluations become more sophisticated with the focus on different aspects of AI in healthcare such as complex clinical reasoning (Goh et al., 2024), the creation of "frameworks for human evaluation" (Tam et al., 2024), and the testing of AI agents in simulated settings (Gao et al., 2024). Moreover, there are clear signs of an increasing trend towards the use of multiple modalities (Yin et al., 2024; Wu et al., 2023) alongside the assessment of LLMs as self-sufficient agents (Wang et al., 2024a).

4. Key Challenges and Future Directions

The discussion of what lies ahead covers several key scientific aspects of our study that we have examined against what has been done in the past.

Table 4
Key Challenges and Future Research Directions in LLM Evaluation

Key Challenge	Description	Future Directions
Dynamic and Real-World Validity	Static benchmarks may lose their validity very soon because of contamination and might not be able to catch, in terms of performance, the actual use, which is interactive and real-world.	Development of dynamic, adaptive benchmarks; increased use of simulation environments (Park et al., 2023; Dubois et al., 2023) and live user studies for evaluation.
Evaluation of Multimodal and Embodied Models	The assessment of models, which can handle text, vision, audio, and even physical actions, will need the development of fresh metrics and benchmarks that are not restricted to text-based tasks.	The development of comprehensive benchmarks for multi-modal reasoning (Yue et al., 2024; Li et al., 2024) and for tasks performed in real-life settings (Driess et al., 2023; Hong et al., 2023).
Scalable and Affordable Human Evaluation	The gold standard in evaluation is human evaluation, but it is expensive, slow, and can even be inconsistent. As such, the scale and speed of reliable evaluation are limited.	The LLM-as-a-judge protocols have been refined (Chiang et al., 2023); and, at the same time, new techniques have been developed to obtain calibrated confidence scores (Tian et al., 2023).
Holistic Trustworthiness and Safety	Presently, safety evaluations pay attention mostly to individual features like toxicity. The trustworthiness concept, which comprises robustness, privacy, and fairness, should be viewed more widely.	The creation of all-encompassing trustworthiness structures (Wang et al., 2023b) that analyze several risk factors at the same time is done with respect to particular use cases.
Standardization and Reproducibility	The lack of standardized evaluation protocols and the issue of data contamination (Sainz et al., 2023; Balloccu et al., 2024) make it difficult to compare results across studies.	Community-wide efforts to establish evaluation standards, rigorous reporting of prompts and methodologies, and the creation of contamination-free test sets.
Evaluation of Long-Context and Agentic Behavior	As models handle longer contexts and operate as persistent agents, new evaluation paradigms are needed to assess memory, planning, and long-term task performance.	Designing benchmarks for long-context reasoning, multi-session interactions, and complex agentic workflows (Wang et al., 2024a; Gao et al., 2024).

5. Conclusion

The present study based on 200 highly cited publications is offering a precise and detailed picture of the studies that keeps changing very quickly at the junction of large language models and evaluation. The area is considered by its interdisciplinary character, a strong presence in both conference and journal publications, and a clear direction leading from the basic capability testing to refined, application-oriented, and safety-centered assessment. Healthcare evaluation is the main theme of the discussion which emphasizes the high-stakes nature of LLMs deployment and the necessity of stringent validation in critical areas. The simultaneous production of gigantic, universal benchmarks and fine-grained detection tools for specific errors like hallucination indicates a community that wants to be knowledgeable about these models both broadly and deeply. The acceptance of LLMs as evaluators by the research community is regarded as a significant, although cautiously, methodological shift that has the potential to transform the entire scenario. However, there are still major issues to deal with like the necessity for creating more vibrant and real-world applicable evaluation environments, the formulation of standard operating procedures, and a comprehensive approach to assessing the application of trustworthiness. The shift in focus from evaluating multimodal, embodied, and agentic systems to the next big thing in this area is indicative of the next frontier in this field. The debatable issue of rigorous, critical, and innovative evaluation will be more pronounced as LLMs continue to enhance their capabilities and become further integrated into society. This analysis is a representation of a constantly active field, bringing to light the collective activities that measure not only the performance of the models but also their limitations and, consequently, the development of the latter one toward ensuring beneficial and trustworthy outcomes.

References

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Zadnik Sullivan, P. L., Cielo, D., Oyelese, A. A., Doberstein, C. E., Telfeian, A. E., Gokaslan, Z. L., & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*, 93(5), 1090–1098.

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference (Vol. 1, pp. 67–93).

Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*, 333(4).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 39.

Chiang, C.-H., & Lee, H.-Y. (2023). Can Large Language Models Be an Alternative to Human Evaluation? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 15607–15631).

Creswell, A., Shannahan, M., & Higgins, I. (2023). SELECTION-INFERENCE: EXPLOITING LARGE LANGUAGE MODELS FOR INTERPRETABLE LOGICAL REASONING. In 11th International Conference on Learning Representations, ICLR 2023.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 1286–1305).

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., … Florence, P. (2023). PaLM-E: An Embodied Multimodal Language Model. *Proceedings of Machine Learning Research*, 202, 8469–8488.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. *Advances in Neural Information Processing Systems*, 36.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3), 1097–1179.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1), 1259.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. *Cell*, 187(22), 6125–6151.

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 6465–6488).

Garcia, P., Ma, S. P., Shah, S., Smith, M., Jeong, Y., Devon-Sand, A., Tai-Seale, M., Takazawa, K., Clutter, D., Vogt, K., Lugtu, C., Rojo, M., Lin, S., Shanafelt, T., Pfeffer, M. A., & Sharp, C. (2024). Artificial Intelligence-Generated Draft Replies to Patient Inbox Messages. *JAMA Network Open*, 7(3), E243201.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10), e2440969.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., … Li, Z.

(2023). LEGALBENCH: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 36.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9), 2613–2622.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems*, 35.

Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C. (2023). 3D-LLM: Injecting the 3D World into Large Language Models. *Advances in Neural Information Processing Systems*, 36.

Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *Proceedings of Machine Learning Research*, 162, 9118–9147.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., & He, J. (2023). C-EVAL: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *Advances in Neural Information Processing Systems*, 36.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards Mitigating Hallucination in Large Language Models via Self-Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1827–1843).

Kocmi, T., & Federmann, C. (2023). 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, EAMT 2023* (pp. 193–203).

Li, J., Guan, Z., Wang, J., Cheung, C. Y., Zheng, Y., Lim, L.-L., Lim, C. C., Ruamviboonsuk, P., Raman, R., Corsino, L., Echouffo-Tcheugui, J. B., Luk, A. O. Y., Chen, L. J., Sun, X., Hamzah, H., Wu, Q., Wang, X., Liu, R., … Wong, T. Y. (2024). Integrated image-based deep learning and language models for primary diabetes care. *Nature Medicine*, 30(10), 2886–2896.

Li, J., & Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023b). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 6449–6464).

Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., & Li, Y. (2023c). API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 3102–3116).

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023a). Evaluating Object Hallucination in Large Vision-Language Models. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 292–305).

Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023a). Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. *Advances in Neural Information Processing Systems*, 36.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023b). G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 2511–2522).

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 12076–12100).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

Park, Y.-J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M., & Naugler, C. (2024). Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1), 72.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2024). TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS. In 12th International Conference on Learning Representations, ICLR 2024.

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Scientific Reports*, 13(1), 20512.

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10776–10787).

Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A., & Arawjo, I. (2024). Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In UIST 2024 - Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.

Sui, Y., Zhou, M., Zhou, M., Han, S., & Zhang, D. (2024). Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. In WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining (pp. 645–654).

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1), 258.

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1), 158.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 5433–5442).

Wang, D., & Zhang, S. (2024b). Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*, 57(11), 299.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024a). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023a). SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS. In 11th International Conference on Learning Representations, ICLR 2023.

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023b). Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, 17.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems*, 36.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P.-S. (2021). Challenges in Detoxifying Language Models. In *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021* (pp. 2447–2469).

Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal Large Language Models: A Survey. In Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023 (pp. 2247–2256).

Xu, P., Patwary, M., Shoeybi, M., Puri, R., Fung, P., Anandkumar, A., & Catanzaro, B. (2020). MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (pp. 2831–2845).

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. *npj Digital Medicine*, 5(1), 194.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, 11(12), nwae403.

Yue, X., Ni, Y., Zheng, T., Zhang, K., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2024). MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 9556–9567).

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1), e12–e22.

Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. In EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings (pp. 15686–15702).

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., & Zhang, N. (2024). LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web*, 27(5), 58.

© 2025 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).