Physics of Life Reviews 53 (2025) 294–296

Contents lists available at ScienceDirect

Physics of Life Reviews

journal homepage: www.elsevier.com/locate/plrev

# Leveraging large language models for complex systems analysis: Opportunities, challenges, and future directions: Comment on "LLMs and generative agent-based models for complex systems research" by Y. Lu et al.

Yanan Wang [a], Xiaofang Duan [b], Xing Pan [a,*], Yini Geng [c,d,*]

[a] School of Reliability & Systems Engineering, Beihang University, Beijing, 100191, China
[b] School of Mathematics and Statistics, Xidian University, Xi'an, Shaanxi 710126, China
[c] Key Laboratory of Computing and Stochastic Mathematics (Ministry of Education), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, China
[d] Key Laboratory of Applied Statistics and Data Science, School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, China

A complex system is composed of numerous interacting fundamental units, whose collective behavior cannot be straightforwardly inferred from the properties of individual components [1]. Instead, such behavior emerges through nonlinear dynamics, feedback mechanisms, and self-organization processes. Representative examples of complex systems include neuronal networks in the brain, ecological systems, financial markets, transportation networks, and social interaction structures [2]. The study of complex systems has revealed a set of universal principles, such as power-law distributions, self-organized criticality, emergent phenomena, and phase transitions. A deeper understanding of these principles contributes to the prediction and regulation of complex system dynamics, thereby enhancing system controllability and resilience against external perturbations and risks [3].

In recent years, the emergence and advancement of artificial intelligence, particularly Large Language Models (LLMs) such as GPT-4 and ChatGPT, have catalyzed significant progress in both natural and social sciences, introducing a novel paradigm for the analysis of complex systems [4]. LLMs possess powerful data processing and pattern recognition capabilities, enabling them to extract latent structures from large-scale datasets and facilitate the modeling and optimization of intricate systems [5]. Notably, Generative Agent-Based Models (GABMs), which integrate LLMs to simulate human behavior, have garnered increasing attention for their potential to capture complex interactions within diverse artificial environments. The application of GABMs enhances the fidelity of individual behavior representation and group dynamics modeling in complex systems, thereby offering innovative methodologies for social system analysis, policy evaluation, market simulation, and other domains.

This extensive literature review encompasses various study domains, including network science, evolutionary game theory, social dynamics, and epidemic modeling, offering a thorough and systematic analysis [target paper]. Lu et al. elaborated on how LLMs can augment the functionalities of agent-based modeling, which is highly progressive. The findings indicated that LLMs can emulate human-like behaviors, including fairness, cooperation, and adherence to social norms, while also possessing distinct advantages such as efficiency, scalability, and ethical simplification. The research elucidated the benefits of LLMs while also identifying the obstacles they encounter, including hallucination, prompt sensitivity, and unanticipated biases.

* Corresponding authors.
*E-mail addresses:* panxing@buaa.edu.cn (X. Pan), 202201106@hunnu.edu.cn (Y. Geng).

https://doi.org/10.1016/j.plrev.2025.04.003
Received 17 April 2025; Accepted 20 April 2025
Available online 21 April 2025

A critical challenge faced by Large Language Models in the implementation of complex systems is the issue of hallucination. Specifically, when the generated text diverges from the original intent (faithfulness) or contradicts factual information (factualness), it signifies the presence of hallucination-related errors [6]. As noted by Zhou et al. [7], the knowledge embedded within LLMs is primarily acquired during the pre-training phase. The presence of noisy data, including misinformation within the pre-training corpus, can undermine the parametric knowledge of LLMs, thereby exacerbating hallucination issues [8]. These phenomena pose a substantial obstacle to the reliability and trustworthiness of LLMs in real-world applications.

Existing research has extensively explored the identification, explanation, and mitigation of hallucinations [9–11]. A direct and effective approach involves integrating external information or tool-generated feedback with user queries before feeding them into LLMs for processing [12,13]. This method, often referred to as contextual knowledge integration, is not only straightforward to implement but also demonstrates significant efficacy in reducing hallucinations [14]. Studies indicate that LLMs exhibit strong in-context learning capabilities, enabling them to extract critical insights from contextual knowledge and rectify previously generated inaccuracies [15].

The second primary issue encountered by LLMs in the context of complex systems is prompt sensitivity, whereby minor alterations in the input can result in markedly divergent outcomes [16]. He et al. investigated the influence of various prompt formats on the efficacy of large language models [17]. Formatting the same context into various templates, including plain text, Markdown, JSON, and YAML, and assessing it with OpenAI's GPT model in tasks such as natural language inference, code generation, and translation revealed that the performance of GPT-3.5-turbo in code translation tasks fluctuates by as much as 40 % based on the prompt template used. Larger models, such as GPT-4, exhibited greater resilience to these alterations. The findings indicated that various prompt formats can substantially influence model performance, necessitating the selection of the suitable prompt format for distinct tasks. This could result in an absence of standardized specifications for the prompts utilized in various investigations, complicating the reproduction of experimental outcomes. Future research may utilize open datasets and benchmarks or develop standardized rapid engineering methodologies to mitigate experimental bias resulting from input variations and to maintain study uniformity.

A third critical challenge associated with Large Language Models is the presence of unintended biases. Bias in LLMs refers to the unjust or preferential perspectives embedded in the generated text, which may lead to skewed or discriminatory outputs [18]. Such biases typically originate from the training data, which encompass historical records, literary works, social media content, and various other textual sources. Given that these corpora often reflect prevailing societal norms and prejudices, LLMs may inadvertently perpetuate gender stereotypes, social discrimination, and inequitable beliefs [19]. To mitigate these biases, future research can explore strategies such as the development of debiased datasets and the implementation of fairness-aware model tuning. These approaches aim to enhance the neutrality and ethical robustness of LLMs, thereby ensuring their fair and responsible deployment in practical applications.

In summary, although LLMs have brought great convenience to complex systems research, they also present a range of challenges and limitations. Issues such as hallucinations, prompt sensitivity, and inherent biases can impact the reliability and fairness of their applications in complex systems. Therefore, future research should focus on developing methods to mitigate these drawbacks, such as enhancing contextual understanding, standardizing prompt engineering, and implementing fairness optimization techniques. By addressing these limitations, researchers can ensure that LLMs-driven complex system studies are conducted with greater scientific rigor, accuracy, and reliability, ultimately improving their practical applicability and trustworthiness.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This investigation was supported by the National Natural Science Foundation of China (Grant Nos. 72171008, 72071011 and 12301673). This work are supported by the Scientific Research Foundation of Yunnan Education Department (No.2025J0579) and the Yunnan Fundamental Research Projects (202501AU070193). This work are supported by the Fundamental Research Funds for the Central Universities and supported by the Innovation Fund of Xidian University (YJSJ25009).

## References

[1] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U. Complex networks: structure and dynamics. Phys Rep 2006;424(4–5):175–308.
[2] Sun G-Q, Li L, Pei Y-S. Advancing epidemic modeling: the role of LLMs and generative agent-based models comment on LLMs and generative agent-based models for complex systems research by Lu et al. Phys Life Rev 2025;52:175–7.
[3] Li X, Zhu Q, Zhao C, Duan X, Zhao B, Zhang X, et al. Higher-order granger reservoir computing: simultaneously achieving scalable complex structures inference and accurate dynamics prediction. Nat Commun 2024;15(1):2506.
[4] Zhao W.X., Zhou K., Li J., Tang T., Wang X., Hou Y., et al. A survey of large language models. arXiv preprint arXiv:230318223, 2023.
[5] Zhang Q., Ding K., Lyv T., Wang X., Yin Q., Zhang Y., et al. Scientific large language models: a survey on biological & chemical domains. arXiv preprint arXiv: 240114656, 2024.
[6] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1–38.
[7] Zhou C., Liu P., Xu P., Iyer S., Sun J., Mao Y., et al. Lima: less is more for alignment. arXiv preprint arXiv:230511206, 2023.
[8] Zhang Y., Li Y., Cui L., Cai D., Liu L., Fu T., et al. Siren's Song in the AI Ocean: a survey on hallucination in large language models. arXiv preprint arXiv: 230901219, 2023.
[9] Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Sci China Technol Sci 2020;63(10):1872–97.

[10] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 2020;21(1):5485–551.
[11] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824–37.
[12] Shi W., Min S., Yasunaga M., Seo M., James R., Lewis M., et al. Replug: retrieval-augmented black-box language models. arXiv preprint arXiv:230112652, 2023.
[13] Mallen A., Asai A., Zhong V., Das R., Khashabi D., Hajishirzi H. When not to trust language models: investigating effectiveness of parametric and non-parametric memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023, p.9802–22.
[14] Shi W., Han X., Lewis M., Tsvetkov Y., Zettlemoyer L., Yih S.W. Trusting your evidence: hallucinate less with context-aware decoding. arXiv preprint arXiv: 230514739, 2023b.
[15] Dong Q., Li L., Dai D., Zheng C., Ma J., Li R., et al. A survey on In-context learning. arXiv preprint arXiv:230100234, 2022.
[16] Chatterjee A., Renduchintala HSVNSK, Bhatia S., Chakraborty T. POSIX: a prompt sensitivity index for large language models. arXiv preprint arXiv:241002185, 2024.
[17] He J., Rungta M., Koleczek D., Sekhon A., Wang F.X., Hasan S. Does prompt formatting have any impact on LLM performance? arXiv preprint arXiv: 241110541, 2024.
[18] Wei X, Kumar N, Zhang H. Addressing bias in generative AI: challenges and research opportunities in information management. Information & Management 2025;62(2):104103.
[19] Navigli R., Conia S., Ross B. Biases in large language models: origins inventory and discussion. ACM J Data Inf Qual, 2023.