Opportunities and Pitfalls with Large Language Models for Biomedical Annotation

Cecilia Arighi
Department of Computer and Information Sciences, University of Delaware, Ammon-Pinizzotto Biopharmaceutical Innovation Building, 590 Avenue 1743, Newark, DE19713, US
Email: arighi@udel.edu

Jin-Dong Kim[1]
Database Center for Life Science (DBCLS), DS-ROIS, ROIS, 178-4-4 Wakashiba Kashiwa, Chiba 277-0871, Japan
Email: jdkim@dbcls.rois.ac.jp

Zhiyong Lu[2]
NCBI, NLM, NIH, Bethesda, MD 20894 Bethesda, MD20894, US
Email: zhiyong.lu@nih.gov

Fabio Rinaldi
IDSIA USI-SUPSI, Polo universitario Lugano - Campus Est, Via la Santa 1, CH-6962 Lugano - Viganello, Switzerland
Email: fabio.rinaldi@idsia.ch

Large language models (LLMs) and biomedical annotations have a symbiotic relationship. LLMs rely on high-quality annotations for training and/or fine-tuning for specific biomedical tasks. These annotations are traditionally generated through expensive and time-consuming human curation. Meanwhile LLMs can also be used to accelerate the process of curation, thus simplifying the process, and potentially creating a virtuous feedback loop. However, their use also introduces new limitations and risks, which are as important to consider as the opportunities they offer. In this workshop, we will review the process that has led to the current rise of LLMs in several fields, and in particular in biomedicine, and discuss specifically the opportunities and pitfalls when they are applied to biomedical annotation and curation.

Keywords: large language model; LLM; biomedical curation; generative AI; biomedicine and health; education; ethics.

[1] Work supported by the Database Integration Coordination Program Funced by NDBC of JST.
[2] This research is partly supported by the NIH Intramural Research, National Library of Medicine.

1. Background

High-quality, well-annotated biomedical data is crucial for training LLMs to understand and process scientific information. These annotations can include labeling entities (genes, proteins), relations (interactions), and other relevant information. By incorporating annotated data, LLMs can learn specific domain knowledge and improve their accuracy in tasks like information extraction, knowledge base creation, and text summarization. Diverse and unbiased annotations can help mitigate bias in LLMs, ensuring their outputs are fair and representative of the underlying data. At the same time, LLMs can be used to automate some aspects of annotation, such as identifying potential entities or suggesting relevant relations. This can significantly reduce the workload for human annotators. LLMs can identify areas of uncertainty in the data and suggest which annotations would be most valuable for improving their performance. This creates a feedback loop where LLMs guide the annotation process for optimal results. Finally, LLMs can be used to check the consistency and accuracy of annotations, identifying potential errors or inconsistencies.

A recent survey of LLMs for data annotation [1] describes how advanced large language models (LLMs), like GPT-4, can transform data annotation by automating and improving accuracy in this traditionally labor-intensive process. It categorizes the methods used for LLM-based data annotation, explores the effectiveness of LLM-generated annotations, and discusses learning strategies incorporating these annotations. The paper also highlights the challenges and limitations of using LLMs in this field, offering guidance for future research and development in automating data annotation. Goel et al [2] proposes a method that uses Large Language Models (LLMs) combined with human expertise to speed up medical text annotation for information extraction, significantly reducing human labor while maintaining high accuracy in generating labeled datasets. Several recent approaches exploit the in-context learning capabilities of LLMs based on a limited number of examples (few-shot) to create annotations, using suitably engineered prompts [3,4,5]. Other recent works discuss the usage of LLMs for knowledge distillation [6,7], or even how LLMs could themselves be used as evaluators [8]. Finally, several studies evaluate the reliability of the annotations generated by LLMs [9,10].

While opportunities with LLMs are actively being explored, it is equally important to be aware of the potential pitfalls that may arise during their use. The limitations and risks associated with using LLMs have been thoroughly examined in previous studies [11]. Some research has explored these challenges within the contexts of biology and medicine [12,13], offering more specific case studies and proposing mitigation strategies. These insights provide invaluable guidance that should be shared with researchers in the field to help avoid unnecessary risks and complications.

2. Workshop

The years 2022 and 2023 marked the emergence of Large Language Models (LLMs). Reflecting this pivotal shift, PSB2024 organized a workshop entitled "Large Language Models (LLMs) and ChatGPT for Biomedicine," aimed at providing introductory insights into LLMs within the realm of Biomedicine. In the meantime, a wealth of diverse experiences with LLMs has been accumulated, and the emphasis of the workshop will be on sharing these varied encounters. As such, presentations showcasing a spectrum of application cases of LLMs have been considered, encompassing both successful implementations and instances where expectations were not met.
The intention is to focus in particular on the impact of LLMs on biomedical annotation and curation. Some of the issues and questions to be addressed in the workshop include but not limited to:

- Are annotation and curation still necessary in the age of LLMs?
- Can LLMs replace those completely?
- How can we assess the quality of automated annotations?
- What are the limitations?

By addressing these challenges this workshop aims to clarify the potential and limits of LLMs in advancing biomedical research and knowledge discovery.

3. Conclusions

LLMs are already making major inroads in our social fabric, rapidly changing the way several highly skilled activities are performed, and leading to serious challenges to societal organization and profound questions about how to best make use of their capabilities for the advantage of humanity. We hope that this workshop will offer a valuable contribution to this ongoing discussion.

4. Acknowledgments

We are grateful to the PSB 2025 organizing committee for enabling us to organize this workshop.

References

1. Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., … Liu, H. (2024). Large Language Models for Data Annotation: A Survey. doi:10.48550/arXiv.2402.13446
2. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L.H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I. & Feder, A.. (2023). LLMs Accelerate Annotation for Medical Information Extraction. Proceedings of the 3rd Machine Learning for Health Symposium, in Proceedings of Machine Learning Research, 225:82-100 doi:10.48550/arXiv.2312.02296

3. Choi, J., Lee, E., Jin, K., & Kim, Y. (2024, March). GPTs Are Multilingual Annotators for Sequence Generation Tasks. In Y. Graham & M. Purver (Eds). Findings of the Association for Computational Linguistics: EACL 2024 (pp. 17–40). https://aclanthology.org/2024.findings-eacl.2
4. He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., … Chen, W. (2024, June). AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In Y. Yang, A. Davani, A. Sil, & A. Kumar (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 165–190). doi:10.18653/v1/2024.naacl-industry.15
5. Smith, R., Fries, J. A., Hancock, B., & Bach, S. H. (2024). Language Models in the Loop: Incorporating Prompting into Weak Supervision. ACM / IMS J. Data Sci., 1(2). doi:10.1145/3617130
6. Tan, S., Tam, W. L., Wang, Y., Gong, W., Zhao, S., Zhang, P., & Tang, J. (2023, July). GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model. In S. Sitaram, B. Beigman Klebanov, & J. D. Williams (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track) (pp. 134–148). doi:10.18653/v1/2023.acl-industry.15
7. Gu, Y., Dong, L., Wei, F., & Huang, M. (2024). MiniLLM: Knowledge Distillation of Large Language Models, The Twelfth International Conference on Learning Representations, https://openreview.net/forum?id=5h0qf7IBZZ.
8. Chiang, C.-H., & Lee, H.-Y. (2023). Can Large Language Models Be an Alternative to Human Evaluations?, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pages 15607–15631, https://aclanthology.org/2023.acl-long.870.
9. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30). doi:10.1073/pnas.2305016120
10. Honovich, O., Scialom, T., Levy, O., & Schick, T. (2022). Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14409–14428,, https://aclanthology.org/2023.acl-long.806
11. OpenAI. (2023) GPT-4 System Card. OpenAI Research Papers. https://cdn.openai.com/papers/gpt-4-system-card.pdf
12. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. (2024) Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med. 2024

Feb;177(2):210-220. doi: 10.7326/M23-2772. Epub 2024 Jan 30. https://www.acpjournals.org/doi/10.7326/M23-2772

13. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, Yang Y, Chen Q, Kim W, Comeau DC, Islamaj R, Kapoor A, Gao X, Lu Z. (2023) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023 Nov 22; 25(1):bbad493. doi:10.1093/bib/bbad493. https://academic.oup.com/bib/article/25/1/bbad493/7505071