Towards Reproducibility and Knowledge Transfer in AI-assisted Data Analysis Code Generation

Robert Haase[a,b,*]

[a]Data Science Center, Leipzig University, Humboldtstraße 25, 04105 Leipzig, Germany
[b]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden / Leipzig
ORCID (Robert Haase): https://orcid.org/0000-0001-5949-2327

Abstract.

Abstract will be added later

1 Introduction

LLMs changing the world. [3] [4] Typically human interacts with AI. Later nobody can reproduce if the human wrote the code or the AI. Solutions for tracking who did what: git, well established in the open source data analysis community. git-bob bridges both worlds: reproducible who did what by interacting with LLM via git and github issues / PRs. As LLMs are capable of solving github-issues [1] and write entire papers [2], such a solution is urgently need to be established as good scientific practice. AS prompts are documented in github-issues,

https://github.com/haesleinhuepf/git-bob

Figure 1. example interaction with git-bob - figure placeholder

2 Features and limitations

git-bob is implemented as github action, hence runs in the IT infrastructure of github.com. No local installation is required. No new graphical interface needs to be learned, it integrates well with preexisting workflows. It allows interaction of multiple humans with the LLM within the same context. E.g. an open source software user can reach out with a question about how to analyse an image, an expert can point out a wage strategy and the LLM implements the details, which can be reviewed be the user and the expert. it allows multi-turn interaction with the LLM using direct text input, additionally file input from a given repository and, when used with a vision language model, also image input. if used with a vision-language model such as OpenAI's GPT4-omni, an LLM that can take an opitional image as input, it can describe image content and with this, guide further analysis. can be configured for different purposes, such as assisting in code writing, [bio-image] data analysis, manuscript writing, code reviewing, but also answering questions of externals reaching out to developers of open source repositories. It can be used in private repositories giving scientists the necessary privacy to work on code and documentation before they eventually publish their work. For example, this manuscript was edited with LLM assistance in a private repository, and the reader can finally see which modifications were done by the human, and how the AI-assistant contributed to the work as shown in in Figure 1C.

3 Conclusion

LLMs are being integrated in contemporary scientific workflows unavoidably. To use LLMs responsibly, documenting how they were used in a specific project seems good-scientific-practice. git-bob allows facilitating this on multiple levels: for code generation, but also for manuscript writing. git-bob is the first implementation that is tightly integrated

Acknowledgements

RH acknowledges the financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI. NS acknowledges support by the BMBF (Federal Ministry of Education and Research) through ACONITE (01IS22065). JKH and CT thank EMBL IT services for their support with Kubernetes.

References

[1] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

[2] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[3] L. A. Royer. The future of bioimage analysis: a dialog between mind and machine. *Nature Methods*, 20(7):951–952, 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01930-y. URL https://doi.org/10.1038/s41592-023-01930-y.

[4] L. A. Royer. Omega – harnessing the power of large language models for bioimage analysis. https://doi.org/10.5281/zenodo.8240289, 2023.

[*] Corresponding Author. Email: robert.haase@uni-leipzig.de