Reflections on the Reproducibility of Commercial LLM Performance in Empirical Software Engineering Studies

Florian Angermeir

Maximilian Amougou, Mark Kreitz, Andreas Bauer, Matthias Linhuber, Davide Fucci, Fabiola Moyón C., Daniel Mendez, Tony Gorschek

Too long, didn’t listen

- Reproduction study of LLM studies
- 85 analyzed LLM studies at ICSE & ASE 2024 -> 69 used OpenAI
- 5 of 69 provided complete enough artefacts to repair and/or execute
- Of 5 we could reproduce 2 partially, 3 not

- Most reproducibility challenges coincide with ”traditional research”
- Deprecation of commercial models is a prevalent challenge

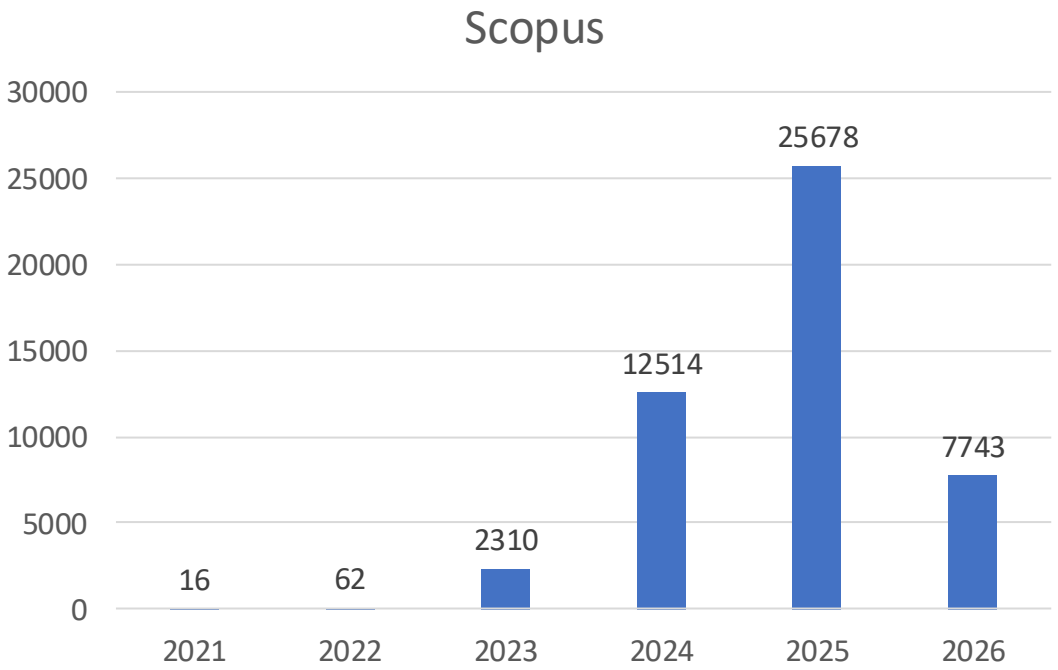
Too long, didn’t listen

- Reproduction study of LLM studies

➔ LLM-studies show similar issues as qualitative studies
➔ For qualitative studies we can rarely reproduce the same context/subject configuration
➔ This restriction does not hold for LLM-centric empirical studies

- Deprecation of commercial models is a prevalent challenge

Motivation



Query: “large language models”

” […] we found that our results are neither reproducible nor reliable when rerunning the same prompts with different seeds […]” [1]

”[…] due to the non-deterministic nature of LLMs, it is unlikely that exact replication of our results will be possible. However, similar results should be obtainable” [2]

“[…] conducting replication experiments using the same models as in the original study, the opacity surrounding model updates […] renders […] replications difficult” [3]

[1] Staudinger, M., Kusa, W., Piroi, F., Lipani, A., & Hanbury, A. (2024). A reproducibility and generalizability study of large language models for query generation.
[2] Muhammad A. A. Pirzada, Giles Reger, Ahmed Bhayat, and Lucas C. Cordeiro. 2024. Llm-generated invariants for bounded model checking without loop unrolling.
[3] Vaugrante, L., Niepert, M., & Hagendorff, T. (2024). A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions

Motivation

Ben Hermann: What We Learned in 15 Years of Artifact Evaluation, ICSE 2026

Disclaimers

1. In this talk reproduction and replication are used interchangeably.

2. Our ambition is not to criticise individual contributions. Their work was guided by what may generally be viewed as high scientific standard. This is a critical reflection on the overall state of reproducibility of LLM-centric SE research in the community.

3. The data we analysed is from 2024, the field is changing fast.

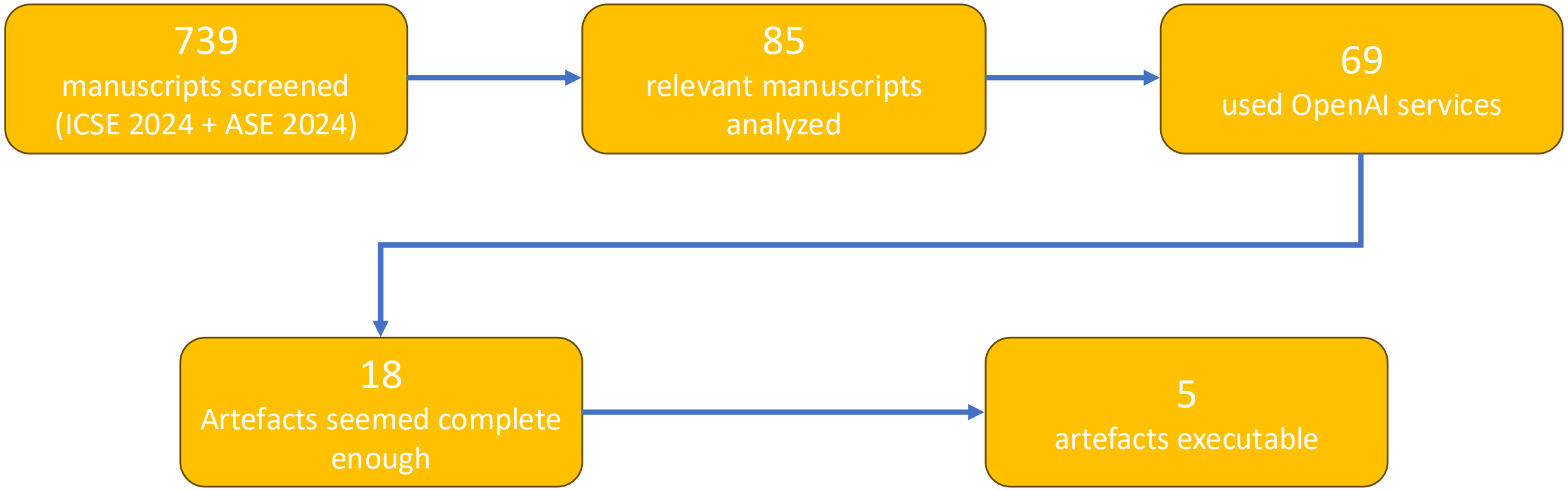
Research Questions

RQ1: To what extend are LLM-centric SE studies reproducible?

RQ2: Which factors impede the reproduction of LLM-centric empirical SE studies?

RQ3: How well do the ACM artefact badges reflect the state of reproducibility of LLM-centric SE studies?

Study Design



Methodology

- Five researchers built containerized reproduction framework
- Attempted to execute studies independently
- 30 repetitions or less due to $500 cap, sometimes subset of experiments
- Bayesian bootstrap & descriptive analysis of results

- Replicable if reported results are in 95% interval

Results: Extend of Reproducibility

- None of five studies could be fully reproduced
- 2 studies could be partially reproduced
- 3 studies could not be reproduced

Study / Metric	Reported Value	2.5th Percentile	97.5th Percentile
Study A / Metric A	79.9	45.6	55.3
Study A / Metric B	85.5	50.1	54.5
Study B / Metric A	16.38	14.51	15.49
Study B / Metric B	16.96	18.95	19.63

Results: Reproduction Challenges

Traditional Problems

- Incomplete Artefacts
- Dependency Version Issues
- General Code Issues
- Incomplete Documentation
- Non-Executable Artefacts
- Lacking Error Handling
- Deprecated External API

LLM Specific Problems

- Deprecated Models (27%)

	Not reported	Not configured
Model Name	9%	-
Model Version	-	56%
Temperature	65%	17%
Configuration Values	84%	44%
Context Window Handling	74%	72%

Results: ACM Badges & Reproducibility

[1]

[1]

Badge	n	Unavailable	Incomplete	Non-functional	Documentation
Only **Artefact Available v1.1**	3	1	1	-	-
Artefact Evaluated – Reusable v1.1	15	-	4	3	1

Requirements for **Artefact Evaluated – Reusable** badge:

- Complete
- Consistent
- Exercisable
- Careful documentation

[1] https://www.acm.org/publications/policies/artifact-review-and-badging-current

Recommendations

- Authors: LLM-Guidelines (llm-guidelines.org)
- Venues: Continue strengthening artefact requirements & evaluation processes
- Funding agencies: Incentivize long-term accessibility and reproducibility

Thank you very much!

Florian Angermeir
angermeir@fortiss.org
angermeir.me