Empirical and **Sustainability** Aspects of Software Engineering Research in the Era of **Large Language Models**: A Reflection

David Williams
University College London

Max Hort
Simula Research Lab

Maria Kechagia
University of Athens

Aldeida Aleti
Monash University

Justyna Petke
University College London

Federica Sarro
University College London

Motivation

1. LLM-based SE Research is Moving Fast

- Surge in research pace since 2022
- Urgency to “be the first”

2. Replicability & Empirical Rigour

Are researchers:

- still considering traditional SE techniques?
- including enough information to make their work replicable?

3. Sustainability

We need to consider:

- Is LLM-based research accessible?
- Are some institutions being left behind?

Scope & Method

1. Retrieving all papers published in ICSE main track between 2023-2025 (total of 692).
2. Filtering papers based on AI-related keywords.
3. Manual selection of **empirical studies featuring LLMs**.
4. Extracting information of **177 papers** and a survey based on the following research questions...

Research Questions

From the 177 LLM-based empirical studies, we answer the following:

*Full paper!

RQ1: Which LLMs are used in SE research and how are they benchmarked?

- Open vs. commercial models
- Which model families?*
- Programming languages*
- Are non-LLM baselines featured?

RQ2: How well do authors tackle the problem of data leakage/contamination?

- Mention of contamination
- Mitigation strategies

RQ3: How replicable are LLM-based studies?

- Mention of model configuration/parameters
- Artefact availability/badges*

RQ4: What are the costs of LLM-based SE research?

- Mention of costs
- Survey distributed to ICSE authors about costs*

Finding #1:
In the past 3 years, the proportion of LLM-based research at ICSE has doubled.

Total number of accepted papers and LLM-based empirical studies in the ICSE main track 2023-2025.

ICSE	# Papers	# LLM SE
2023	210	32 (15.2%)
2024	236	55 (23.3%)
2025	246	90 (36.6%)
Total	692	177 (25.6%)

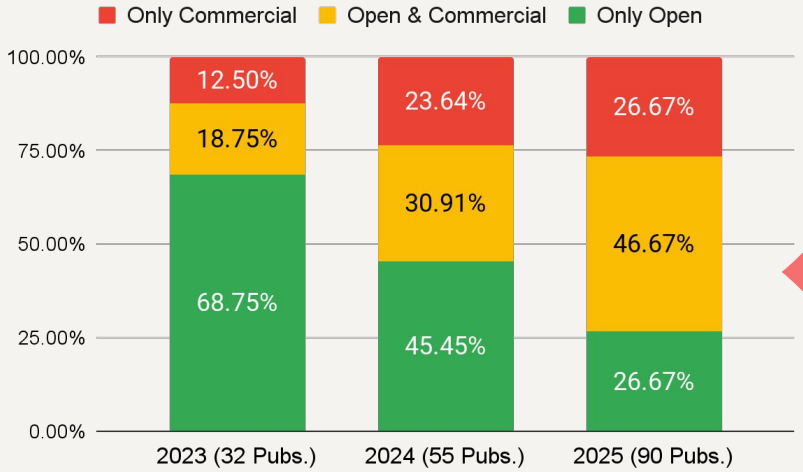
15.2% → 36.6% = ↑ 2.41x from 2023 to 2025

RQ1: Models & Benchmarking

RQ1: Models & Benchmarking

Finding #2:

Commercial models are becoming more prevalent.

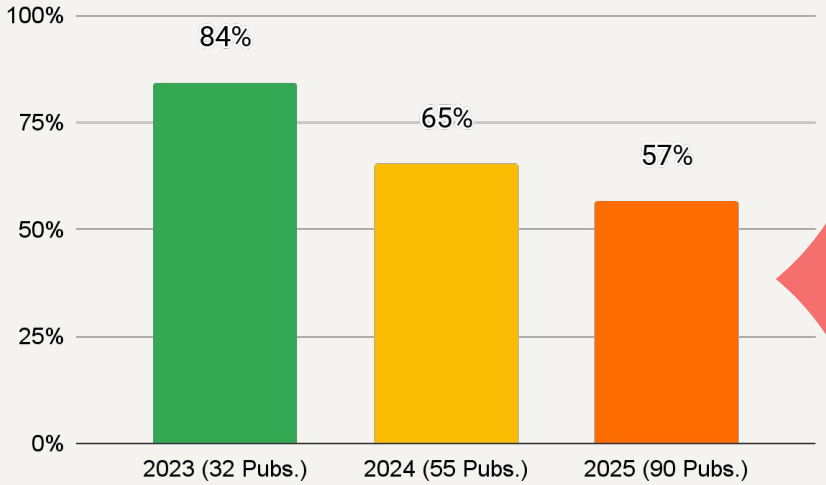


Proportion of papers using only commercial vs. only open vs. both types of models.

RQ1: Models & Benchmarking

Finding #3:

Benchmarking against non-LLM techniques is becoming less popular.



Proportion of papers including non-LLM SE baselines in their evaluations.

RQ2: Contamination

RQ2: Contamination

Finding #4:

Few papers mention contamination.

Finding #5:

Several techniques have been proposed to mitigate contamination.

Reporting

2025: 38 out of 90 (**42.2%**)
2024: 14 out of 55 (**25.5%**)
2023: 6 out of 32 papers (**18.8%**)

Mitigation Strategies

(Within the papers that mention contamination)

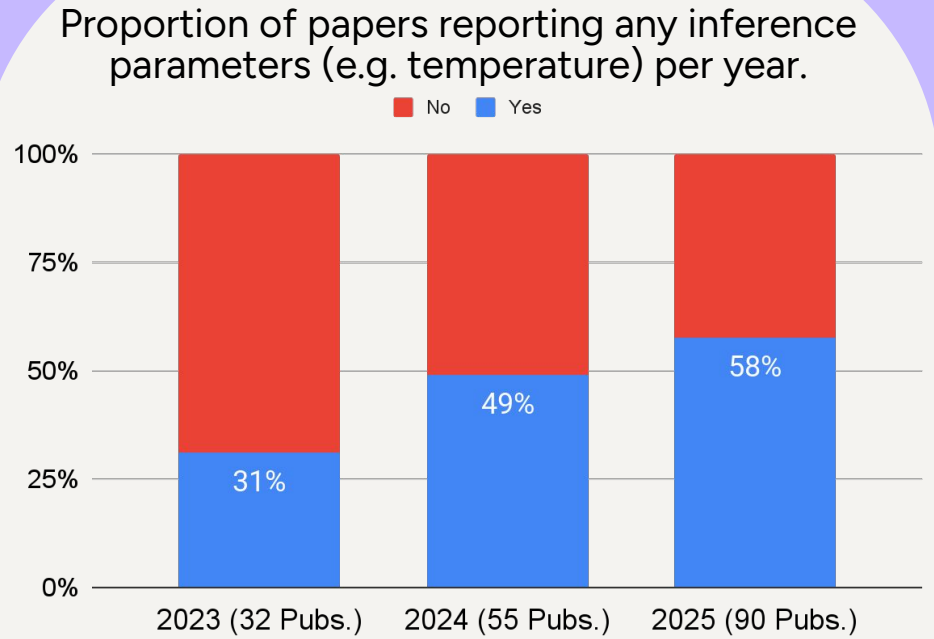
- Temporal filtering
- Code obfuscation
- Multi-dataset evaluation
- Ablation studies
- (Sometimes) None!

RQ3: Replicability

RQ3: Replicability

Finding #6:

Despite improvements, barely half of papers report on inference parameters.



Overall : 50.3% report inference parameters.

RQ4: Sustainability

RQ4: Sustainability

Finding #7:

Costs are rarely reported, and researchers have concerns about sustaining them.

Cost Reporting

Cost Type	# Papers (Prop.)
Hardware	89 (50%)
Time	36 (20%)
Financial	18 (10%)
In/Out Tokens	12 (7%)
Energy/CO2	None

User Study (57 Authors)

"How likely are you to keep using __ models in the next 12 months?"

- Commercial: 89%
- Open: 95%

"Will you be able to continue sustaining the costs of __ models in the next 12 months?"

- Commercial:
  - 56% - "Uncertain"
  - 9% - "No"
- Open:
  - 57% - "Yes"

Thank you for listening! In summary…

Finding 1: Over 3 years, the proportion of LLM-based research at ICSE has doubled.

Finding 2: Papers with only commercial models are increasing (12.5% → 26.6%).

Finding 3: Non-LLM baselines are becoming less popular (84% in 2023 to 57% in 2025).

Finding 4: Less than half of papers (42% in 2025) mention data leakage/contamination.

Finding 5: Approaches for leakage (e.g. temporal filtering, obfuscation, ablation studies).

Finding 6: Barely half (50.3%) of papers report *any* inference parameters.

Finding 7: Costs reporting is rare ($ - 10%, energy - 0%), concerns about future costs.

Insights & guidance in the full paper!

https://arxiv.org/abs/2510.26538

Contact me!

david.williams.22@ucl.ac.uk
https://davejjwilliams.github.io