How-to:
SE Research with LLMs

Dave Williams, 26th November 2025

About Me

2nd Year PhD Student @ UCL
Supervised by Prof. Federica Sarro

Research Interests:

- Human factors of AI Adoption
- (AI) Developer Productivity
- Code Review

Prior Work:

- User-Centric Deployment of APR @ Bloomberg
- Empirical PCA Evaluation
- **Empirical and Sustainability Factors in LLM-based SE Research (in this talk!)**

In this talk...

Chapters

1. The LLM4SE Benchmarking Landscape
2. Looking Closer at LLM Usage in ICSE
3. Looking Forward: Guidelines for Empirical Studies

Chapter 0: Increasing Impact, Helpful and Harmful

How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference

Nidhal Jegham[1 2] Marwan Abdelatti[1 3] Lassad Elmoubarki[2] Abdeltawab Hendawi[1 *]

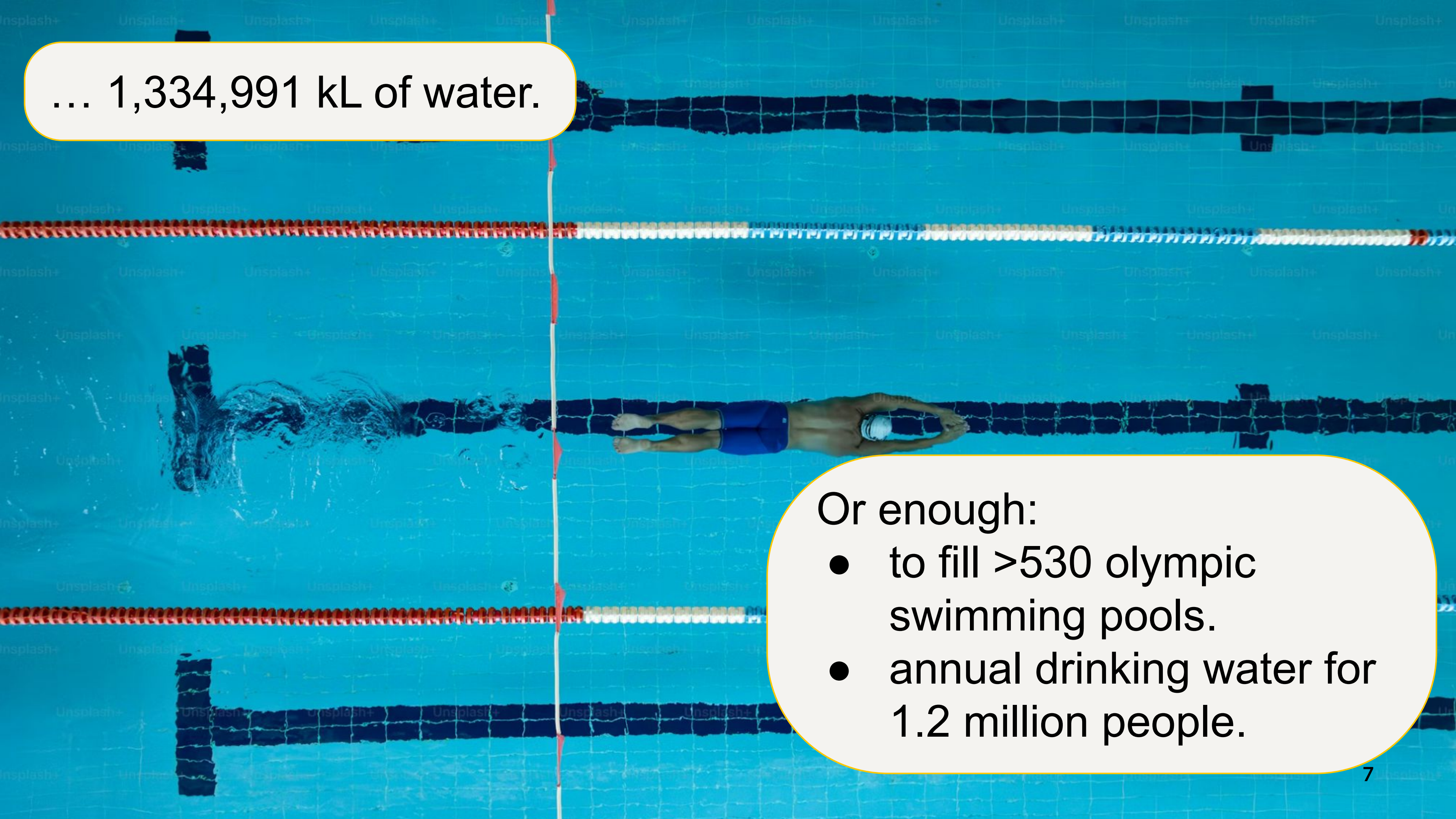
[1]University of Rhode Island [2]University of Tunis [3]Providence College

{nidhal.jegham, hendawi}@uri.edu lassad.elmoubarki@tbs.rnu.tn mabdelat@providence.edu

In a case study estimating the environmental impact of GPT-4o, they found that in a year (or 772 billion queries), GPT-4o uses…

… 391,509 MWh of electricity.

Or enough to power >145,000 UK homes.

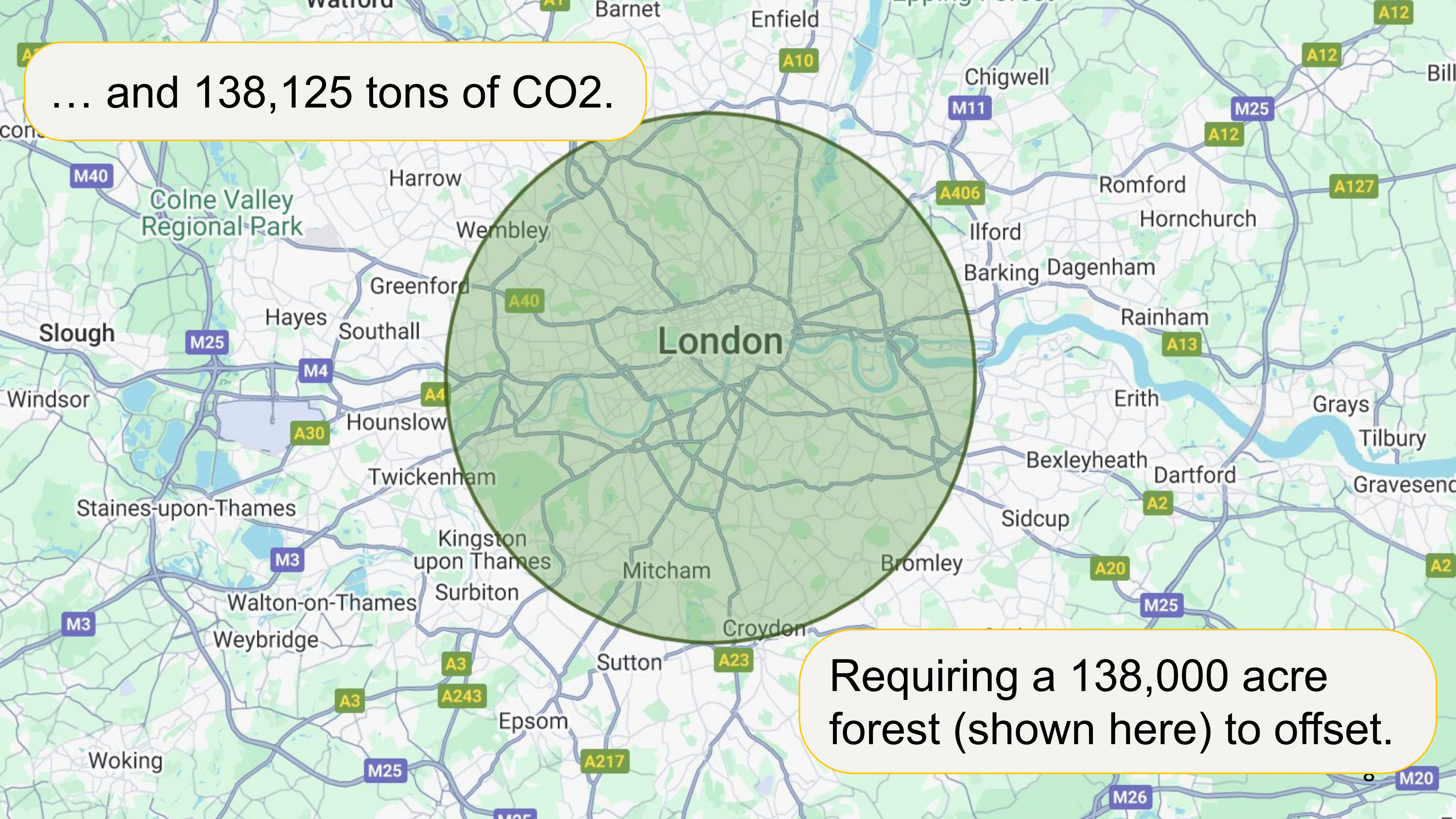


… 1,334,991 kL of water.

Or enough:

- to fill >530 olympic swimming pools.
- annual drinking water for 1.2 million people.

… and 138,125 tons of CO_2.



Requiring a 138,000 acre forest (shown here) to offset.

My (Personal) Thoughts

Don't swat flies with hammers

- "Smaller" LLMs have become extremely performant for many SE tasks, so don't kill your wallet (and the planet) using commercial models if they aren't necessary.

People first, not innovation

- If you are making a practical contribution (tool, technique or approach), don't guess what end users might want. (Before running large-scale synthetic experiments,) your work can always benefit from direct feedback from your target audience!

There's no need to reinvent the wheel

- Many non-LLM alternatives already exist to solve SE problems. While LLMs could potentially stand to improve the state-of-the-art in some domains, consider (and measure) their impacts (e.g. environmentally).

But, if you must use LLMs...

Chapter 1: The LLM4SE Benchmarking Landscape

Surveying the Benchmarking Landscape of Large Language Models in Code Intelligence

MOHAMMAD ABDOLLAHI, York University, Canada
RUIXIN ZHANG, York University, Canada
NIMA SHIRI HARZEVILI, York University, Canada
JIHO SHIN, York University, Canada
SONG WANG, York University, Canada
HADI HEMMATI, York University, Canada

Study Scope

General Criteria

Timeframe: Jan. 2020 - Jun. 2025
Study Types: Peer-reviewed (+ recent ArXiv)
Inclusion Criteria:
- Must be applying LLMs for code intelligence tasks.
- Study should present a benchmark dataset.

Exclusion Criteria:
- Do not involve LLM-based evaluation techniques or a lack of focus on benchmark datasets.

Final Result

The authors identified **142 papers** covering **156 unique benchmarks**.

Research Questions

The authors examined 142 papers to investigate...

RQ1: Current Landscape of Benchmark Datasets

Age, task types & complexity, size, tasks, and programming languages.

RQ2: Characteristics and Quality of Benchmark Datasets

Dataset Structure: Formats, labelling schemes and metadata.

RQ3: Evaluation Metrics and Techniques

Evaluation criteria, metrics, task alignment, and consistency across datasets.

RQ4: Challenges and Limitations

Examining limitations highlighted by dataset users for a subset of 14 datasets.

RQ5: Future Directions and Improvements

Recommendations for crafting more complete and realistic datasets, as well as new tasks, multi-modal approaches, etc.

RQ1: What Is the Current Landscape of Benchmark Datasets?

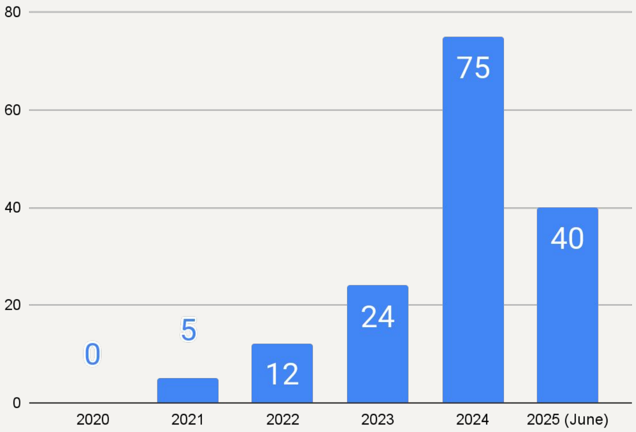
Focus:

- Release timeline
- Programming languages
- Publication venues
- Tasks covered
- Data sources & sizes

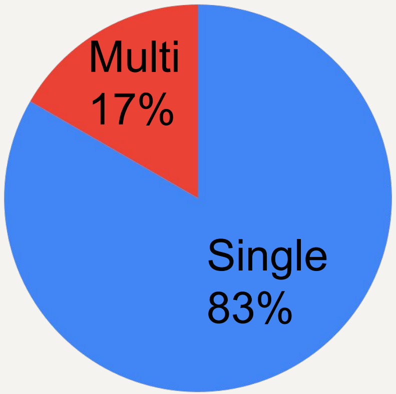
RQ1: Landscape Overview

156 benchmark datasets across **142** papers.

Number of Benchmarks per Year



Single vs. Multi Task Datasets



Programming Languages

Language	# Studies	# Datasets
Python	106	120 (**77%**)
Java	51	59 (38%)
C++	29	37 (24%)

...

48 unique programming languages

Venues

Venue Topic	# Datasets	
	Conf.	Journ.
ML/AI	**36**	2
NLP	**30**	0
SE	21	6
Other	2	2
ArXiv	**57**	

Venues are important to note, since e.g. ML/AI typically focus less on realism than SE.

Many benchmarks are first published on ArXiv.

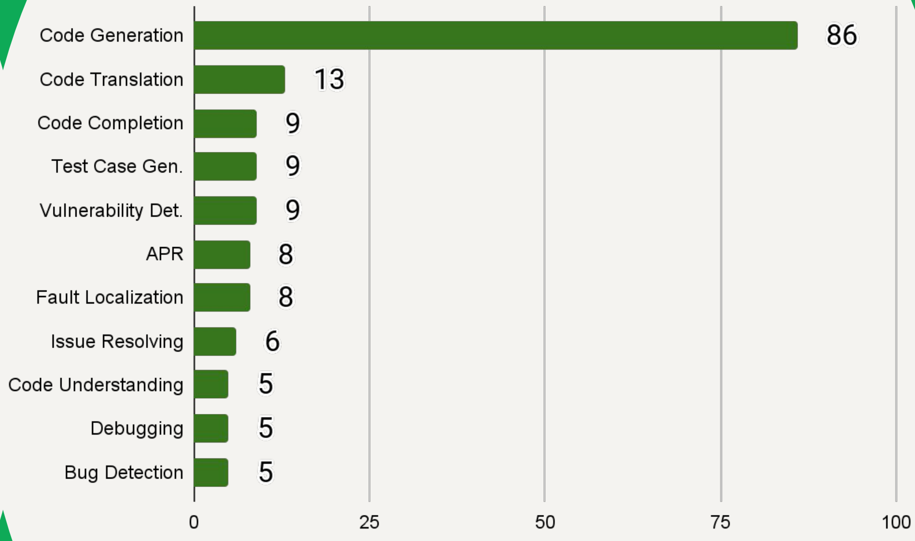
RQ1: Tasks

32 distinct code intelligence tasks identified

Code generation dominates (86 datasets, 55%)

Other common tasks:

- Code translation
- Code completion
- Test generation
- Program repair & debugging
- Vulnerability detection, classification, summarization



Datasets per Task (for tasks with ≥5 datasets)

RQ1: Sizes

Mean: 1631641
Median: 759
Standard Deviation: 15987155

Dataset sizes vary massively, from as few as 14 samples (ProjectDev) to as many as 189 million (DynaCode).

Dataset Size	Frequency
Small (<500 Samples)	64 (**41%**)
Medium (500-5k Samples)	55 (35%)
Large (>5k Samples)	37 (24%)

RQ2: Characteristics and Quality of Benchmark Datasets

RQ2: Sources & Quality

Data Sources

- **66 (46%) benchmarks come from GitHub**
    - SWE-bench, RepoEval, …
- 49 (35%) crafted manually
    - HumanEval, MBPP, …
- 35 (25%) from competitive coding platforms
- 32 (22.5%) from other existing benchmarks
    - e.g. HumanEval+, SWE-Bench+

Quality Control Strategies

- Manual inspection (32%)
- Automated filtering (e.g. deduplication)
- Hybrid approaches
- **43% of benchmarks had NONE**

"Realism" Strategies

- 49% - Sourcing from "real-world" data
- 18% - Workflow-oriented problems

RQ3: Evaluation Metrics and Techniques

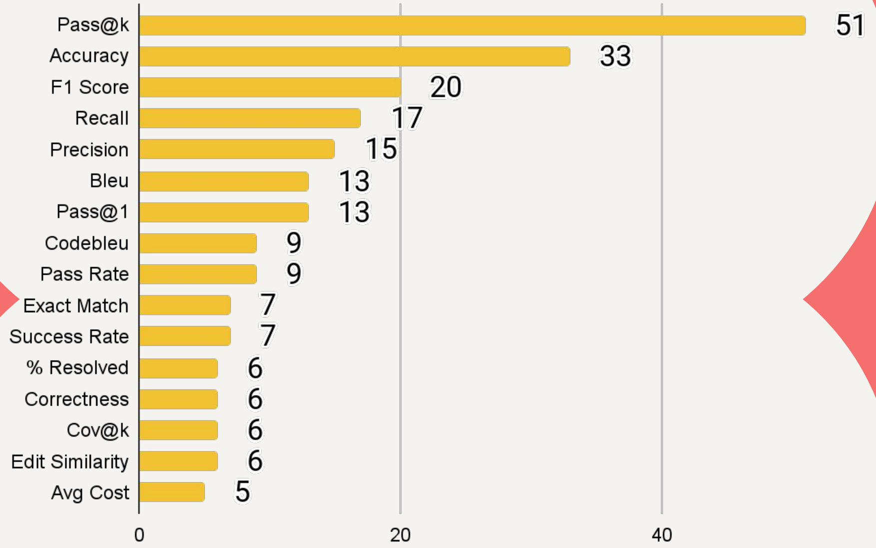
RQ3: Metrics

Pass@k stands out as the most popular (36%).

The authors identified **74 unique metrics**.

Abstracting a little, several strategies emerge:

- Execution-based (e.g. Pass@k, Runtime)
- Similarity checking (e.g. BLEU, BERTScore)
- Human feedback
- LLM-based (i.e. LLM-as-a-Judge)



 # Studies per metric (for metrics in ≥5 studies)

RQ4: Shortcomings & Limitations

1. Limited Task Complexity and Real-World Relevance

- Too much focus solely on simplified algorithmic or competitive programming tasks.
- Many are limited to **single-function** or **single-file contexts**.

2. Data Quality and Bias Issues

- Many rely on LLMs to generate the tasks with little supervision.
- Lack of difficulty distribution.
- Ambiguous task descriptions.

3. Inadequate Test Coverage

- Popular benchmarks such as HumanEval have been found to have incorrect canonical solutions.
- Weak test coverage leads to plausible but incorrect solutions.

4. Evaluation Limitations

- Most benchmarks only evaluate one aspect of generated code (i.e. correctness), ignoring others.
- Synthetic evaluation pipelines do not accurately reflect the complexity of real-world software.

5. Data Leakage Risks

- The validity of benchmarks using popular open-source repositories can be compromised if their contents are in an evaluated model's training set.

RQ5: Future Directions

The authors identify 6 areas to focus on:

- Multimodal
- Domain-specific
- Large-scale/realism
- Assessing reasoning
- Cutoff aware
- Dynamic

Picking a Benchmark

More details in the full paper!

1. Find out what's out there

- Use surveys like this one to identify all benchmarks aligning with your research questions.

2. Check data quality

- How do the benchmark creators ensure the quality of the tasks?

3. Consider evaluation techniques

- What metrics will best convey your arguments?

4. Consider contamination

- Will contamination be a threat to the validity of your study if you use this benchmark?

One More Tip: Living Surveys, e.g.

🤖 Awesome Code Agents

Towards AI-Powered Software 3.0

A curated list of products, benchmarks, and research papers on autonomous code agents.

Beyond coding — they're redefining how software changes the world.

Website · X/Twitter · LinkedIn · Discord · Reddit · GitHub

STARS 15 | FORKS 0 | LICENSE APACHE-2.0 | PAPER ARXIV | CONTRIBUTORS 3

awesome | papers 372 | Commits 44/month | Forks 0 | Issues Closed 0 | Discussions 0 total | PRs Welcome | last commit today

Chapter 2: Looking Closer at LLM Usage in ICSE

Reflecting on Empirical and Sustainability Aspects of Software Engineering Research in the Era of Large Language Models

David Williams
University College London
United Kingdom
david.williams.22@ucl.ac.uk

Max Hort
Simula Research Laboratory
Norway
maxh@simula.no

Maria Kechagia
National and Kapodistrian University of Athens
Greece
makechag@ba.uoa.gr

Aldeida Aleti
Monash University
Australia
aldeida.aleti@monash.edu

Justyna Petke
University College London
United Kingdom
j.petke@ucl.ac.uk

Federica Sarro
University College London
United Kingdom
f.sarro@ucl.ac.uk

Motivation

LLM-based SE Research is Moving Fast

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

Replicability & Empirical Rigour

Are researchers:

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

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

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

- Open vs. commercial
- Which families?
- Non-LLM baselines
- Programming languages

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 configuration/ parameters
- Artefact availability/badges

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

- Mention of costs
- Survey distributed to ICSE authors

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

Table: Num. papers in ICSE main track 2023-2025

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

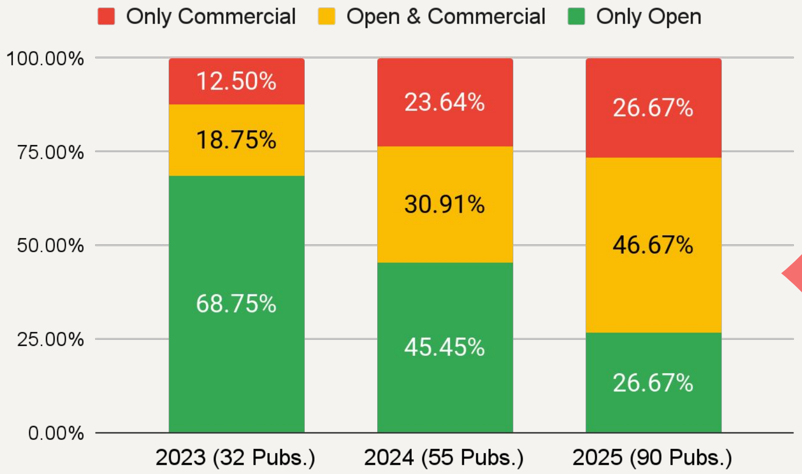
15.2% → 36.6% = ↑ 2.41x
from 2023 to 2025

RQ1: Models & Benchmarking

RQ1: Models & Benchmarking

Finding #2:

Commercial models are becoming more prevalent.

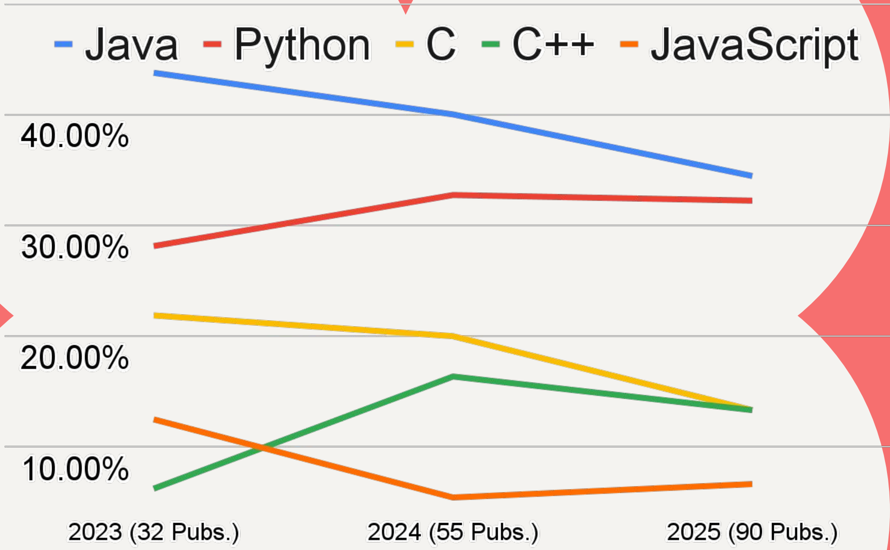


Prop. of papers using only commercial vs. only open vs. both types of models.

RQ1: Models & Benchmarking

Finding #3:

Languages targeted are shifting (towards Python).

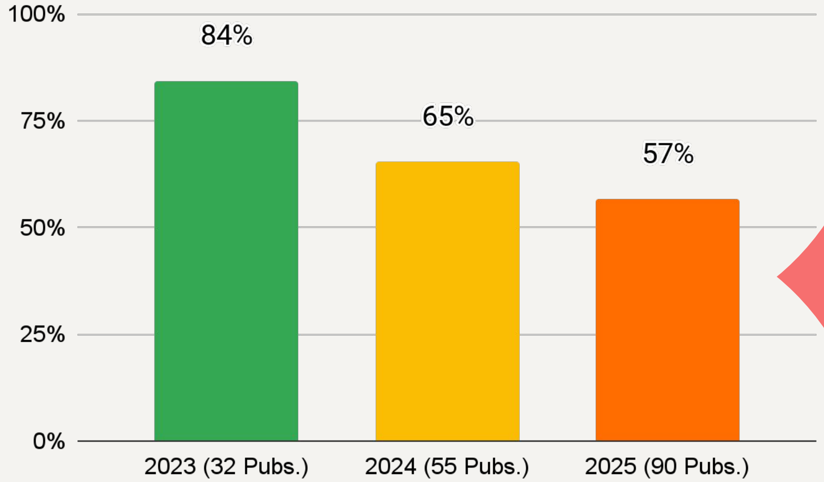


Prop. of papers targeting top 5 PLs per year.

RQ1: Models & Benchmarking

Finding #4:

Benchmarking against non-LLM techniques is becoming less popular.



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

RQ2: Contamination

RQ2: Contamination

Finding #5:

Less than half of papers mention contamination.

Finding #6:

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:)

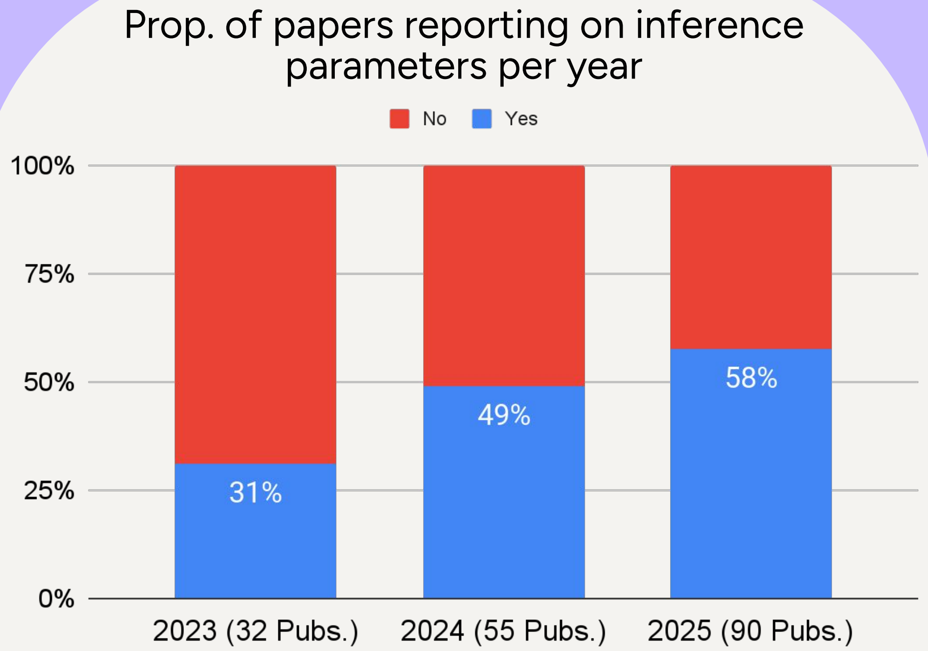
- None!
- Temporal filtering
- Code obfuscation
- Multi-dataset evaluation & ablation

RQ3: Replicability

RQ3: Replicability

Finding #7:

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



Overall : 50.3% report inference parameters.

RQ4: Sustainability

RQ4: Sustainability

Finding #8:

Costs are rarely reported, and researchers are nervous 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?"

- Commercial:
  - 65% "Uncertain"
  - 9% "No"
- Open:
  - 65% "Yes"

If I had to give a few suggestions…

Read more in the ArXiv Preprint!

Benchmark using open models (too).

- Will your results still be replicable when commercial APIs deprecate the closed model you used?
- Other researchers may be interested in your work, but may not have the finances to try it out.

If a non-LLM technique exists, try it!

- In using LLMs, are you ignoring a rich prior literature on viable non-LLM techniques?

Report every parameter/prompt you can.

- How can anyone replicate your work if they can't configure the models in the same way?
- If there is no room due to conference page limitations, include these in a separate doc reporting on parameters, prompts used, etc.

If you can, measure and report costs!

- Help future researchers decide if your technique is financially/computationally viable for their work.

But, in case you don’t want to take my advice…

Chapter 3: Guidelines for Empirical Studies

Guidelines for Empirical Studies in Software Engineering involving Large Language Models

Sebastian Baltes
University of Bayreuth, Germany
sebastian.baltes@uni-bayreuth.de

Florian Angermeir
fortiss, Germany
BTH, Sweden
angermeir@fortiss.org

Chetan Arora
Monash University, Australia
chetan.arora@monash.edu

Marvin Muñoz Barón
Chunyang Chen
TU Munich, Germany
{marvin.munoz-baron,chunyang.chen}@tum.de

Lukas Böhme
Hasso-Plattner-Institut, Germany
University of Potsdam, Germany
lukas.boehme@hpi.de

Fabio Calefato
University of Bari, Italy
fabio.calefato@uniba.it

Neil Ernst
University of Victoria, Canada
nernst@uvic.ca

Davide Falessi
University of Rome Tor Vergata, Italy
falessi@ing.uniroma2.it

Brian Fitzgerald
Lero, Ireland
University of Limerick, Ireland
brian.fitzgerald@ul.ie

Davide Fucci
BTH, Sweden
davide.fucci@bth.se

Marcos Kalinowski
PUC Rio de Janeiro, Brazil
kalinowski@inf.puc-rio.br

Stefano Lambiase
Daniel Russo
Aalborg University, Denmark
{stla,daniel.russo}@cs.aau.dk

Mircea Lungu
IT University Copenhagen, Denmark
mlun@itu.dk

Lutz Prechelt
Freie Universität Berlin, Germany
prechelt@inf.fu-berlin.de

Paul Ralph
Dalhousie University, Canada
paulralph@dal.ca

Rijnard van Tonder
Independent, Antigua and Barbuda
rvantonder@gmail.com

Christoph Treude
SMU, Singapore
ctreude@smu.edu.sg

Stefan Wagner
TU Munich, Germany
stefan.wagner@tum.de

Approach & Scope

- Collaborative effort starting at ISERN 2024, followed by a position paper at WSESE 2025.
- Authors focus on textual models.

Taxonomy of LLM-Based SE Study Types

1. LLMs as Tools for SE Researchers
   a. LLMs as Annotators
   b. LLMs as Judges
   c. LLMs for Synthesis
   d. LLMs as Subjects
2. LLMs as Tools for Software Engineers
   a. Studying LLM Usage in SE
   b. LLMs for New SE Tools
   c. Benchmarking LLMs for SE Tasks

8 *Must/Should* Guidelines for Using LLMs in Empirical Studies in SE

Guideline #1: LLM Usage and Role

***MUST* report**:
Whether an LLM was used at all

***SHOULD* report**:
The purpose, automated tasks and expected benefits.

Guideline #2: Model Version, Configuration, & Customisations

Example:
"We integrated a gpt-4 model in version 0125-Preview via the Azure OpenAI Service, and configured it with a temperature of 0.7, top_p set to 0.8, a maximum token length of 512, and the seed value 23487. We ran our experiment on 10th January 2025 (system fingerprint fp_6b68a8204b).

MUST **Report**

- Exact LLM model/tool version.
- Configuration parameters
- Experiment dates

If fine-tuning:

- Fine-tuning goals
- Datasets
- Procedure

SHOULD **Report**

- Default parameters
- Reasoning for model choices
- Comparisons of base and fine-tuned models
- Fine-tuning data & weights

Guideline #3: Tool Architecture Beyond Models

MUST Report

- Full architecture of novel LLM-based tools.
- Hosting setup/hardware
- Confidential/proprietary components (as a threat to reproducibility)

If autonomous agents are used:

- Agent roles
- Reasoning frameworks
- Communication flows

SHOULD Include

- Architectural diagrams
- Justification for design decisions

Guideline #4: Prompts, their Development, and Interaction Logs

MUST **Report**

- ALL PROMPTS
  - Structure, formatting, dynamic components (variables)
- Token optimisation techniques
- Prompt reuse across models and configs

For dynamically/user generated* prompts:

- Generation and collection process

SHOULD **Report**

- Prompt revisions
- Pilot-testing insights
- Full interaction logs (prompts and responses)*

*if privacy and confidentiality can be ensured

Guideline #5: Human Validation for LLM Outputs

SHOULD **Report**

- Consideration of human validation early in the study design
- Measuring Instruments
- Results of a statistical power analysis
- Mitigation of confounding factors

When aggregating LLM judgements:

- Methods and rationale
- Inter-rater agreements

Guideline #6:
Use an Open LLM as a Baseline

When using commercial LLMs, authors…

SHOULD **Report**

- Results using an open LLM as a baseline
- Inter-model agreement
- Full step-by-step replication instructions as part of the supplementary material.

Guideline #7: Suitable Baselines, Benchmarks, and Metrics

MUST **Report**

- Justification for the choice of benchmark.
- Why the metrics are suitable for the specific study.

SHOULD **Report**

- Summary of benchmark structure, task types, limitations.
- Results of non-LLM baselines.
- Results of experiment repetitions AND result distribution.

Guideline #8: Limitations and Mitigations

MUST Report

- Study limitations:
  - Impact of non-determinism
  - Generalisability constraints
- Whether generalisation across models was measured (& differences observed)
- Model outputs
- Sensitive data handling, ethics approvals
- Justification for using LLMs at all

Guidelines in a Nutshell

Full paper for
in-depth examples!

1. Say if you're actually using LLMs.
2. Versions, configurations, and customisations.
3. How the LLM interacts with other components.
4. What prompts did you use?
5. Involve people to validate LLM outputs.
6. Use open models (as well as commercial)!
7. Pick your benchmarks (and metrics) wisely.
8. Highlight limitations (contamination, generalisability, etc.) and how you mitigated them.

Thank You!

Dave Williams

davejjwilliams.github.io
david.williams.22@ucl.ac.uk

Paper 0: How Hungry is AI?

Paper 1: Benchmarking LLM4SE Landscape

Paper 2: Reflecting on LLMs in ICSE 2023-2025

Paper 3: Guidelines for LLM-based SE Research