ChatGPT has been globally **exposed** to at least **4.7M samples** from **263 benchmarks**

Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

Simone Balloccu   Patrícia Schmidtová   Mateusz Lango   Ondřej Dušek
{balloccu, schmidtova, *lango, odusek*}*@ufal.mff.cuni.cz*

CHARLES UNIVERSITY   ÚFAL

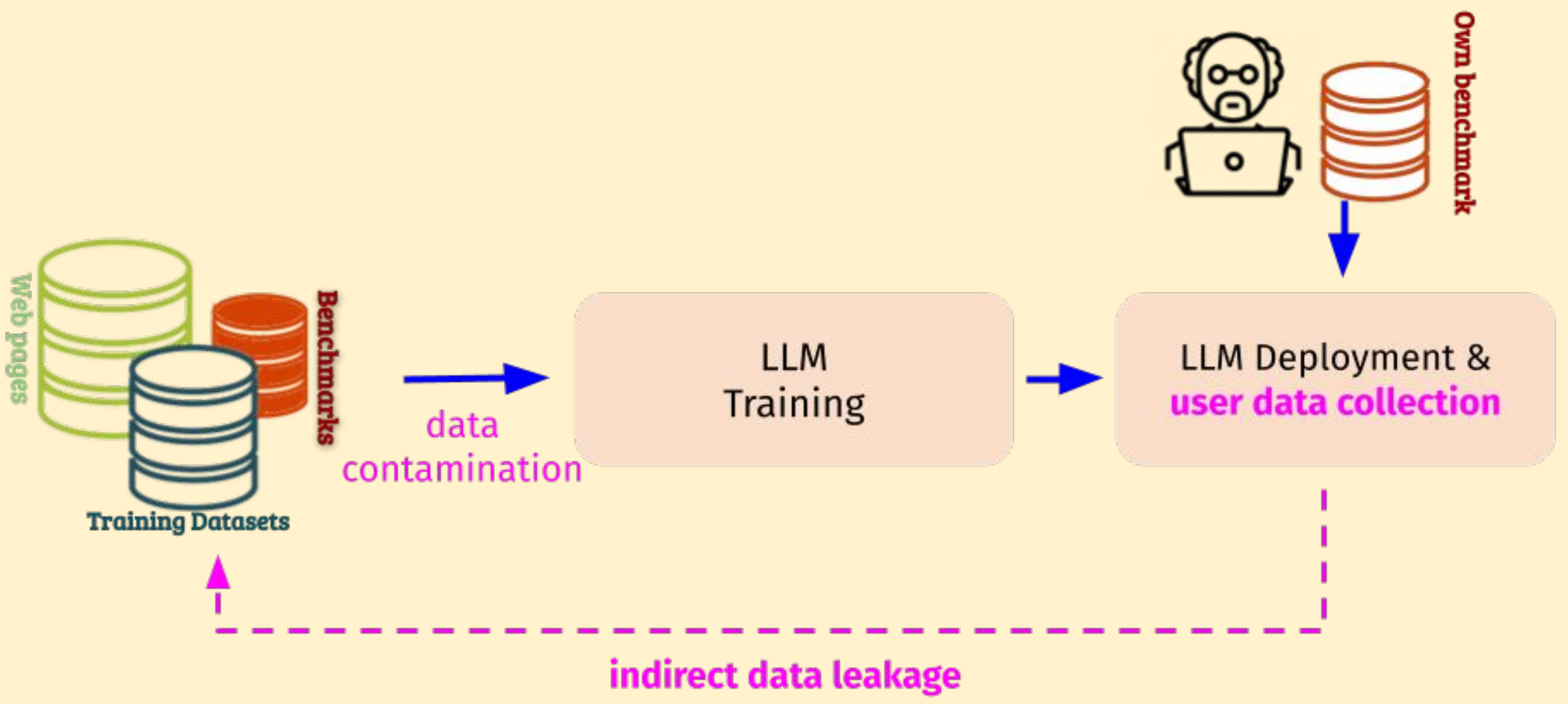
Data leakage in closed-source LLMs

Motivation

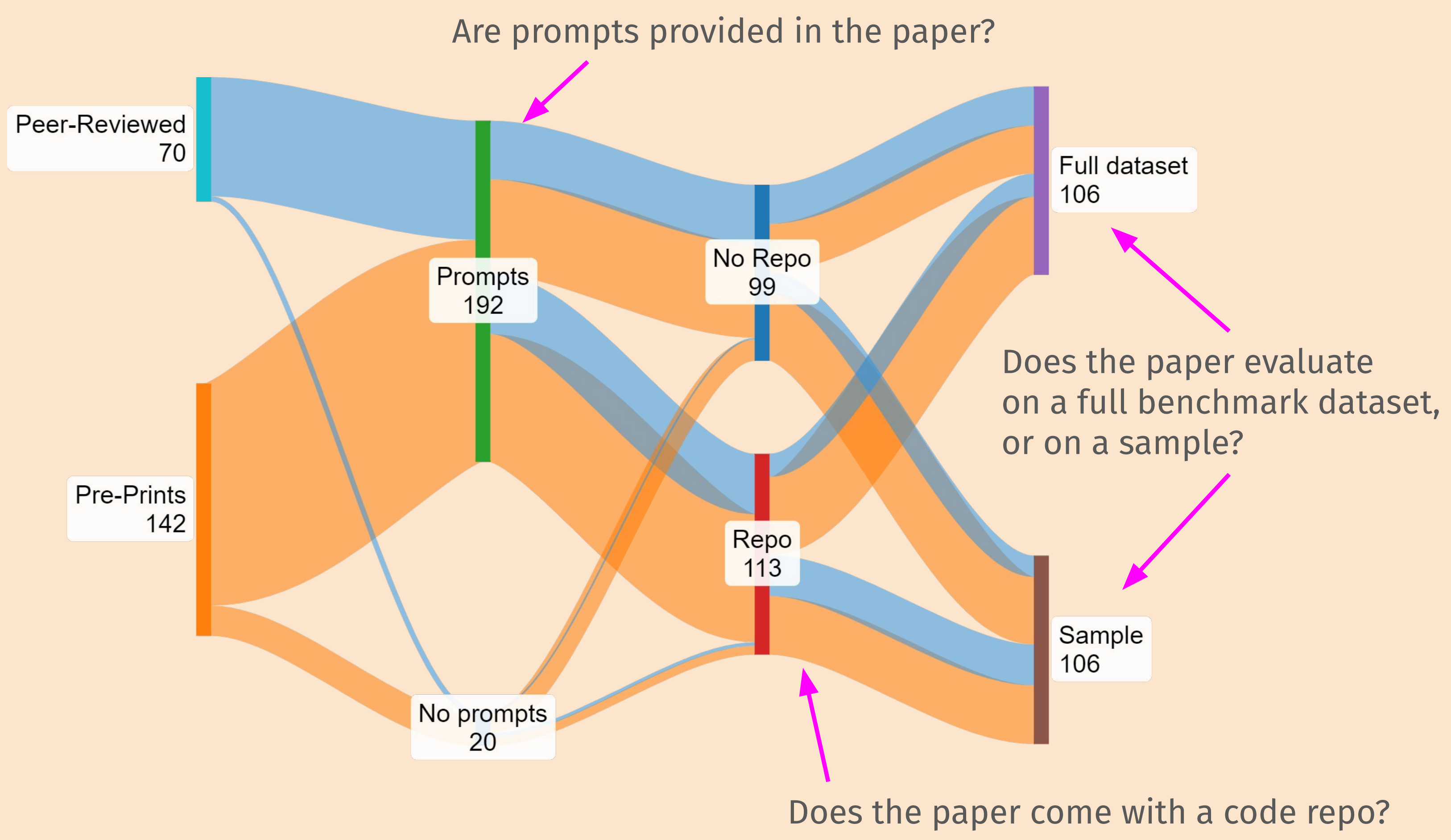
- some popular LLMs are **closed-source**
- lack of information on training data raises questions about the **credibility of LLM performance evaluation**
- several attempts to address this issue overlook the problem of **indirect data leaks**

Methodology & Findings

- We analyse **255 papers** evaluating OpenAI's GPT-3.5 and GPT-4 on a variety of tasks
- We conclude that **~42% of the relevant reviewed papers leaked data** to GPT-3.5 and GPT-4, for a total of **~4.7M samples** across **263 benchmarks**



Evaluation reproducibility



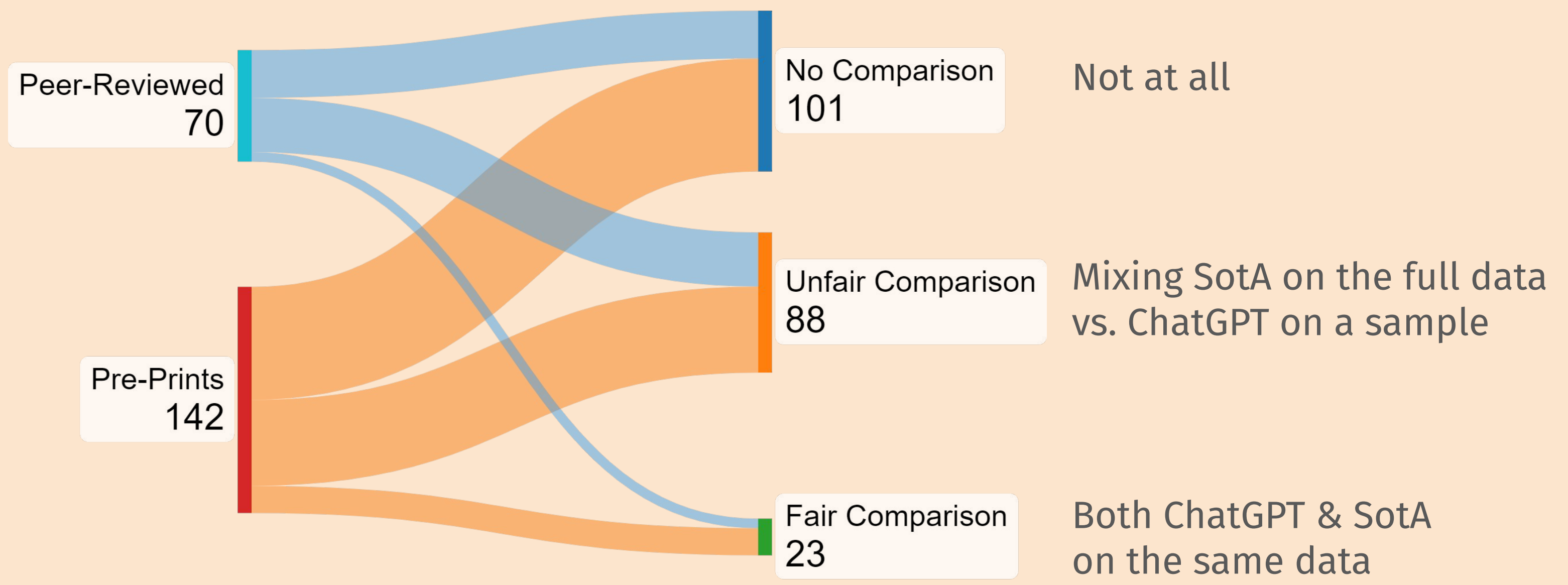
Are prompts provided in the paper?

Does the paper evaluate on a full benchmark dataset, or on a sample?

Does the paper come with a code repo?

Evaluation fairness

Does the paper compare with previous SotA?

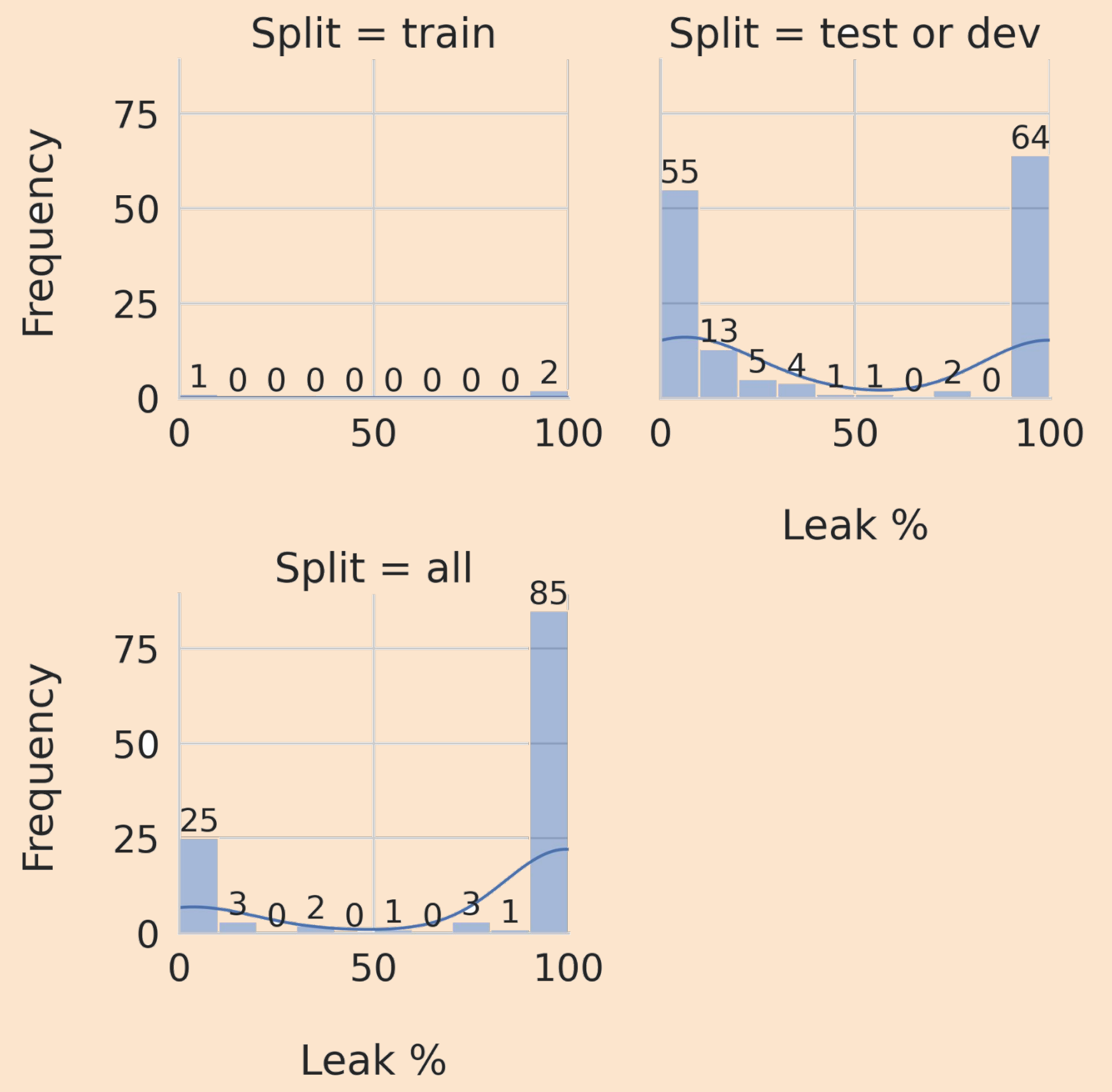


Not at all

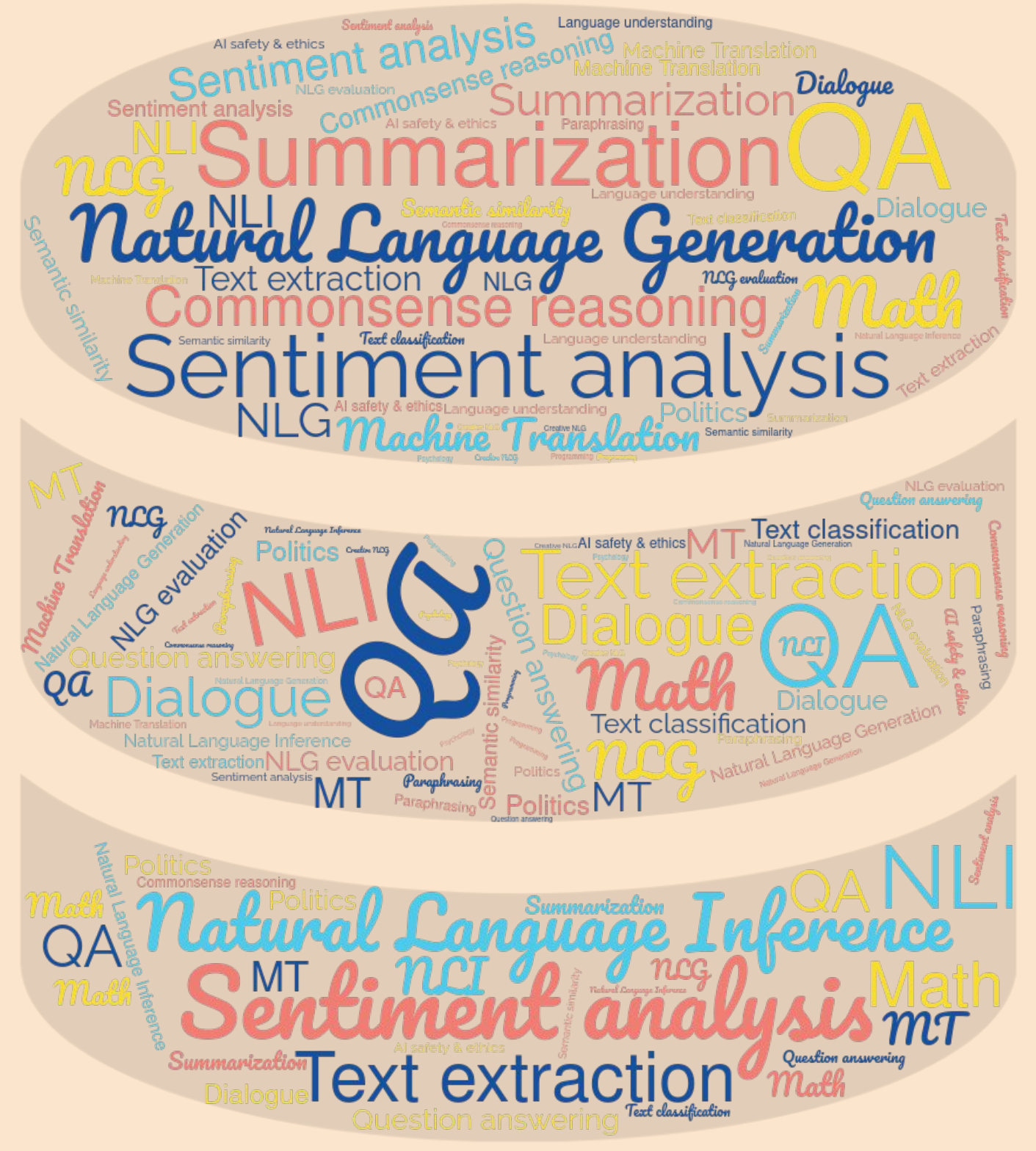
Mixing SotA on the full data vs. ChatGPT on a sample

Both ChatGPT & SotA on the same data

Indirectly leaked datasets



Number of times (y) we observed a specific percentage of data leaked (x) from the given split



NLP tasks: size/frequency corresponds to the number of leaked datasets.

Suggested practices in closed-source LLM evaluation

- **Access models in a non-leaky way:** use the API or opt-out of data collection (check vendor policy)
- **Interpret performance with caution:** incredible performance may be explained by data leakage
- **When possible, avoid using closed-source models**
- **Use a fair comparison:** sufficiently large samples, same for all methods, re-run experiments – do not copy & paste results
- **Make evaluation reproducible:** release code, prompts, model versions, sample selection (if subsampling)
- **Report indirect data leakage**, e.g. on our webpage!

Presented at EACL 2024, Malta

https://leak-llm.github.io/

Supported by the European Research Council (Grant agreement No. 101039303 NG-NLG). This work used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101). Patrícia Schmidtová was also supported by the Women in Quant Finance Network grant awarded by G-Research.