EACL 2024

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

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

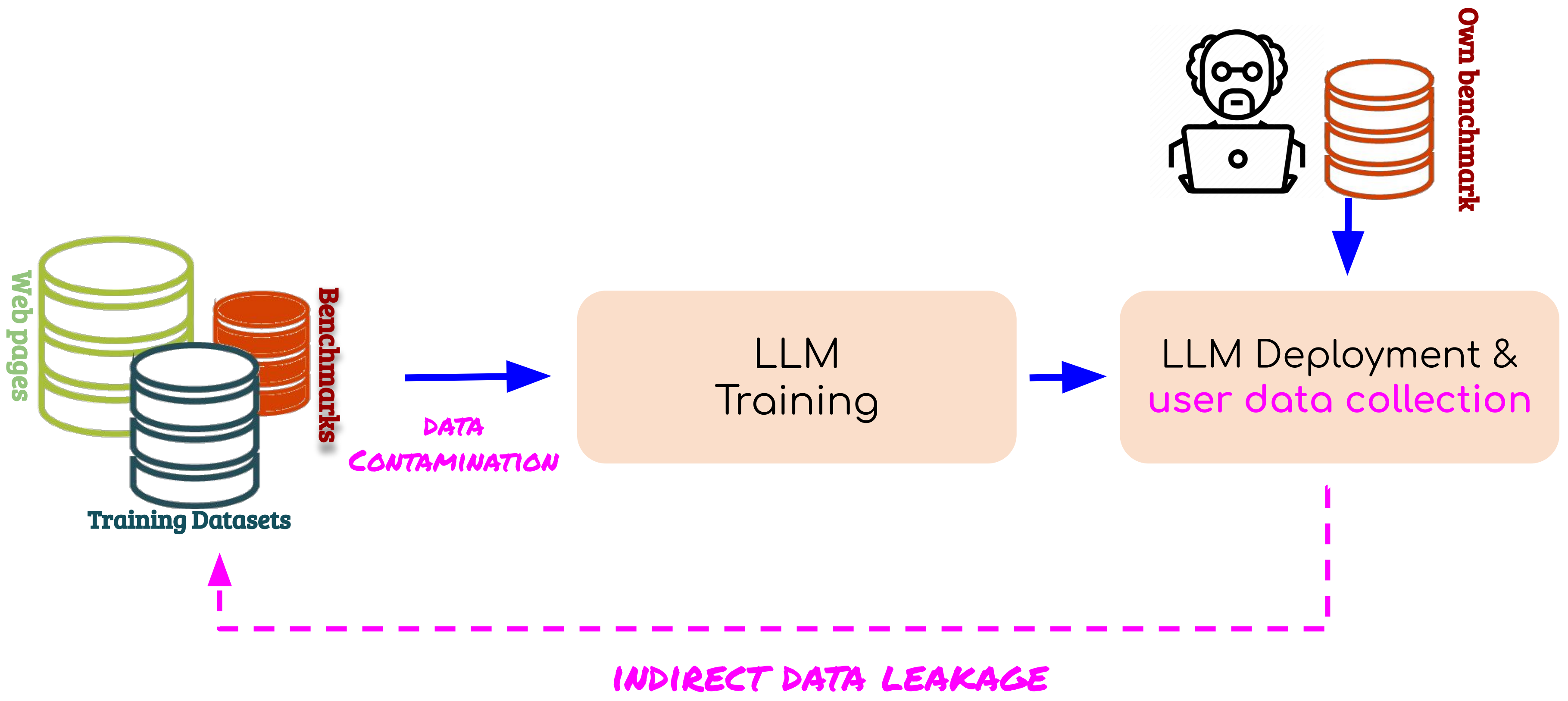
Overview

- The lack of details on training data for closed-source LLMs raised concerns on the issue of data contamination.
- Existing research overlooks when this happens indirectly - for example when models are updated from user data containing benchmarks.
- We review 255 papers causing an indirect data leak by evaluating GPT-3.5 and GPT-4 through the ChatGPT interface.
- We find that these models have been exposed to millions of samples from hundreds of NLP benchmarks.

Closed-Source LLMs & Data Contamination

- **Closed-Source**: LLMs only accessible via APIs or UIs
- For such models, researchers don't have access to:
  - Model weights
  - **Training data**
  - Other infrastructural details
- **Data contamination**: pre-training data may contain training, validation and **test sets** of NLP benchmarks

Indirect Data Leakage



Why is Indirect Data Leakage important?

1. It’s more difficult to trace due to possible subtle alterations

2. It comes with instructions included

Methodology

1. Identifying relevant work
2. Assessing quality and relevance
3. Summarizing the evidence
4. Evaluating reproducibility and fairness

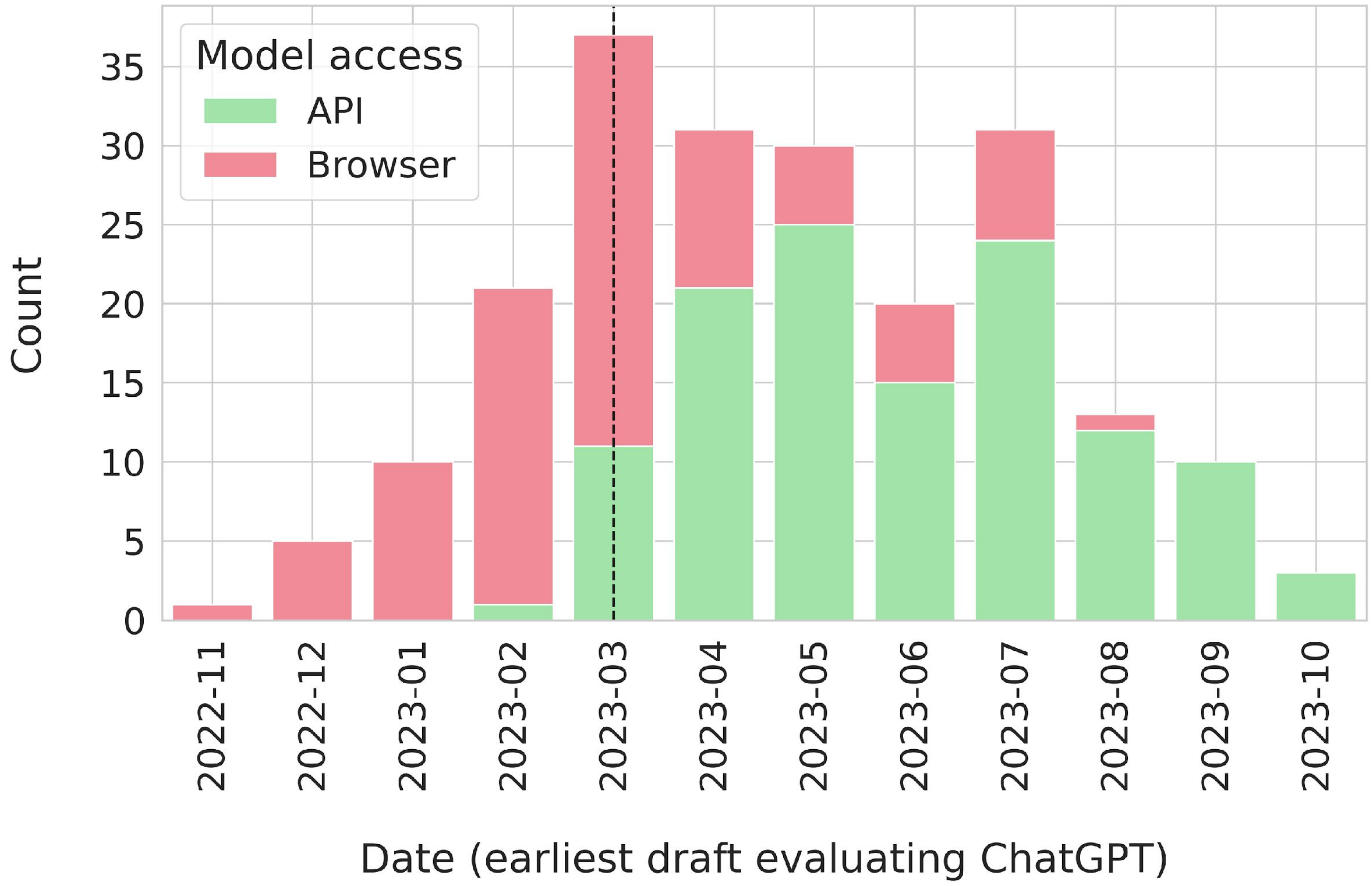
Results

We examined **255** papers, **212** of them interacted with closed-source models.

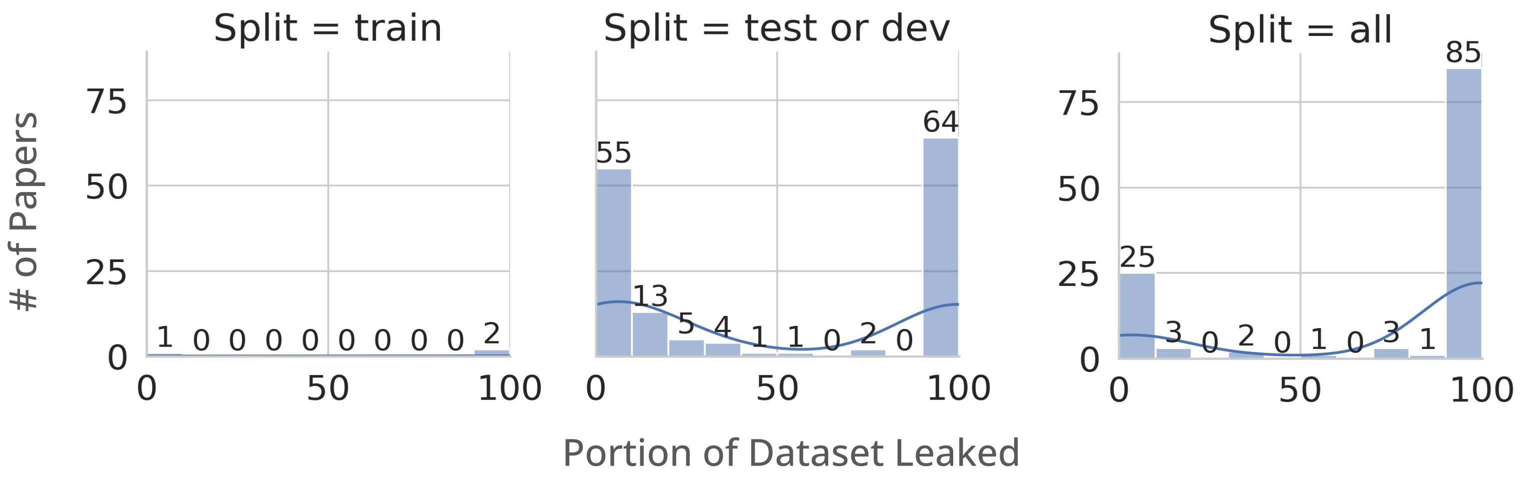
Out of these **212** papers, **90** (~**42%**) indirectly leaked data.

90 papers leaked ~**4.7M** samples form **263** NLP benchmarks.

Timeline of Documented ChatGPT Access



Results



- Leak overview:
  - **< 5%** for **66** datasets (~**25%**)
  - **5-50%** for **47** datasets (~**18%**)
  - **50-95%** for **10** datasets (~**4%**)
  - **> 95%** for **142** datasets (~**53%**)

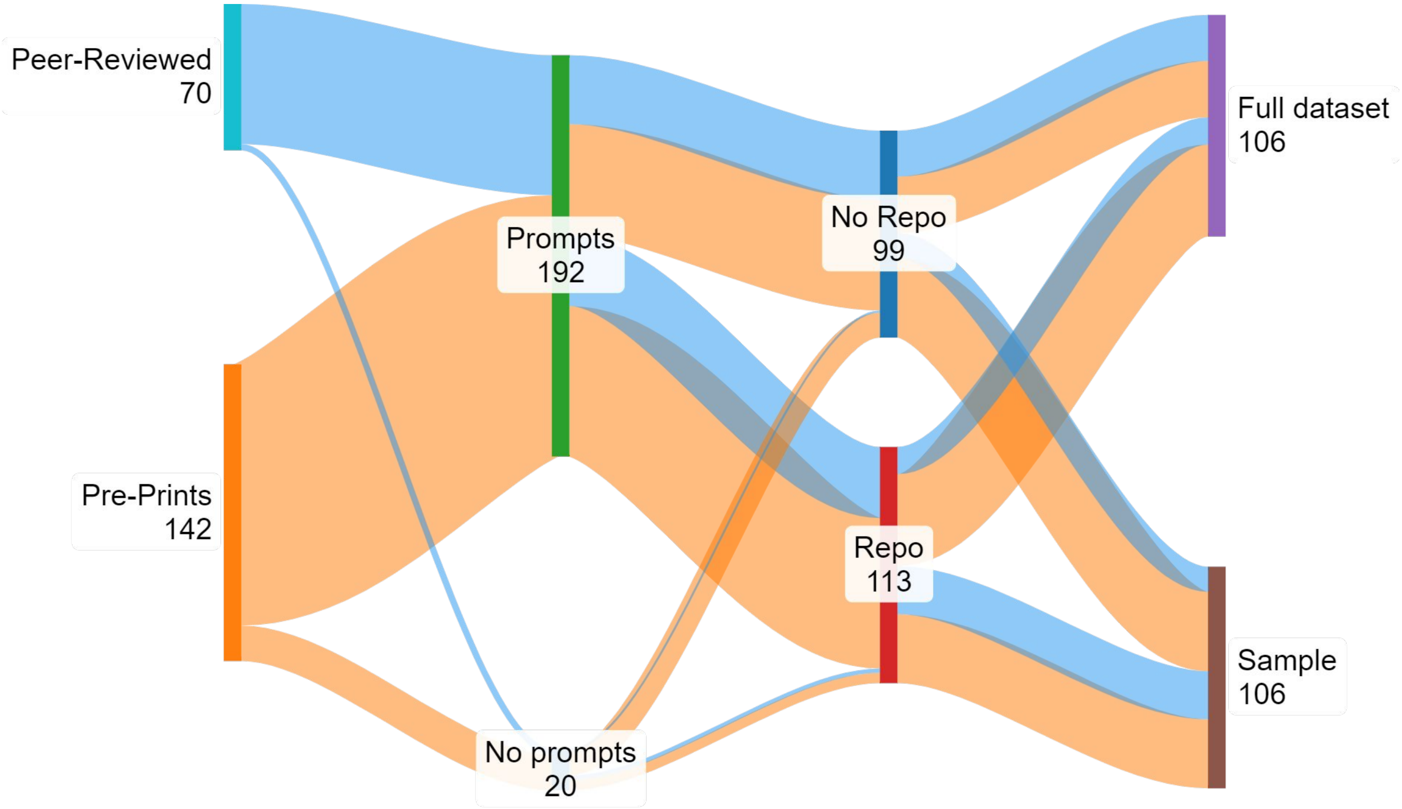
Tasks suffering the highest leaks:

- Natural Language Inference
- Question Answering
- Natural Language Generation

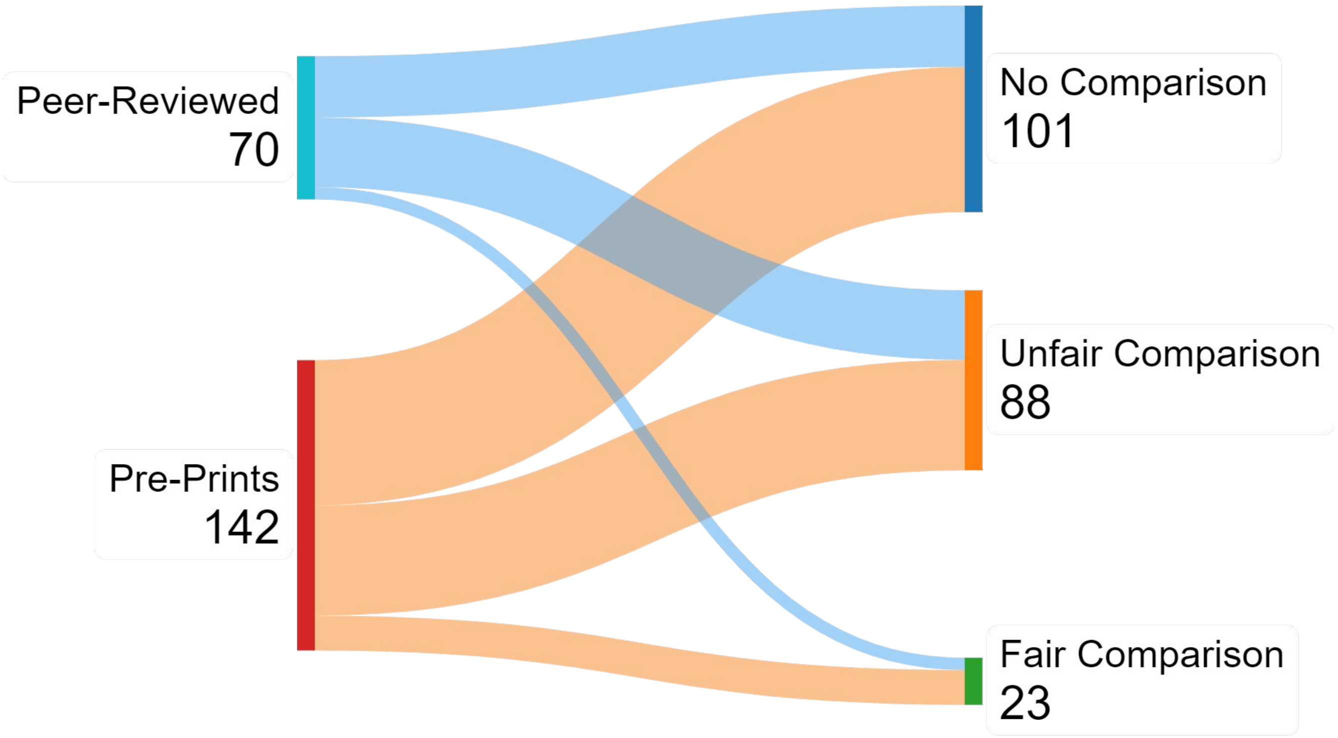
Results – Indirect Data Leak

- Mainly highly popular NLP benchmarks, e.g.:
  - Semeval2016 Task 6 (Stance Detection)
  - SAMSum (Dialogue Summarization)
  - MultiWOZ 2.4 (Dialogue)
- Smaller number: high-quality custom datasets
  - Often exams e.g., medicine, physics or law
  - Not all released publicly – only the authors and OpenAI now have access

Results – Reproducibility



Results – Fairness



Unfair comparison: comparing the performance on different samples of a dataset.

Suggested practices

- Access the model in a way that does not leak data
- Interpret performance with caution
- When possible, avoid using closed-source models
- Adopt a fair and objective comparison
- Make the evaluation reproducible
- Report indirect data leakage

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




balloccu schmidtova lango odusek

@ufal.mff.cuni.cz

We are worried about indirect data leakage, and you should be too!

Please help us document data that has been leaked:

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

Supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) & Charles University Project SVV 260 698.
Using the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).
Patrícia Schmidtová was supported by the Women in Quant Finance Network grant awarded by G-Research.

European Research Council

Established by the European Commission