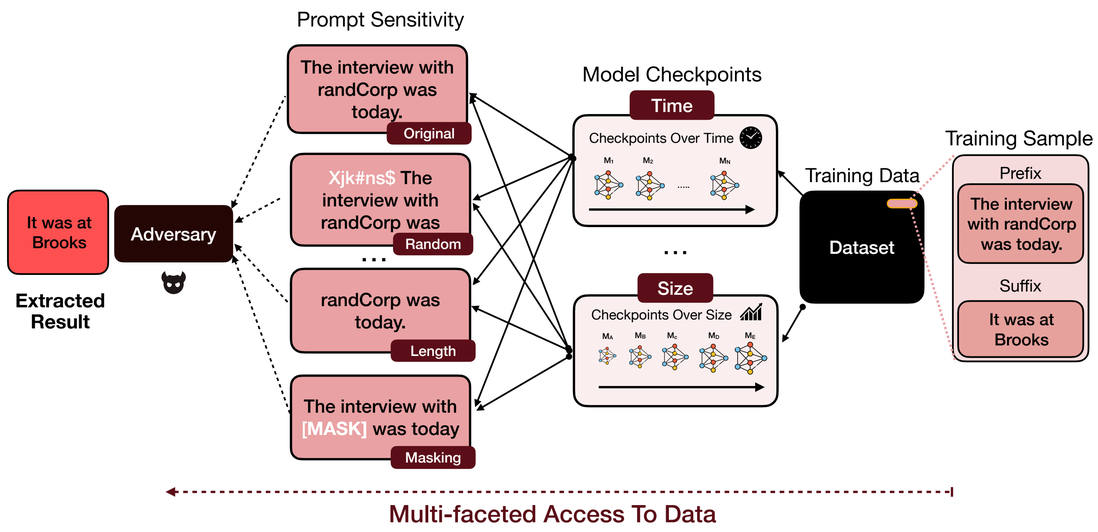
Towards More Realistic Extraction Attacks

Yash More* [1 2] Prakhar Ganesh* [1 2] Golnoosh Farnadi [1 2]

[1]Mila – Quebec AI Institute [2]McGill University *equal contribution

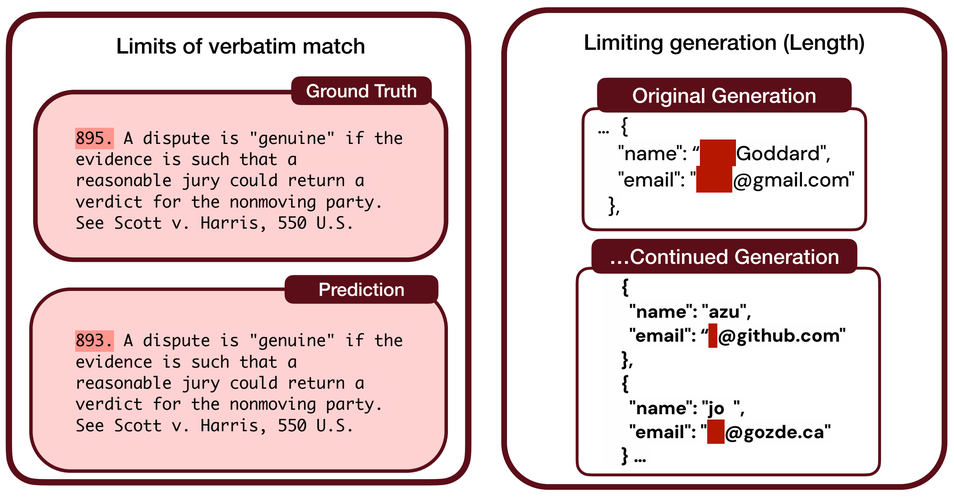
Composability in Current LLM Systems

In the real world, the adversary has multi-faceted access to a dataset through (a) varying ways of prompting the models and (b) accessing several models trained on the same data. **What if an adversary could exploit this?**



Evaluation Challenges: Extraction Attacks

- Extraction attacks, specifically discoverable memorization, are often evaluated in the literature using a verbatim match
- Existing metrics do not take into account the brittleness of LLM generations which can introduce noise, marking otherwise successful attacks unsuccessful.
- Existing literature chooses a fixed generation length to evaluate the attack (50 tokens), skipping varying lengths of both more and fewer tokens.
- To navigate these challenges, **we propose using an increased generation length and a continuous measure of risk that does not rely on verbatim match.**



Metric	Advantages	Disadvantages
Verbatim match: Exact overlap between generated and original text.	Strong evidence of memorization. No hyperparameter.	Highly brittle to noisy generation. Limited in scope.
Levenshtein Ratio: Minimum single-character edits (insertions, deletions or substitutions) between two sequences.	Can recognize fragments of extracted information. Robust to noise. Continuous metric.	Ambiguity at smaller thresholds due to coincidental similarity. Lacks interpretability.
Longest Common Substring (LCS): Longest common substring present in both sequences.	Compromise between edit distance metrics and verbatim match. Interpretable.	Does not account for the presence of multiple substrings.
n-gram similarity: Ratio of n-grams common between two sequences.	Granular. Can also detect multiple fragments of information.	Highly sensitive to the hyperparameter n.

Table 1. Overview of various similarity metrics in the context of extraction attacks.

We seek to answer the following questions

1. **Can adversaries exploit repeated prompting?** Extraction attacks are highly sensitive to the prompt design, **extracting over 20% more data with even minor changes to the prompt.**

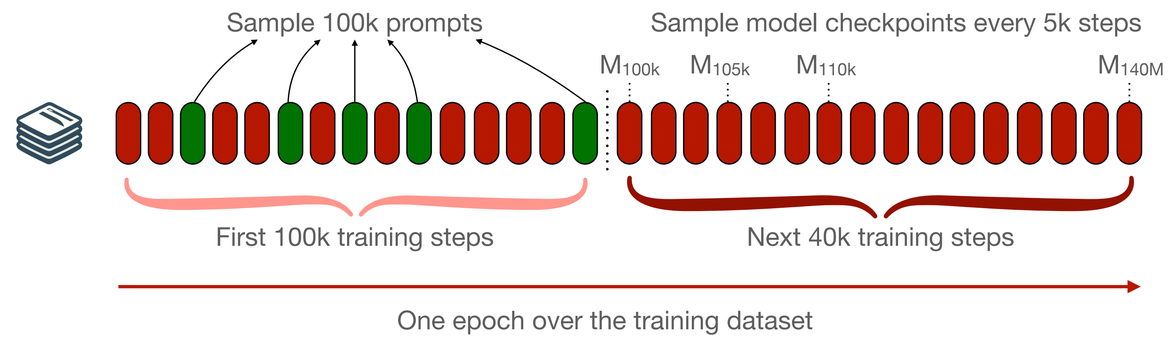
2. **Does access to multiple checkpoints increase extraction risk?** An adversary with access to multiple checkpoints can **increase the extraction rates up to $2\times$,** significantly heightening the risk of information leakage.

3. **Is verbatim match a reliable measure of information leakage?** We show the limitations, and resulting underestimation of extraction risk. **As alternatives, we propose the use of approximate matching to estimate the risks better**

4. **How effective is data deduplication in reducing the extraction risks?** A powerful adversary with repeated prompting and access to checkpoints can still extract information, passing rates of extraction even without-deduplication.

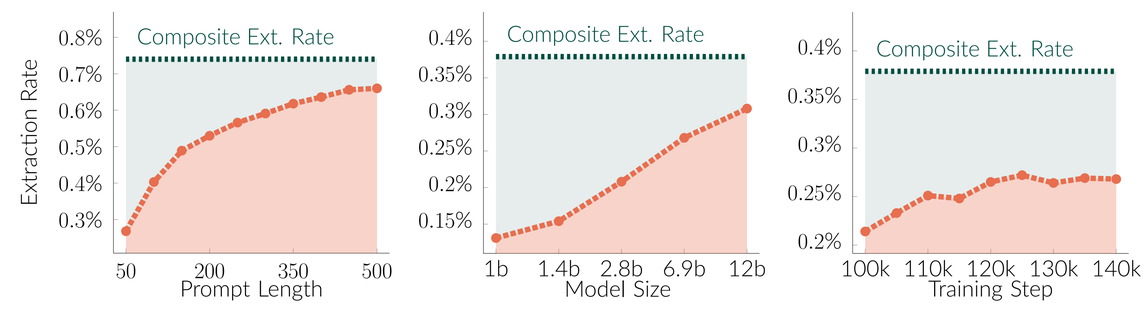
Evaluation Setup

We use the Pythia suite for all our experiments. We choose Pythia due to (a) access to models of various sizes, (b) across multiple checkpoints, and (c) training data to simulate the real-world availability of models.



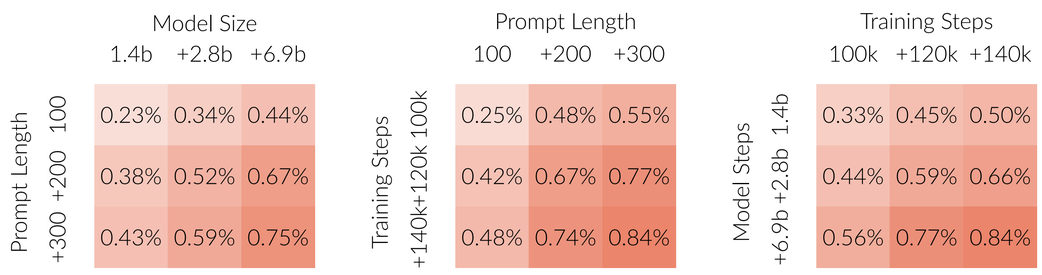
Extraction Rates across Dimensions

- Increasing prompt length results in better extraction rates, with the composite extraction rate better than even the largest prompt length (500). Thus, an adversary can exploit repeated prompting to extract more information.
- Larger models tend to memorize more information, however, the composite extraction rate is higher than the extraction rate of any single model size. Releasing models in different sizes creates an increased risk of leakage.
- Attacking models at later stages of training is more successful, and an adversary can nearly double their extraction rate with a composite attack. Thus, even regularly updating models can exacerbate the risks of extraction attacks.

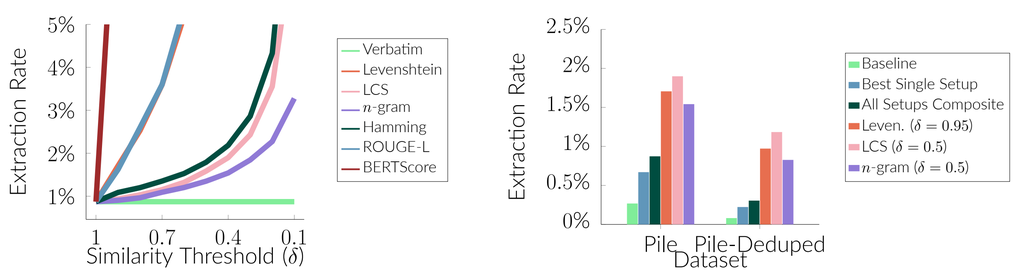


Combining Churn: Realistic extraction attacks

A real-world adversary can combine multiple axes and significantly increase their extraction rates. Analyzing two axes at a time, the overall composite extraction rate (bottom right) is 3-4 times higher than the base setup (top left) and nearly double the composite extraction rates along one axis (top right and bottom left).



Combining all three axes of variability, along with approximate matching metrics, can increase the extraction rate by more than $4\times$. While data deduplication does lower the extraction rate, our trends persist, and a powerful adversary can extract significantly more information even after data deduplication.



Future Work

By highlighting the reality of an adversary performing extraction attacks in the current LLM landscape, our work reveals a severe underestimation of information leakage in the existing literature. We list potential future research directions below:

- **Understanding composability:** We believe the composability of information leakage should be researched in more detail, especially while understanding the risks posed by data extraction attacks.
- **Pursuing Applications**: Application of our study in real-world settings with concrete adversarial scenarios or areas like Identifying copyright claims, assessing the PII etc.
- **Rethinking privacy:** There is a need to explore how the study's findings on discoverable memorization translate to other privacy attacks.
- **Mitigation techniques:** Further work on **navigation** of risks posed by such adversaries would allow us to develop robust safety and privacy mechanisms.

yash.more, prakhar.ganesh@mila.quebec