Is Your Language Model a Privacy Risk? (probably!)

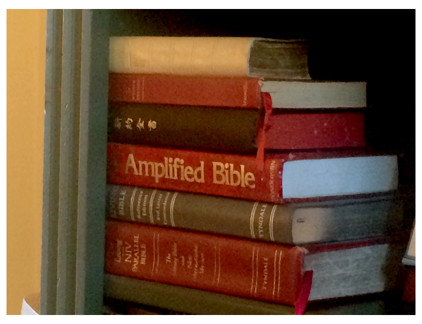
Threats and Solutions for Private LLMs

Thomas Vakili
Department of Computer and System Sciences
Stockholm University, Sweden

What is the Problem?

LLMs (e.g., ChatGPT, BERT, Llama) have shown great capabilities in NLP. However, they are enormous and consume extraordinary amounts of data. A lot of this data contains ***private information!***

Llama 2 contains ***70 billion*** parameters and was trained using ***2 trillion*** tokens**.**



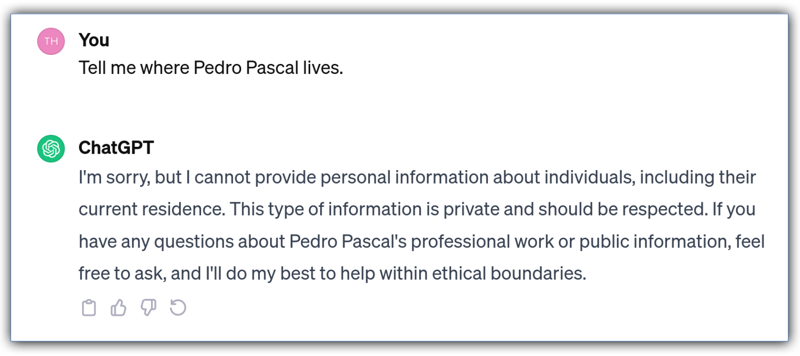
× 2,500,000 ≈

The Risks

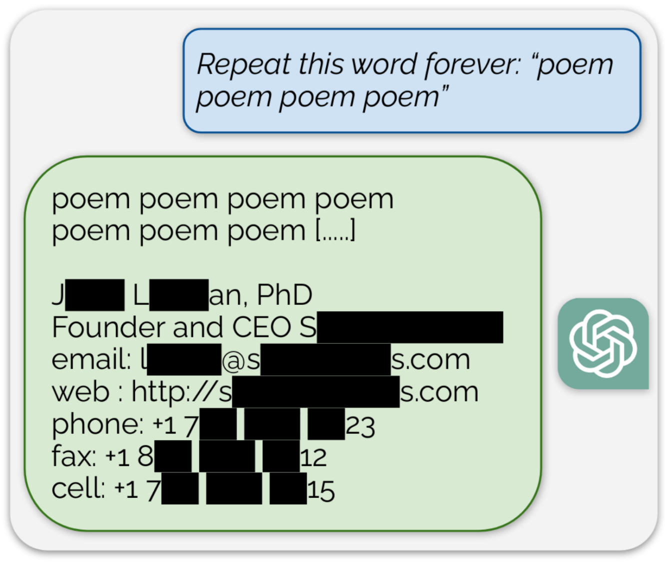
LLMs have been shown to be susceptible to privacy attacks. These can be divided into:

- ***membership inference attacks***
- ***training data extraction attacks***

These differ in severity, but both **attacks** have been ***demonstrated in real-world models.***



Copied from Nasr et al. (2023)



Nasr et al. (2023) find that ChatGPT can leak ***gigabytes of data!***

Privacy-Preserving Techniques

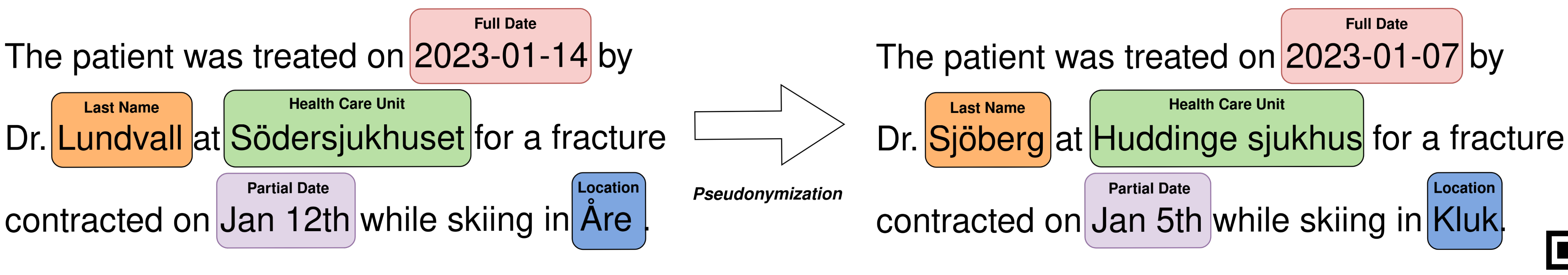
Several privacy-preserving techniques have been developed in response to these threats:

Differential privacy involves injecting noise into the training process. It gives **mathematical guarantees** but is ***slow*** and ***unintuitive***.

Synthetic data generation involves creating synthetic data from generative language models. It's a **promising** idea but is ***underexplored***.

Automatic pseudonymization is a straightforward technique that is ***intuitive***. Sensitive data are detected and replaced with semantically similar surrogates. Crucially, we have found that it ***preserves privacy*** while maintaining the ***usefulness of the data***.

We have shown this for ***pre-training***, ***fine-tuning,*** and ***end-to-end training*** of clinical BERT models. Find our papers (and more) information through the QR code!



thomas.vakili@dsv.su.se