Towards security and privacy in document understanding models

Sonia Vanier and Jérémie Dentan – École Polytechnique

Introduction

- LLM are trained on massive data scraped from the net
- It is infeasible to properly sanitize these datasets to remove personal or sensitive information
- Models have been attacked in production, exposing sensitive data used during training

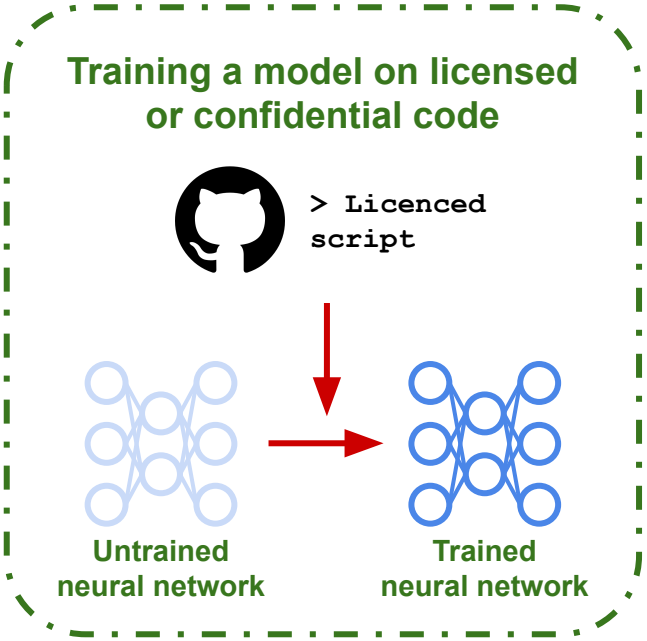
→ We develop a new privacy attack against document understanding models

→ We use it to analyze model's vulnerability

→ Long-term goal: protect models at minimal cost

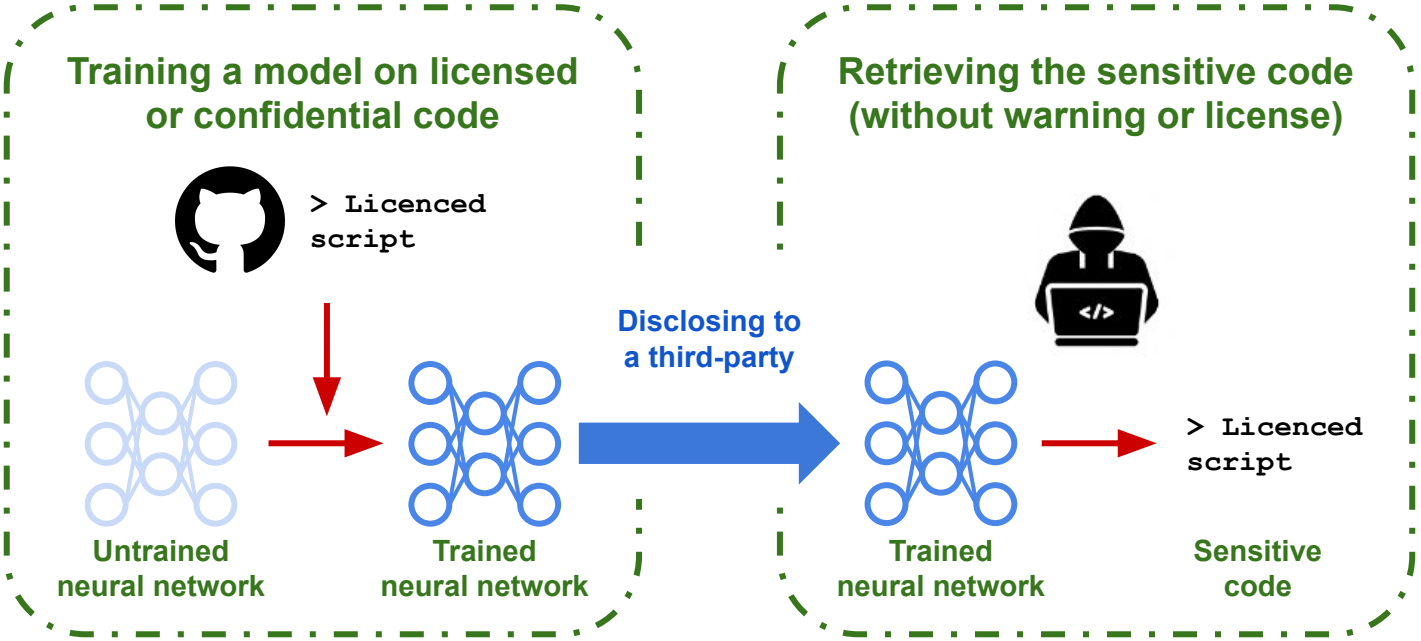
Privacy issues in Language Models

Language models memorize information from their training set and can disclose it at inference time.



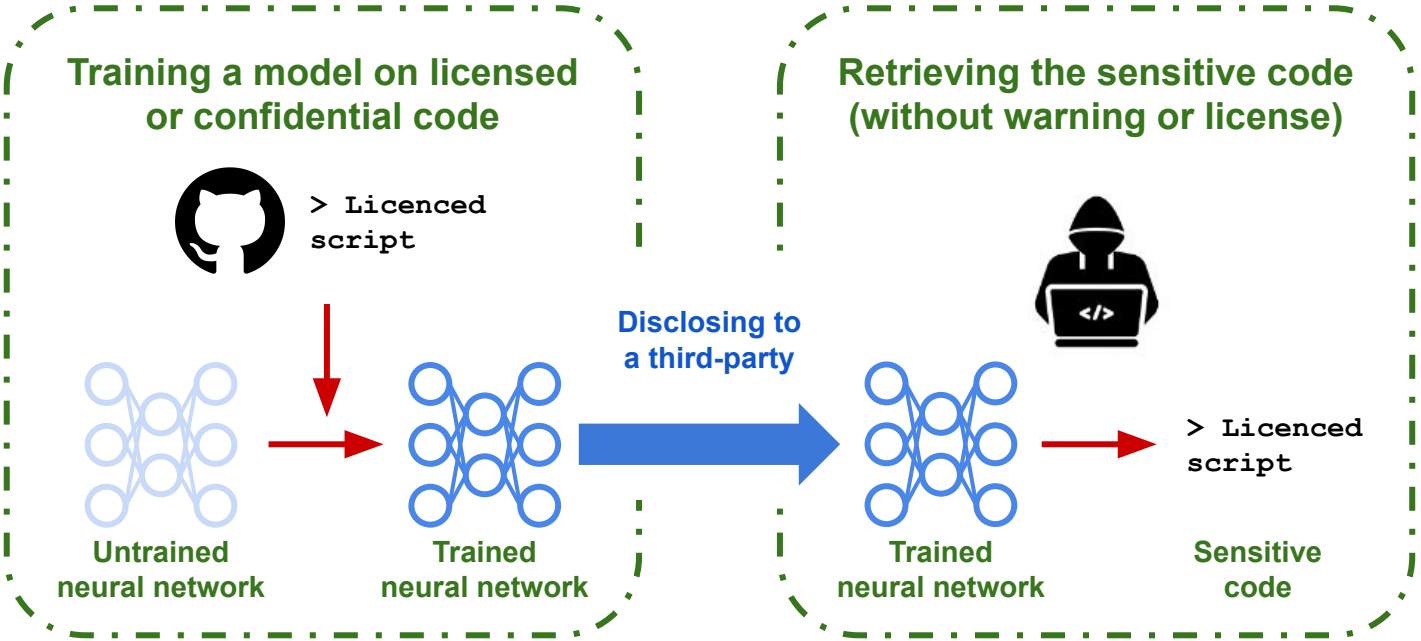
Privacy issues in Language Models

Language models memorize information from their training set and can disclose it at inference time.



Privacy issues in Language Models

Language models memorize information from their training set and can disclose it at inference time.



It can happen:

- By mistake
- On purpose

What kind of models are vulnerable to privacy attacks ?

Decoder-only, generative language models

Tasks: next token prediction
(Gemini, Llama, Copilot, GPT etc.)

- Very powerful abilities, many applications
- Easier to do privacy attacks
- Larger models memorize more

[1] Nicholas Carlini, Daphne Ippolito, *et al.* Quantifying Memorization Across Neural Language Models. ICLR. 2023.

What kind of models are vulnerable to privacy attacks ?

Decoder-only, generative language models

Tasks: next token prediction
(Gemini, Llama, Copilot, GPT etc.)

- ✅ Very powerful abilities, many applications
- ❌ Easier to do privacy attacks
- ❌ Larger models memorize more

Encoder-only, discriminative language models

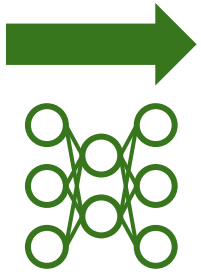
Tasks: classification, entity extraction, etc.
(BERT, RoBERTa, etc.)

- ❌ Specific applications
- ✅ Harder to do privacy attacks
- ❌ Underexplored domain

[1] Nicholas Carlini, Daphne Ippolito, *et al.* Quantifying Memorization Across Neural Language Models. ICLR. 2023.

We developed a new attack against some encoder-only models

The first reconstruction attacks against document understanding models



Document understanding model

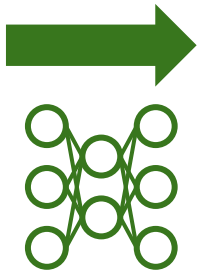
≈ BERT + 2D position encoding + visual features

(text) *(layout)* *(image)*

[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.

We developed a new attack against some encoder-only models

The first reconstruction attacks against document understanding models



1. Name
2. Surname
3. Birth date
4. Document ID

Document understanding model

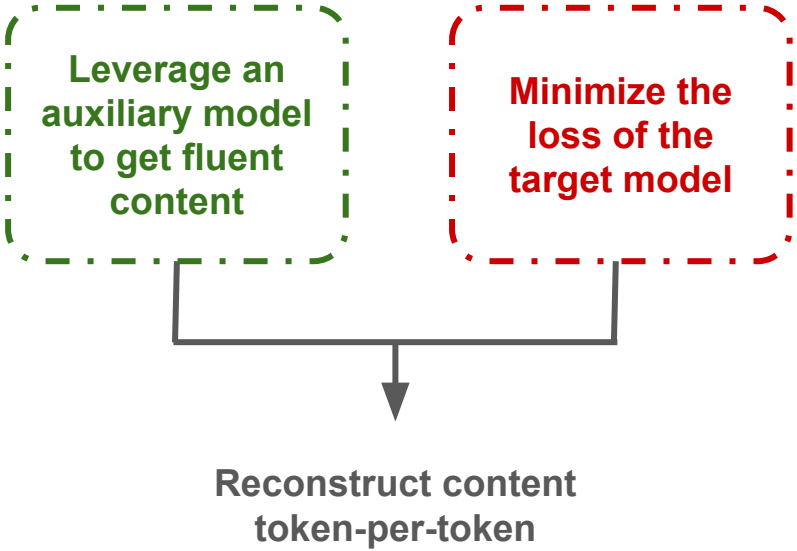
≈ BERT + 2D position encoding + visual features

(text) *(layout)* *(image)*

[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.

Our attack

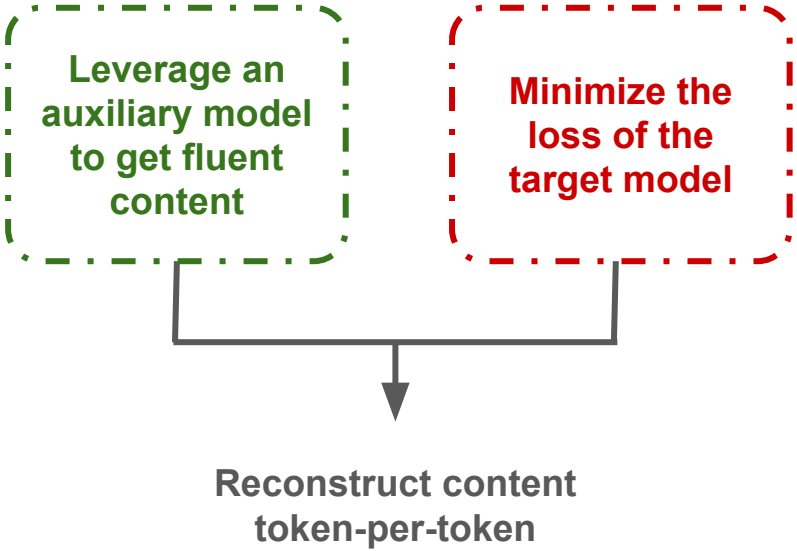
How do we reconstruct data ?



[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.

Our attack

How do we reconstruct data ?



Strong empirical results

- **Experiments in many settings**
  (2 architecture, 2 datasets, 4 tasks)
- **Perfectly reconstruct up to 4.1% of the fields in the training set**
  (including names, dates, addresses, 7-digit numbers...)

[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.

Insight #1: Does our attack require overfitting?

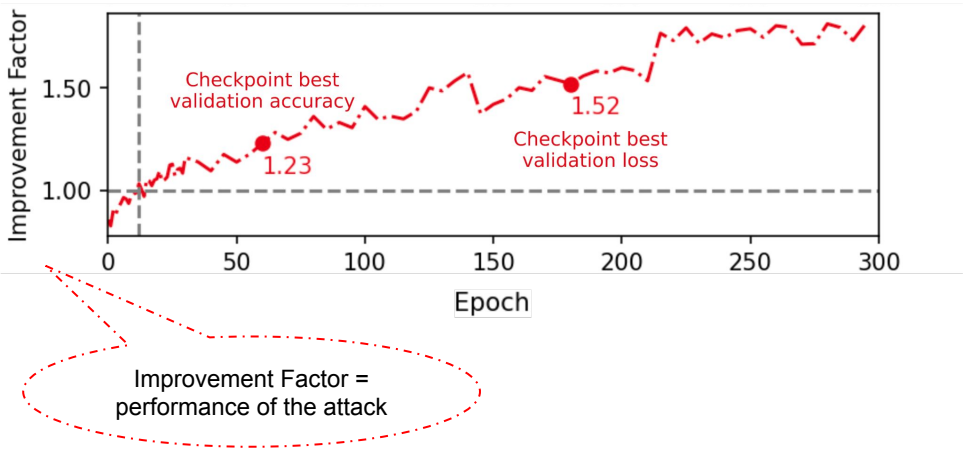
No, it does not.

[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Insight #1: Does our attack require overfitting?

No, it does not.

- Memorization starts well before overfitting.
- Overfitting contributes to memorization, but it is not necessary.
- Consistent with other works such as [3]



[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Insight #2: Does the visual modality contribute to the attack?

Yes, it does.

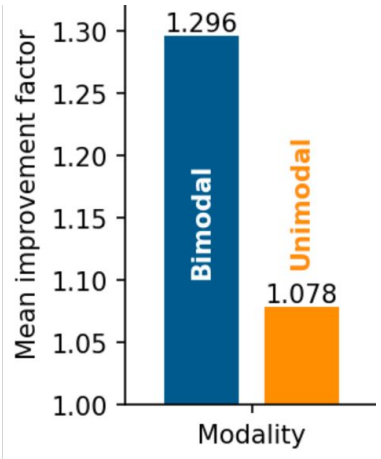
[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Insight #2: Does the visual modality contribute to the attack?

Yes, it does.

- **Pixel/token associations are memorized by the model.**

Document model
≈ **BERT + 2D position encoding** + **visual features**
(text) *(layout)* *(image)*



[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Insight #3: Does the layout contributes to the attack?

Yes, it does.

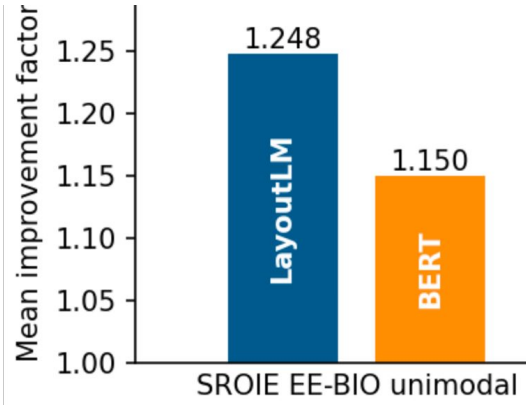
[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Insight #3: Does the layout contributes to the attack?

Yes, it does.

- **Layout/token associations are memorized by the model.**

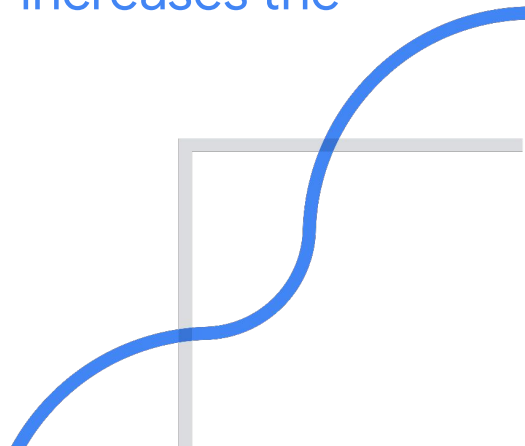
Document model
≈ **BERT** + **2D position encoding**
(text) *(layout)*



[2] Jérémie Dentan, Arnaud Paran, Aymen Shabou. Reconstructing training data from document understanding models. Usenix Security. 2024.
[3] Chiyuan Zhang, Samy Bengio, *et al.* Understanding deep learning requires rethinking generalization. ICLR. 2017.

Conclusions

- Many types of model memorize their training data
- We developed the first privacy attack against document models
- Attacks are realistic even without overfitting
- Multimodality increases the privacy risk



Thank You

Acknowledgements

This work received financial support from Crédit Agricole SA through the research chair "Trustworthy and responsible AI" with École Polytechnique.