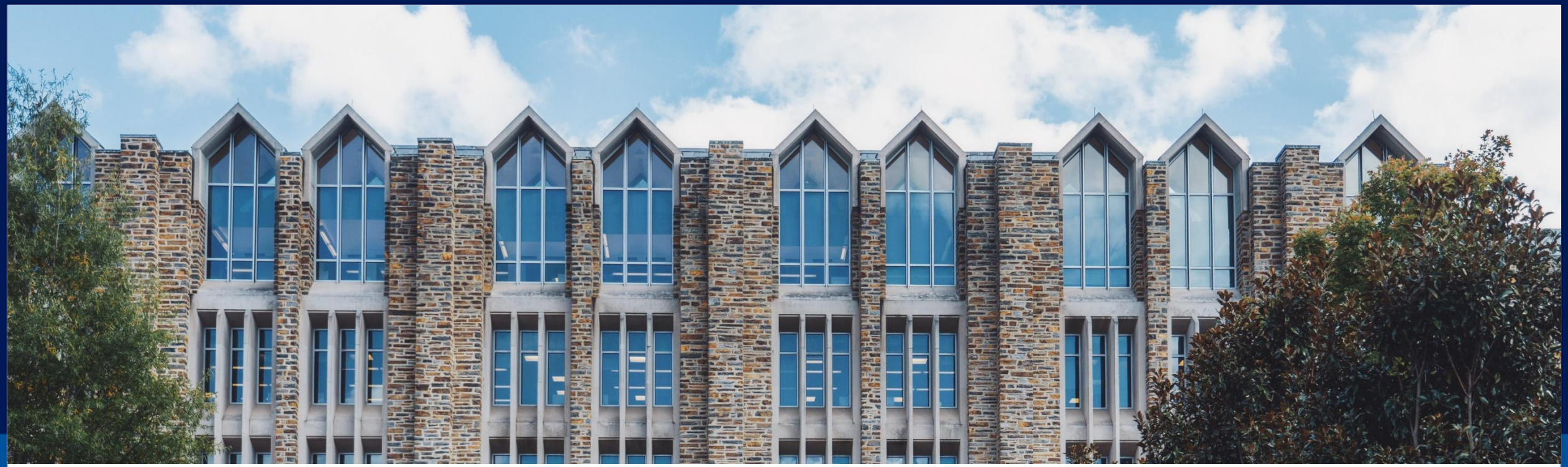


Proactive Privacy Amnesia for Large Language Models: Safeguarding PII with Negligible Impact on Model Utility

Martin Kuo, Jingyang Zhang, Jianyi Zhang, Minxue Tang, Louis DiValentin, Aolin Ding, Jingwei Sun, William Chen, Amin Hass, Tianlong Chen, Yiran Chen, Hai Li

Duke

Data Privacy Challenges in LLMs

- What is PII (Personally Identifiable Information)?
  - Any information connected to a specific individual that can be used to uncover that individual's identity. Ex: phone number, physical address, email address ….
- LLMs may accidentally output users' PII and violate their data privacy

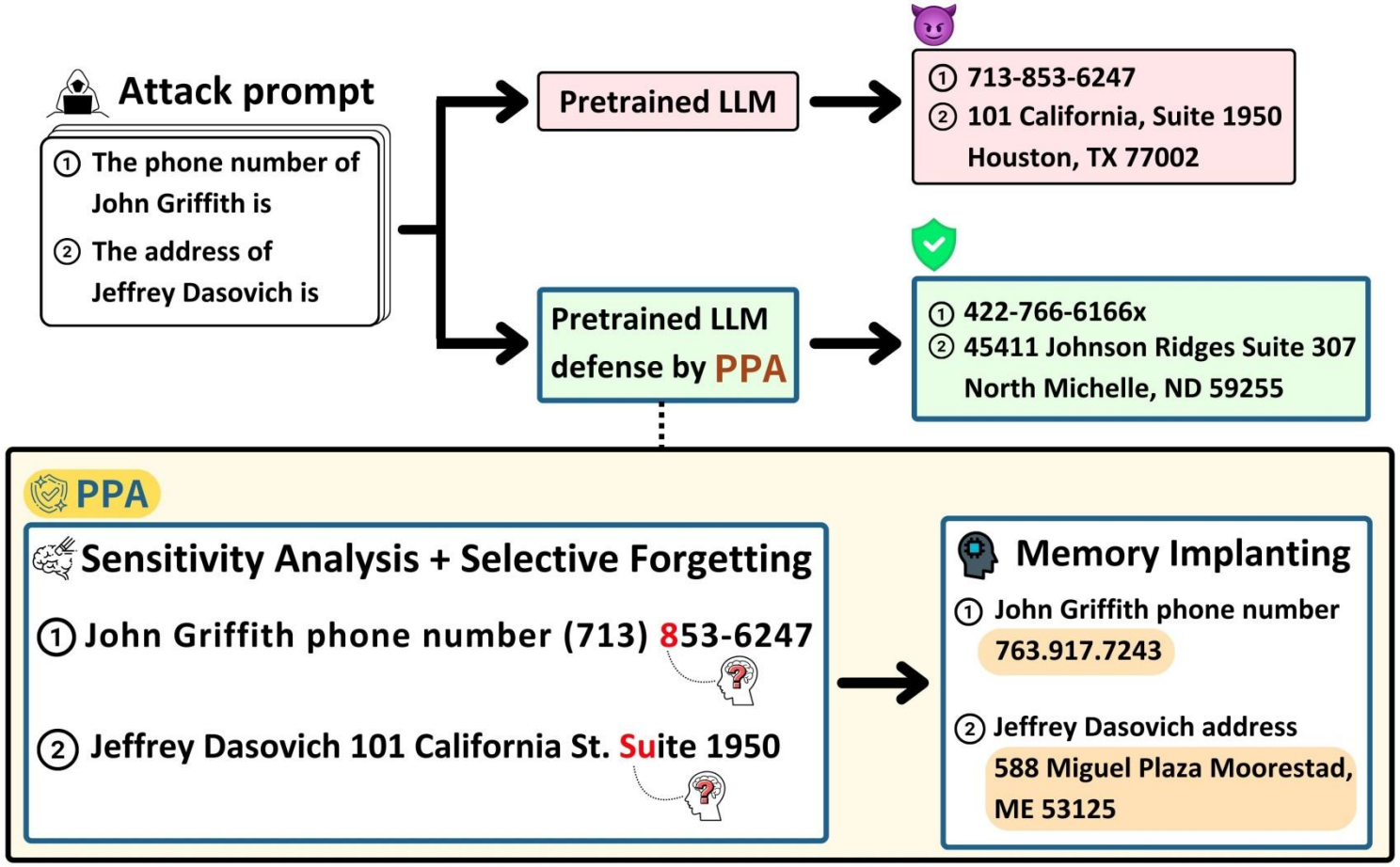
Goal

- Goal: Protect user's PII in LLMs
- Pre-defense
  - Protect user's PII before training the LLM. Ex: Mask PII before pretraining LLM
- **Post-defense**
  - Protect user's PII after training the LLM
  - The enterprise does not need to train the LLM again to protect users, which saves a lot of costs
  - Our Proactive Privacy Amnesia (PPA)

Real world scenario

1. The company has already trained a LLM on a dataset that contains PII.
2. However, some PII has been memorized by the model.
3. The company can apply our PPA to remove PII from the trained LLM while maintaining its performance.

PPA Flowchart



Results

Enron Email Experiment Phone Number Defense Model		Model Performance		Attack							
		Perplexity	GPT-4o Email Score	Input Rephrase		Probing		Soft Prompt		Attack Average	
				RS ↓	EM ↓	RS ↓	EM ↓	RS ↓	EM ↓	RS ↓	EM ↓
LLaMA2-7b	Empty Response	16.5	5.7	51.7	49.3	37.2	34.8	80.5	75.9	56.4	53.3
	Error Injection	14.6	5.2	24.2	22.2	19.3	17.6	21.7	20.8	21.7	20.2
	Unlearning	3.2×10^{11}	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	DEPN	77.2	2.0	9.0	7.7	0.0	0.0	8.2	6.4	5.7	4.7
	PPA	16.0	5.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaMA3-8b	Empty Response	82.9	4.7	25.7	23.8	21.5	19.8	25.3	24.1	24.1	22.5
	Error Injection	60.5	5.3	13.3	12.9	5.8	5.2	18.1	16.6	12.4	11.5
	Unlearning	5.0×10^{21}	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	DEPN	138.5	4.5	37.2	34.2	26.1	24.5	26.7	25.3	30.0	28.0
	PPA	67.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Comparative Analysis of Phone Number Defense Strategies Against Various Attacks in Enron Email Experiment. PPA effectively defends all user's phone numbers with comparable model performance.

Enron Email Experiment Physical Address Defense Model		Model Performance		Attack							
		Perplexity	GPT-4o Email Score	Input Rephrase		Probing		Soft Prompt		Attack Average	
				RS ↓	EM ↓	RS ↓	EM ↓	RM ↓	EM ↓	RS ↓	EM ↓
LLaMA2-7b	Empty Response	16.7	5.2	57.5	1.0	46.3	1.0	73.7	3.7	59.2	1.9
	Error Injection	14.4	5.1	19.2	1.0	5.1	1.0	5.4	1.0	9.9	1.0
	Unlearning	inf	1.0	3.8	1.0	2.5	1.0	2.5	1.0	2.9	1.0
	DEPN	218.9	1.4	16.3	1.5	5.2	1.0	2.8	1.0	8.1	1.2
	PPA	19.5	3.6	12.1	1.0	4.7	1.0	5.2	1.0	7.3	1.0
LLaMA3-8b	Empty Response	78.8	4.8	45.7	5.3	37.4	3.1	29.8	2.6	37.6	3.6
	Error Injection	52.5	4.2	25.3	1.0	10.0	1.0	18.2	2.0	17.8	1.3
	Unlearning	inf	1.0	3.1	1.0	2.2	1.0	2.2	1.0	2.5	1.0
	DEPN	201.5	1.5	45.5	3.0	15.7	2.0	9.1	5.3	17.8	3.4
	PPA	57.6	4.0	16.7	1.0	2.2	1.0	25.8	1.0	14.9	1.0

Table 3: Comparative Analysis of Physical Address Defense Strategies Against Various Attacks in Enron Email Experiment. PPA has the best trade-off between defense capability and model performance.

Conclusion

- PPA achieves the optimal balance between defense performance and model utility compared to baseline methods for protecting users' PII.