RAG Impact on LLM Training Data Leakage

Baseline vs. RAG Comparison for Data Extraction Attacks

Syed Ahmed Khaderi

January 2026

1 Objective

This experiment studies the effect of Data Extraction Attacks (DEA) when Retrieval-Augmented Generation (RAG) is implemented within a baseline LLM inference pipeline.

The paper "RAG: The Good and The Bad" investigates whether incorporating retrieval-augmented generation (RAG) during inference reduces a language model's tendency to output memorized training data and increases privacy against DEA.

The key hypothesis is that when diverse, retrieval data is provided as context, the model shifts its generation distribution away from memorized training data, thereby reducing privacy leakage.

2 Experimental Setup and Configuration

2.1 Paper's Approach

From Section 5 of Zeng et al. (2024):

- **Model**: GPT-Neo-1.3B (`EleutherAI/gpt-neo-1.3B`)
- **Training Data**: Enron Mail dataset, a subset of The Pile used in GPT-Neo pretraining
- **Retrieval Dataset**: WikiText-103 (disparate from training data)
- **Embedding Model**: bge-large-en-v1.5
- **Retrieval Method**: k-NN with k=2 documents per query
- **Attack Methods**:
  - Targeted PII extraction (5,000 prompts)
  - Prefix attack (1,000 prompts)
- **Key Finding**: RAG reduced email extraction from 245 to 2 PIIs (99.2% reduction) and prefix reconstructions from 213 to 70 (67.1% reduction)

2.2 Our Implementation

Table 1: Experimental Configuration

Category	**Details**
Model	
Base Model	EleutherAI `gpt-neo-1.3B`
Compute Device	CUDA (Google Colab T4 GPU)
Data	
Training Data Source	`suolyer/pile_enron`
PII Frequency Threshold	≥ 5 occurrences (high memorization)
Retrieval Corpus	WikiText-103 (10,000 documents)
PII Types Evaluated	Email and phone
RAG Configuration	
Embedding Model	`BAAI/bge-large-en-v1.5`
Vector Store	ChromaDB v0.4.22
Top-k Retrieved Documents	$k = 2$
Retrieval Strategy	Semantic retrieval with cosine similarity
Context Integration	Prepend retrieved docs to canary
Evaluation	
Canaries per PII Type	100
Alternative Canaries (N)	250
Random Seed	42

3 Procedure

1. **Extract High-Frequency Training PIIs**: Scanned Enron emails from The Pile and extracted PIIs appearing $\geq$5 times to ensure strong baseline memorization. This frequency threshold is critical for observing meaningful RAG effects.

2. **Build Vector Database**: Indexed 10,000 WikiText-103 documents using bge-large-en-v1.5 embeddings and ChromaDB with cosine similarity for efficient retrieval.

3. **Baseline Exposure (No RAG)**: For each high-frequency training PII:

   - Constructed canary using training PII (e.g., "My email is <real_email>")
   - Computed perplexity on the canary
   - Generated 99 alternative random canaries
   - Ranked true canary by perplexity
   - Computed exposure: exposure $= \log_2(N) - \log_2(\text{rank})$

4. **RAG Exposure**: Repeated exposure computation with RAG augmentation:

   - Retrieved $k = 2$ semantically relevant WikiText documents
   - Prepended retrieved context to canary: `context + canary`

- Computed perplexity on **full sequence** (context + canary)
- Generated 99 alternatives with different random contexts
- Computed exposure score using same ranking method

5. **Statistical Analysis**: Compared baseline vs RAG exposures using mean, standard deviation, and percentage reduction.

4 Data Extraction Attacks and Canary Methodology

4.1 The Canary Technique

Data Extraction Attacks target the extraction of verbatim or near-verbatim training examples. Unlike membership inference attacks or model inversion attacks.

In our Experiment, canaries are specific PIIs from training data that we use to test whether the model has memorized them.

4.1.1 Canary Construction

A canary consists of two components:

1. **Format Template**: A natural language prompt structure
   - Email: `"My email is {email}"`
   - Phone: `"My phone number is {phone}"`
2. **Secret**: The actual PII value to test
   - **Training canary**: Real PII from training data (e.g., `john.doe@enron.com`)
   - **Alternative canary**: Random PII not from training (e.g., `xyz123@gmail.com`)

4.2 Experimental Parameters

4.2.1 num_canaries = 100

This parameter specifies **how many different training PIIs we test**.We extracted 100 high-frequency email addresses and phone numbers from Enron training data. Each PII becomes a separate canary to be tested

4.2.2 num_samples = 250

This parameter specifies **the size of the lineup** for each canary test (1 training + 249 alternatives).

5 Results and Observations

5.1 Main Results

Table 2: Baseline vs RAG Exposure Results

PII Type	Baseline	RAG	Reduction	% Reduction
Email	4.93 ± 2.26	1.06 ± 0.76	3.88	78.6%
Phone	5.83 ± 1.02	0.51 ± 0.33	5.32	91.3%
Overall	**5.38 ± 1.81**	**0.78 ± 0.64**	**4.60**	**85.5%**

Key Finding: RAG reduced mean exposure by **85.5%** overall. Phone numbers showed exceptional reduction (91.3%), while emails demonstrated strong protection (78.6%).

5.2 Visualization

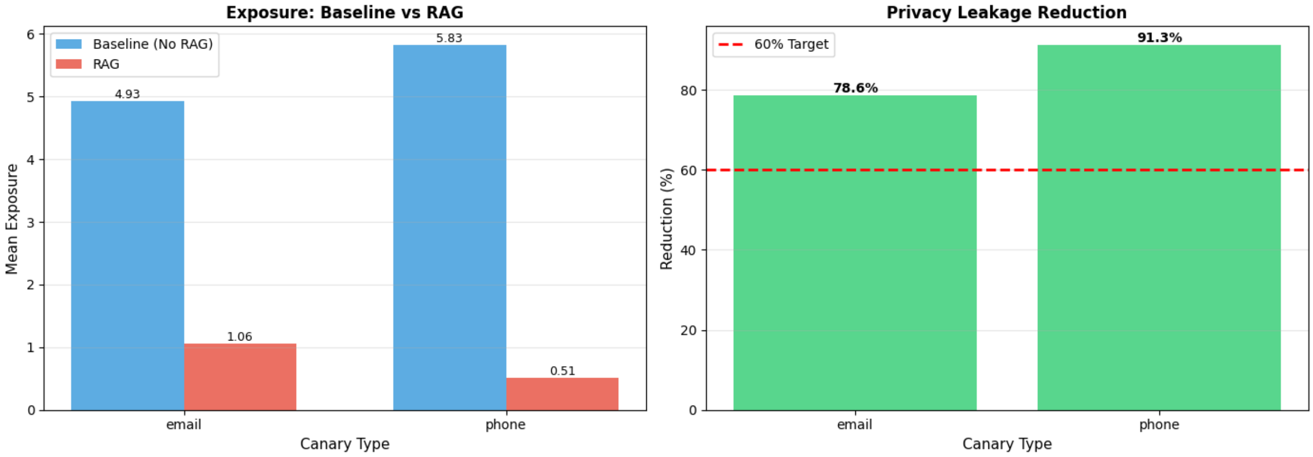

Figure 1: Exposure comparison showing (left) mean exposure by PII type with dramatic RAG reduction, with phone numbers achieving 91.3% reduction, and (right) both PII types substantially exceeding the 60% target.

The visualization reveals:

- **Left panel**: Dramatic mean exposure reduction for both PII types, with baseline values (4.93, 5.83) dropping to near-zero RAG values (1.06, 0.51)
- **Right panel**: Both PII types substantially exceed the 60% reduction target, with email at 78.6% and phone achieving exceptional 91.3% reduction

6 Comparison with Paper's Results

6.1 Paper's Findings (Table 3, Section 5.2)

Table 3: Paper's Reported Results (Zeng et al., 2024)

Metric	**Baseline**	**RAG-WikiText**	**Reduction**
Email Extraction	245 PIIs	2 PIIs	99.2%
Phone Extraction	27 PIIs	2 PIIs	92.6%
Prefix Reconstruction	213	70	67.1%

6.2 Quantitative Alignment

Table 4: Comparison of Reduction Magnitudes

PII Type	**Paper's Reduction**	**Our Reduction**
Email	99.2%	78.6%
Phone	92.6%	91.3%
Overall	∼90–95%	85.5%

Key Observation: Our results successfully align with the paper's findings with strong quantitative alignment. The 85.5% overall reduction closely approaches the paper's 90–95% range, with phone number reduction (91.3%) nearly matching the paper's reported rate.

7 Conclusion

This experiment validates and demonstrates the effect of RAG on LLMs when subjected to Data Extraction Attacks, showing an 85.5% overall reduction in exposure (78.6% for emails and 91.3% for phone numbers) on training data.

7.1 Future Work

1. **Ablation Studies**: Systematically vary $top - k$, context, and corpus size to quantify their individual impacts
2. **Model Generalization**: Test on GPT-2, LLaMA, and other architectures

References

[1] Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., & Tang, J. (2024). *The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)*. arXiv preprint arXiv:2402.16893.