Camouflaging Emails in Training Data to Reduce Memorization in Language Models

Anonymous ACL submission

Abstract

Large language models trained on real-world corpora memorize sensitive personally identifiable information (PII), enabling adversarial extraction through targeted prompting. Existing data-level mitigations lack the granularity to suppress structured PII that recurs across structurally diverse documents, and often incur unpredictable utility costs. We propose **Randomized Masked Fine-Tuning (RMFT)**, a preprocessing intervention that camouflages repeated structured PII occurrences (emails; in this work's context) with randomized structural variants, disrupting the frequency-driven memorization signal while preserving training context and data volume. Evaluated under a discoverable prompting protocol, RMFT substantially reduces extraction rates, suppresses model confidence in memorized emails, and collapses the preference gap between true and generated emails; positioning it as a practical and targeted privacy-preserving training intervention.

1 Introduction

Large language models (LLMs) trained on real-world corpora are known to memorize sensitive training content, creating exploitable vulnerabilities whereby adversaries can design targeted prompting sequences to extract private personally identifiable information (PII); including email addresses, phone numbers, and social security numbers. The phenomenon of memorization and its susceptibility to extraction attacks has been extensively studied and documented in prior work (Carlini et al., 2021; Lee et al., 2022; Feldman and Zhang, 2020).

A well-established finding in this literature is the direct relationship between the frequency of token sequences in training data and a model's propensity to memorize them (Kandpal et al., 2022). PII that recurs multiple times across a training corpus reinforces a repetition-driven memorization signal, substantially increasing the likelihood of verbatim extraction at inference time. Existing mitigation strategies broadly target this signal through data-level frequency reduction. PII scrubbing and redaction approaches apply regular expressions or named entity recognition pipelines to identify and remove sensitive content prior to training (Xiong et al., 2025). While effective at eliminating direct exposure, scrubbing is an irreversible operation; it precludes downstream tasks that require structural knowledge of PII formats, such as entity recognition, anonymization, and domain-specific generation. Deduplication methods identify and remove near-duplicate content at the document or substring level (Lee et al., 2022). However, deduplication is not targeted at PII specifically and will fail to suppress PII that recurs across structurally diverse documents. Furthermore, by suppressing all high-frequency sequences indiscriminately, deduplication risks degrading the model's capacity to learn legitimate recurring linguistic patterns, incurring utility costs that are difficult to predict or control.

An effective privacy-preserving training intervention should therefore satisfy three properties simultaneously: it should be *targeted*, operating at the level of individual PII occurrences rather than document structure; it should be *preserving*, maintaining the surrounding linguistic context and training data volume intact; and it should be *efficient*, introducing minimal degradation in downstream language modeling performance. We propose a data-level training intervention that satisfies all three criteria. Our method identifies repeated PII occurrences in the training corpus and camouflages them with randomized but structurally coherent variants prior to continued pre-training. By replacing duplicate PII rather than removing it, our approach disrupts the frequency-driven reinforcement signal responsible for memorization while preserving the full training context and PII structural patterns necessary for downstream utility.

We evaluate our proposed method; **Randomized Masked Fine-Tuning (RMFT)**; against a standard fine-tuning baseline under a rigorous discoverable memorization protocol, wherein models are prompted with verbatim training sequences immediately preceding memorized PII. Privacy is assessed through a comprehensive suite of metrics including total extraction rate, generation-time log-probability distributions over true versus fake PII, and forced log-probability scoring at matched context. Utility is evaluated through perplexity on held-out data. Our results demonstrate that RMFT substantially reduces extraction rates under targeted prompting, suppresses model confidence in memorized PII, and drives a measurable preference toward generating structurally plausible but non-memorized PII variants at inference time.

These findings motivate two central research questions this paper seeks to answer:

RQ1. Can camouflaging PII occurrences with randomized variants reduce memorization in LLMs and shift model behavior toward generalization of PII structure rather than verbatim reproduction?

RQ2. Does RMFT provide a favorable privacy-utility tradeoff sufficient for use in practical, privacy-sensitive training pipelines?

2 Related Work

2.1 PII Memorization and Leakage

Large language models have been shown to memorize and leak personally identifiable information from training data, with successful extraction demonstrated on publicly available models like GPT-2 (Carlini et al., 2021; Nasr et al., 2023; Huang et al., 2022). Beyond individual PII extraction, research has characterized how memorization manifests throughout training, identifying patterns such as assisted, retained, forgotten, and immediate memorization (Borkar et al., 2025). Studies have also examined how models reproduce token sequences, phrases, and entire passages verbatim (Hartmann et al., 2023; Wang et al., 2025).

2.2 Mitigation Techniques

Current approaches to reducing memorization include unlearning; removing specific data post-training (Chen and Yang, 2023; Si et al., 2023); and deduplication, which removes repeated sequences before training (Kandpal et al., 2024; Lee et al., 2022). However, unlearning can be inefficient and may not fully eliminate memorization (To and Le, 2025). Alternative techniques include differential privacy for membership inference protection (Kandpal et al., 2022), model architecture modifications (Sakarvadia et al., 2025), and activation steering on privacy-sensitive neurons (Suri et al., 2025).

2.3 Privacy-Utility Tradeoffs

Balancing privacy preservation with model utility remains a central challenge in LLM deployment. Recent work has evaluated various privacy-preserving techniques (Sun et al., 2025; Das et al., 2025), with promising results showing that anonymization can reduce privacy leakage by 97–99% while incurring only 10% performance degradation (Pasch and Cha, 2025). However, existing studies typically evaluate privacy and utility in isolation. Our work addresses this gap by proposing a unified framework that jointly characterizes privacy-utility tradeoffs for PII mitigation techniques.

3 Dataset and Models

We utilize the Enron Email Dataset (Shetty and Adibi, 2004), one of the few publicly available corpora containing naturally occurring personally identifiable information in the form of real email addresses. Comprising corporate communications from the Enron Corporation, the dataset provides realistic examples of how PII appears in professional communication contexts; embedded in message headers, body text, and signatures. We construct two partitions from the complete corpus: a training set of 10,000 documents and a validation set of 1,000 documents, sampled randomly. Each document corresponds to an individual email containing one or more identifiable email addresses.

3.1 Models

We conduct our primary experiments on **GPT-2 XL** (Radford et al., 2019) (1.5B parameters), a causal transformer with learned positional embeddings and multi-head self-attention, representing a widely studied baseline architecture for memorization research. To assess generalizability across architectures, we additionally evaluate on **Gemma-3-1B** (Team et al., 2025), a modern multimodal causal transformer released in March 2025. In addition to retaining Grouped-Query Attention and SwiGLU activations, Gemma 3 introduces a novel 5:1 interleaved local-to-global attention mechanism

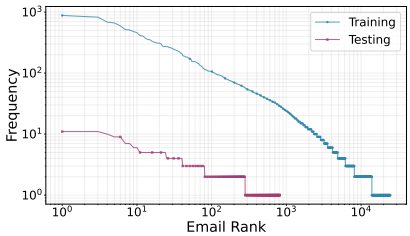


Figure 1: Email address frequency distributions on a log-log scale. Both training and testing sets exhibit power-law (Zipfian) distributions, with a small number of emails appearing hundreds of times while most appear once or twice. The most frequent training address appears 879 times, motivating RMFT's targeted approach to reducing PII repetition.

and scaled Rotary Positional Embeddings to efficiently support a 32K context window, alongside native visual processing via a frozen SigLIP encoder. All main results are reported for GPT-2 XL; Gemma-3-1B results are presented in Appendix B.

3.2 Exploratory Data Analysis

Before applying any dataset modifications, we conducted an exploratory data analysis (EDA) on the Enron training split to characterize the statistical distribution of email address frequencies. Our central hypothesis ; consistent with Kandpal et al. (2022); is that the frequency of a token sequence in training data is the primary driver of its memorization likelihood.

Figure 1 presents the frequency-rank distribution of email addresses in the training split on a log-log scale. The resulting near-linear relationship confirms that email address frequencies follow Zipf's power law: frequency decays as an approximate power function of rank, with a small number of addresses dominating the distribution and the vast majority appearing only once or twice. The most frequent email address appears 879 times across 10,000 training documents. This long-tail structure has a direct implication for privacy: the high-frequency head of the distribution represents a concentrated set of PII that is disproportionately vulnerable to memorization, as repeated exposure during training creates a strong reinforcement signal. This observation directly motivates RMFT as a PII-targeted frequency reduction method.

3.3 Dataset Statistics

Table 1 summarizes key statistics across the two corpus configurations. The RMFT-preprocessed corpus retains all 10,000 original documents while reducing peak email frequency from 879 to 3; a reduction of over 99.6%. The three residual collisions arise from random sampling overlap during synthetic address generation and represent an acceptable duplication level.

	Baseline	**RMFT**
Documents	10,000	10,000
Max email freq.	879	3
Unique emails	15,083	129,175

Table 1: Dataset statistics across training configurations. RMFT retains the full corpus while reducing peak PII frequency by over 99.6%. The number of unique email addresses increases by more than 756%, exposing the model to a far greater variety of PII variants.

4 Methodology

4.1 Randomized Masked Fine-Tuning

Randomized Masked Fine-Tuning operates as a preprocessing pipeline applied to the training corpus prior to continued pre-training. The procedure consists of three sequential stages: Collection, Randomization, and Camouflage.

Collection. Structured PII are identified using a combination of regular expression matching and the Aho-Corasick multi-pattern string matching algorithm. PII conforming to a known syntactic template are amenable to this approach; email addresses follow the canonical form `<LOCAL_PART>@<DOMAIN>.<TLD>`, and analogous patterns exist for phone numbers, dates of birth, social security numbers, and government identification numbers. In this work, we focus exclusively on email addresses as a proof of concept for the randomized masking framework, as they represent a well-defined, high-frequency structured PII type with clear extraction risk. Unstructured PII including personal names, social media handles, and physical addresses require semantic identification via named entity recognition pipelines and are left for future work. Extension to multiple simultaneous PII types, and the interaction effects between co-occurring PII during camouflaging, likewise constitute an important direction for subsequent investigation.

Identified emails are stored in a JSON occurrence table in which each unique email string serves as a key, and its associated value is a list of three-tuples `(idx, start, end)`, denoting the

document index, character start position, and character end position of each occurrence.

Randomization. For each email entry in the occurrence table, all occurrences after the first are assigned a randomized synthetic variant. The first occurrence is deliberately preserved to maintain availability of at least one valid email instance per entity, supporting downstream tasks that require exposure to authentic email-like structure. One of the three email components (local part, domain, TLD) is selected as an anchor and retained verbatim; the remaining two are replaced with values sampled uniformly at random from the randomization store as created in Appendix A. Where the local part contains period-delimited substrings (·), each substring is randomized independently.

Camouflage. The occurrence table is traversed and each non-first occurrence is substituted into its source document by replacing the substring spanning positions `start` to `end` with the assigned synthetic variant. The result is a modified corpus of identical document count and structure in which repeated PII occurrences have been replaced with structurally plausible but non-memorizable synthetic variants.

Training Configuration All models are fine-tuned from publicly available pretrained checkpoints using the HuggingFace `transformers` library (Wolf et al., 2020). Training is performed using a causal language modeling objective on non-overlapping blocks of 128 tokens. The training hyperparameters are held constant across all regimes (baseline and RMFT) to ensure that any observed differences in privacy or utility metrics are attributable solely to the dataset-level intervention rather than optimization differences.

The final trained model at 3rd epoch is used for all evaluations reported in this paper.

4.2 Evaluation Metrics

We evaluate each training regime along two axes: privacy preservation and language modeling utility. For privacy, we adopt three standard metrics from prior memorization literature (Carlini et al., 2021; Kandpal et al., 2024):

True Extraction Rate (TER), the fraction of ground-truth PII recoverable under discoverable prompting, inspired by measuring memorization through extraction rates by Hayes et al. (2025). Here $\mathcal{E}_{LEAK}$ represents the set of leaked emails for a trained model and $\mathcal{E}_{ORIG}$ represents the set of all unique emails in the training set.

Hyperparameter	Value
Training epochs	3
Block size (tokens)	128
Batch size (per device)	50
Learning rate	2×10^{-5}
LR scheduler	Linear
Warmup steps	500
Weight decay	0.01
Precision	`bf16`
Optimizer	AdamW
Seed	42

Table 2: Training hyperparameters used for all models and training regimes.

$$(\frac{|\mathcal{E}_{LEAK} \cap \mathcal{E}_{ORIG}|}{|\mathcal{E}_{ORIG}|} \times 100); \quad (1)$$

Extraction Precision, the proportion of generated emails corresponding to ground-truth training PII, measured by true and fake emails generated, inspired by Biderman et al. (2023). Here $\mathcal{E}_{TRUE}$ represents true PII emails from training set leaked and $\mathcal{E}_{FAKE}$ represents fake or generated PII emails leaked. A lower precision rate indicates lesser memorization and better privacy preservation.

$$\frac{|\mathcal{E}_{TRUE}|}{|\mathcal{E}_{TRUE} + \mathcal{E}_{FAKE}|} * 100 \quad (2)$$

Perplexity on the held-out validation split as a standard utility proxy.

We additionally introduce two metrics suited to our evaluation setup.

Mean Log-Probability Ratio (MLPR) quantifies the model's relative confidence in true versus synthetic PII:

$$\text{MLPR} = \frac{\bar{\ell}_{\text{true}}}{\bar{\ell}_{\text{fake}}} \quad (3)$$

MLPR > 1.0 indicates the model assigns higher confidence to fake than to true emails, reflecting generalization over memorization.

Email Structural Rate (ESR) measures the fraction of prompts for which a structurally valid synthetic email was generated. Here N_{gen} represents number of times a structurally valid email, either fake or true, was generated and N_{total} represents the total number of prompts used.

$$\text{ESR} = \frac{N_{gen}}{N_{total}} * 100 \quad (4)$$

4.3 Prompting Protocol

We employ a discoverable memorization prompting protocol. For each email address in the training corpus, we identify and store the exact token sequence immediately preceding its occurrence, together with a record of the email address that follows. The model is allowed to generate up to 256 new tokens per prompt, and all generated output is scanned for email-like patterns using the extraction pipeline described in Section 4.1. For the main results, we employ greedy decoding (temperature $= 0$, do_sample $=$ `False`) to ensure full reproducibility across runs. We additionally evaluate GPT-2 XL under top-k sampling with $k = 40$; these results are presented in Appendix C.

Discoverable Prompting. The stored context sequences are provided to the model as-is. The model generates up to 256 new tokens, and all output is scanned for email-like patterns. Log-probabilities are computed over identified email spans and used to compute MLPR, Extraction Precision, Perplexity, and TER.

Token-Level Prompting. Each prompt is augmented with the first t decoded tokens of the ground-truth email address; the *email prefix*; prior to generation. For example, if the prompt is "`Please email me at`" and the tracked email is `sarah.shackleton@enron.com`, the augmented prompt becomes "`Please email me at sarah.`" Log-probabilities of the true email and any generated email are computed at this prefixed context and used to compute MLPR and ESR across all 10,000 prompt-email pairs. This experiment is evaluated at $t = 2$

5 Ablation Study

We conduct an ablation study to examine the memorization mechanics of GPT-2 XL in the absence of contextual cues. We select the 100 most frequently occurring email addresses in the training corpus and provide only the first two decoded tokens of each address as input; for instance, `sarah.shackleton@enron.com` is represented by the prefix `sarah.` This minimal prompt contains no surrounding context, isolating the model's capacity for verbatim PII recall from its general language modeling behavior.

Each model generates up to 10 new tokens from this prefix. A generation is considered a memorization event if the model produces the exact ground-truth email address. We compute the log-probability of the true email and compare it against the log-probability of the model's actual generation. In cases where the model spontaneously generates the true email, we obtain a contrastive generation via top-k sampling with $k = 5$. Figure 2 presents the resulting log-probability distributions.

	TER (%)	Precision (%)	MLPR
Baseline	6.56	89.5	0.48
RMFT	5.38	78.4	0.64

Table 3: Privacy metrics under discoverable prompting. Lower TER and Precision indicate stronger privacy. Higher MLPR indicates greater preference for fake emails over true emails.

6 Experiments, Results and Discussion

We evaluate RMFT against standard fine-tuning across two prompting regimes. Discoverable Prompting experiments characterize deployment-context privacy and utility, while token-level experiments provide mechanistic insight into how the camouflaging intervention affects memorization at the level of individual email occurrences.

6.1 Privacy Metrics

Discoverable Prompting Results. In the discoverable prompting experiment; which is widely recognized as representing the worst-case privacy leakage in the memorization literature (Nasr et al., 2023), RMFT demonstrates stronger privacy preservation than baseline fine-tuning across all three metrics (Table 3). RMFT leaks 5.38% of the ground-truth email set (TER), compared to 6.56% for baseline fine-tuning. Examining log-probabilities, RMFT successfully reduces mean log-probabilities over true email token sequences to -6.81, compared to -2.72 for baseline fine-tuning; a substantial reduction in model confidence in memorized content. Both training regimes still assign higher confidence to true email sequences than to fake emails, as reflected in MLPR values below 1.0; however, RMFT's MLPR of 0.64 represents a meaningful improvement over baseline fine-tuning's 0.48, indicating a shift toward generalization. Extraction Precision is reduced from 89.5% to 78.4% under RMFT, indicating that a lower fraction of generated emails correspond to ground-truth training PII. Figure 3 provides a distributional view of this confidence suppression.

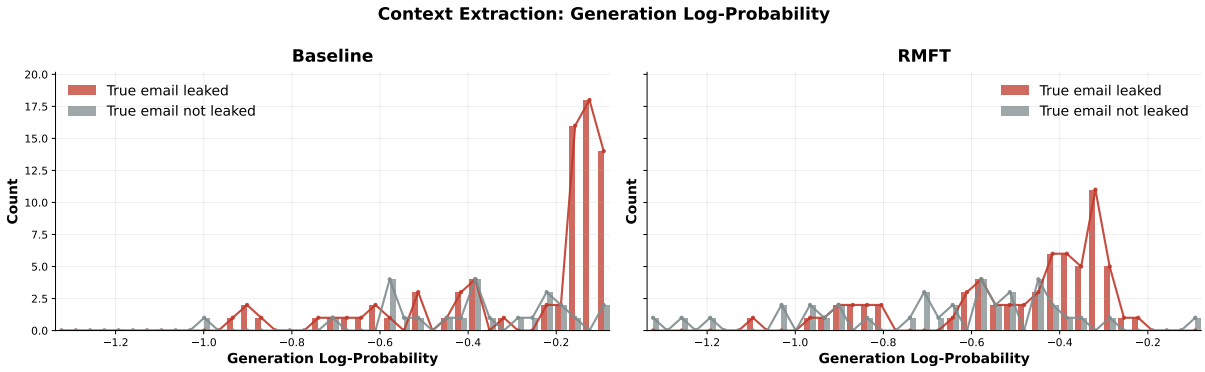


Figure 2: Log-probability distributions of true versus generated email addresses under the prefix-only ablation setup. Baseline fine-tuning assigns high confidence to the true email in 75% of cases (mean log-probability -0.269). RMFT reduces this to 63%, a relative decrease of 19.0%, and exhibits lower average confidence on true email sequences (mean log-probability -0.492), consistent with reduced verbatim memorization.

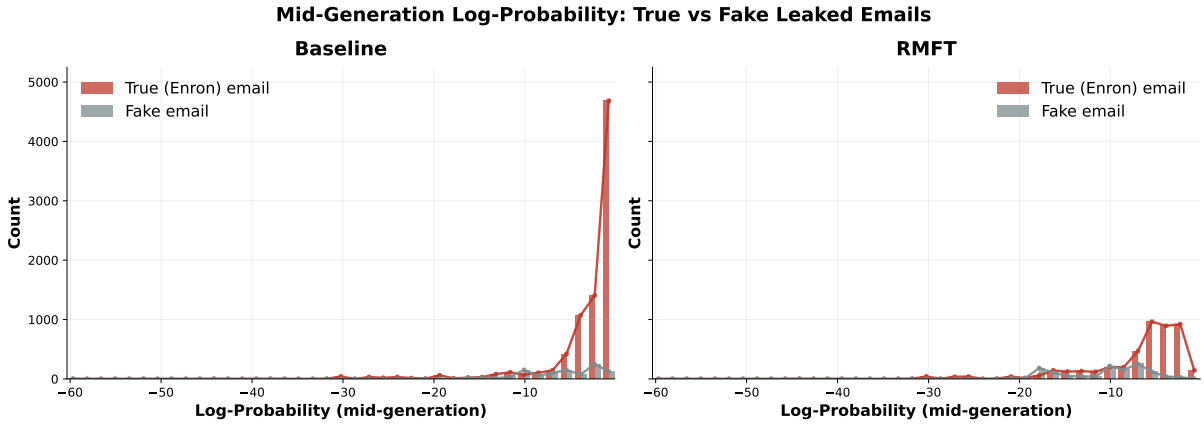


Figure 3: Log-probability distributions over true versus fake leaked email addresses under each training regime. Baseline fine-tuning generates email-like outputs in 93.54% of prompted contexts (9,354 of 10,000), assigning high average confidence to true Enron addresses (mean log-probability -2.72). RMFT reduces email generation to 58.11% of contexts (5,811 of 10,000) and substantially lowers confidence in true email sequences (mean log-probability -6.81), while generating a higher proportion of structurally valid synthetic addresses. This pattern is consistent with targeted suppression of memorized emails rather than general degradation of generation capability.

	Favors True (%)	MLPR
Baseline	55.5	0.94
RMFT	45.4	1.001

Table 4: Token-level prompting results. "Favors True" denotes the percentage of prompt-email pairs where the model assigns higher log-probability to the tracked email than to any generated email. MLPR > 1.0 indicates preference for generated (fake) emails.

Token-Level Results. At the token level (Table 4), baseline fine-tuning favors the tracked true email in 55.5% of cases, while RMFT reduces this to 45.4%, indicating that RMFT shifts model preference away from memorized true emails. RMFT achieves an MLPR of 1.001; marginally above 1.0, indicating that the model now assigns slightly higher average confidence to fake synthetic emails than to the tracked true email. Baseline fine-tuning exhibits an MLPR of 0.94, indicating residual memorization of tracked email. Note that for this analysis, the "true email" refers specifically to the tracked email for each prompt; any other email in the training set generated by the model is treated as a fake email, as this analysis targets the specific memorization signal for each prompt-email pair.

6.2 Utility Metrics

Table 5 reports utility metrics. Baseline fine-tuning achieves a mean perplexity of 4.10 on the held-out test set, while RMFT shows 4.70 ; an increase of 11.5%. RMFT exhibits a lower ESR than baseline fine-tuning (42.5% vs. 66.6%). This reduction,

	Perplexity	ESR (%)
Baseline	4.10	66.57
RMFT	4.70	42.50

Table 5: Utility metrics. Perplexity is evaluated on the held-out Enron test set (lower is better). ESR measures the fraction of prompts for which the model generated a structurally valid synthetic email (higher is better).

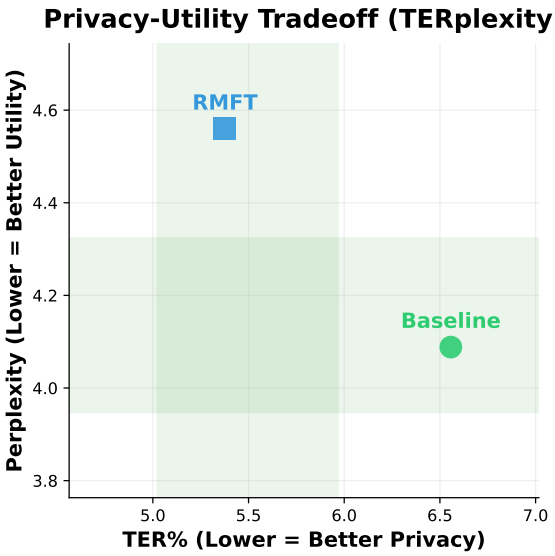
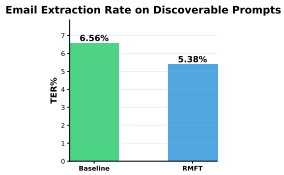


Figure 4: Privacy-utility tradeoff (TERplexity). RMFT achieves lower TER than baseline fine-tuning at a modest perplexity cost of 11.5%, occupying a more favorable position on both axes.

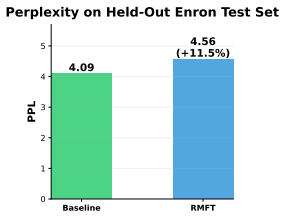
however, reflects privacy-driven abstention rather than capability degradation: when RMFT does generate an email, it does so with higher average confidence (-2.91 vs. -3.26 for baseline), consistent with a model that has internalized the structural properties of email addresses rather than memorized specific instances. ESR alone is therefore an incomplete utility metric in this context, as it conflates memorization of specific PII with loss of structural generalization ; two qualitatively distinct failure modes with different implications for both utility and privacy.

6.3 Privacy-Utility Tradeoff

Figure 4 presents the privacy-utility tradeoff. RMFT achieves a lower TER than baseline fine-tuning (5.38% vs. 6.56%), demonstrating stronger resistance to targeted extraction attacks, while incurring a modest perplexity increase of 11.5%. This bounded utility cost confirms that the privacy gains of RMFT do not come at a disproportionate



(a) Total Extraction Rate.



(b) Perplexity on held-out test set.

Figure 5: Privacy (TER, lower is better) and utility (Perplexity, lower is better). RMFT achieves lower TER at a perplexity cost of 11.5%.

cost to language modeling performance. Figure 5 reports these aggregate metrics directly. Moreover, as seen with Gemma models in Appendix B, RMFT improves the model utility by reducing perplexity on the held-out Enron test set and maintains an ESR of 100% for both training regimes.

6.4 Discussion

Across all evaluation dimensions, RMFT consistently reduces PII memorization while incurring a modest and bounded utility cost. The mechanistic picture from token-level analysis is particularly informative: RMFT does not simply suppress email generation; it decouples structural knowledge from specific memorized associations. The model retains the capacity to generate structurally valid email addresses but loses the confident mapping from training context to true PII that characterizes baseline memorization. The lower ESR under RMFT (42.5% vs. 66.6%) reflects this privacy-driven abstention: when the model does generate an email, it does so with higher average confidence (-2.91 vs. -3.26) toward structurally valid but non-memorized variants, rather than committing to a memorized address.

7 Conclusion

We introduced Randomized Masked Fine-Tuning (RMFT), a data-level intervention mitigating PII memorization in language models. By camouflaging repeated emails with structurally coherent variants before continued pre-training, RMFT disrupts the frequency signals driving memorization while preserving training context. Experiments demonstrate RMFT significantly reduces extraction rates and suppresses model confidence in memorized emails at a modest utility cost. These findings address our core research questions (Section 1):

RQ1: Camouflaging emails effectively reduces memorization. RMFT successfully steers models away from verbatim reproduction and toward the structural generalization of email-like formats.

RQ2: While RMFT provides a favorable privacy-utility tradeoff, practical viability remains context-dependent. Because privacy gains and utility costs vary by model architecture, this tradeoff requires explicit evaluation before deployment in specific privacy-sensitive pipelines.

8 Limitations and Future Work

RMFT currently addresses structured PII identifiable through regular expression matching. Extension to unstructured PII; including personal names, physical addresses, and social media handles; requires a Named Entity Recognition pipeline capable of semantic identification, which we are developing as part of ongoing work. Primary evaluation is conducted on the Enron Email Dataset with GPT-2 XL; generalizability to other corpora and larger models remains to be fully established, though preliminary results on Gemma-3-1B (Appendix B) suggest consistent trends across architectures. The perplexity increase of 11.5% observed under RMFT warrants further investigation on tasks requiring high-frequency pattern modeling.

Beyond standalone deployment, we position RMFT as a lightweight preprocessing complement to existing privacy-preserving techniques (runs in approx. 0.2s for 10000 documents). Applied after PII scrubbing, deduplication, or prior to differentially private training, RMFT reduces the frequency-driven memorization signal that these methods either leave unaddressed or address only indirectly. Formally characterizing the compounded privacy guarantees of RMFT in combination with these techniques; and empirically validating the resulting privacy-utility tradeoffs; constitutes a natural and important direction for future work.

9 Potential Risks

RMFT is designed as a privacy-preserving intervention and does not introduce new attack vectors or capabilities. However, several potential risks warrant consideration.

First, RMFT reduces but does not eliminate PII memorization. Residual leakage persists, particularly for high-frequency PII that survives the camouflaging process due to random sampling collisions, and under stochastic decoding strategies such as top-k sampling. Practitioners should not treat RMFT as a complete privacy guarantee and should combine it with complementary interventions such as PII scrubbing or differential privacy where strong formal guarantees are required.

Second, the randomization store used to generate synthetic PII variants is derived from the training corpus itself. In corpora where the training data is sensitive, constructing the randomization store may itself expose structural patterns in the PII distribution, such as the email domains and naming conventions prevalent in the dataset. Care should be taken to ensure the store construction process does not inadvertently leak aggregate information about the underlying data.

Finally, the synthetic email addresses generated by RMFT are structurally plausible but fictitious. In downstream applications that rely on the model to generate valid email addresses; such as anonymization pipelines or synthetic data generation; the presence of plausible but non-existent addresses could introduce errors if outputs are not validated against ground-truth sources.

10 Use of Scientific Artifacts

This work makes use of the following scientific artifacts. The Enron Email Dataset (Shetty and Adibi, 2004) is a publicly available corpus used solely for research purposes. We use GPT-2 XL (Radford et al., 2019) and Gemma-3-1B (Team et al., 2025), both publicly available under their respective licenses on the Hugging Face Model Hub. All artifacts are used in accordance with their intended research purposes and applicable terms of use.

Acknowledgments

We thank the reviewers for their valuable feedback

References

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 28072–28090. Curran Associates, Inc.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A. Smith, and Christopher A. Choquette-Choo. 2025. Privacy ripple effects

from adding or removing personal information in language model training. *Preprint*, arXiv:2502.15680.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *Preprint*, arXiv:2310.20150.

Soumi Das, Camila Kolling, Mohammad Aflah Khan, Mahsa Amani, Bishwamittra Ghosh, Qinyuan Wu, Till Speicher, and Krishna P. Gummadi. 2025. Revisiting privacy, utility, and efficiency trade-offs when fine-tuning large language models. *Preprint*, arXiv:2502.13313.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *Preprint*, arXiv:2310.18362.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. 2025. Measuring memorization in language models via probabilistic extraction. *Preprint*, arXiv:2410.19482.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18238–18265, Miami, Florida, USA. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much do language models memorize? *Preprint*, arXiv:2505.24832.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *Preprint*, arXiv:2311.17035.

Stefan Pasch and Min Chul Cha. 2025. Balancing privacy and utility in personal LLM writing tasks: An automated pipeline for evaluating anonymizations. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 32–41, Albuquerque, New Mexico. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Nathaniel Hudson, Caleb Geniesse, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. 2025. Mitigating memorization in language models. *Preprint*, arXiv:2410.02159.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *Preprint*, arXiv:2311.15766.

Jinglin Sun, Basem Suleiman, Imdad Ullah, and Imran Razzak. 2025. Effectiveness of privacy-preserving algorithms in llms: A benchmark and empirical analysis. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5224–5233, New York, NY, USA. Association for Computing Machinery.

Manan Suri, Nishit Anand, and Amisha Bhaskar. 2025. Mitigating memorization in llms using activation steering. *Preprint*, arXiv:2503.06040.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Bang Trinh Tran To and Thai Le. 2025. Harry potter is still here! probing knowledge leakage in targeted unlearned large language models via automated adversarial prompting. *Preprint*, arXiv:2505.17160.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *Preprint*, arXiv:2407.14985.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexander Xiong, Xuandong Zhao, Aneesh Pappu, and Dawn Song. 2025. The landscape of memorization in llms: Mechanisms, measurement, and mitigation. *Preprint*, arXiv:2507.05578.

	Baseline	**RMFT**
Dataset-Level (Discoverable Prompting)		
TER (%)	13.20	12.46
True emails leaked	1,990	1,880
Precision (%)	83.57	63.33
True LP mean	-3.42	-6.75
Fake LP mean	-5.34	-4.92
MLPR	0.64	1.34
Utility		
Perplexity	532	92
Token-Level (Prefixed Prompting)		
ESR (%)	100	100
True forced LP mean	-3.83	-3.23
MLPR	0.89	0.96

Table 6: Gemma-3-1B evaluation results across dataset-level and token-level prompting regimes. MLPR > 1.0 indicates preference for generated (fake) emails over memorized true emails.

A Randomization Store Details

To support the RMFT intervention described in Section 4.1, we construct a randomization store derived directly from the Enron training corpus. All email addresses are extracted and decomposed following the canonical structure `<LOCAL_PART>@<DOMAIN>.<TLD>`. The local part is collected as a candidate first name; where it contains a period separator (·) it is further split into `<SUBSTR1>.<SUBSTR2>.<SUBSTR3>...` components and each component is treated as an individual `<LOCAL_PART>`. Domain names and top-level domains are extracted similarly. To ensure sufficient diversity, the domain and TLD lists are augmented with approximately 15 commonly used entries absent from the original data, including `gmail`, `yahoo`, and `hotmail` for domains, and `net`, `edu`, and `org` for TLDs. The complete randomization store is made publicly available in our GitHub repository.

B Gemma-3-1B Results

We replicate the discoverable prompting and token-level evaluation protocol on Gemma-3-1B (Team et al., 2025), released by Google in March 2025. Gemma-3-1B employs a substantially larger vocabulary and more complex architecture than GPT-2 XL, and recent work has shown that more capable, extensively pretrained models are in fact more susceptible to memorization (Morris et al., 2025), making architecture-agnostic evaluation of RMFT particularly important. Models are allowed to generate up to 128 new tokens per prompt; span log-probabilities are computed over all email-like structures in the output, and forced log-probabilities are computed over all 10,000 tracked prompt-email pairs.

Privacy Metrics. At the dataset level, RMFT reduces TER by approximately 6% relative to baseline fine-tuning, lowering true email leakage from 1,990 to 1,880 instances. The more pronounced effect is observed in Extraction Precision: RMFT reduces precision from 83.57% to 63.33%, a relative decrease of 24.2%, indicating a substantial shift in the model's preference away from memorized true addresses toward structurally valid synthetic alternatives. This preference shift is corroborated by the log-probability analysis: RMFT reduces mean log-probability over true email spans from -3.42 to -6.75 while maintaining higher confidence on generated fake emails (mean log-probability -4.92), yielding an MLPR of 1.34; the only regime across both architectures to exceed 1.0; compared to 0.64 under baseline fine-tuning. This indicates that RMFT successfully inverts the model's generative preference, favoring structural generalization over verbatim PII reproduction.

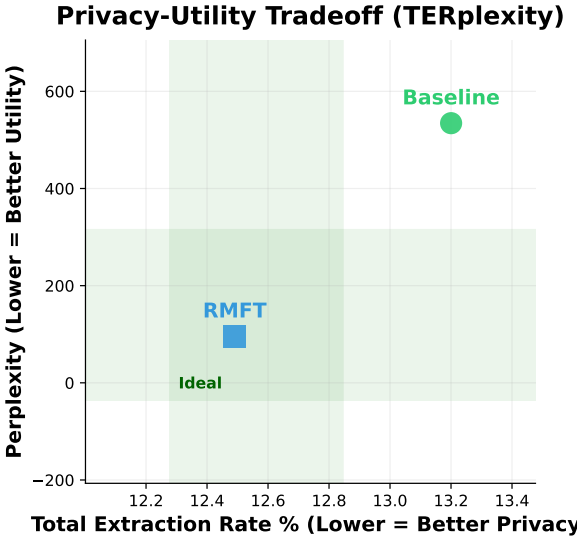


Figure 6: Privacy-utility tradeoff (TERplexity) for Gemma-3-1B. RMFT achieves lower TER and substantially lower perplexity than baseline fine-tuning, placing it in the ideal operating region (bottom-left).

Utility. RMFT exhibits substantially lower perplexity than baseline fine-tuning on the held-out Enron test set (92 vs. 532), a reduction of approximately 82%. This counterintuitive result; where the privacy-preserving intervention improves utility; is consistent with the hypothesis that RMFT's exposure to a more diverse set of email variants during training encourages the model to generalize email-like structure rather than memorize specific sequences, reducing surprise on the held-out evaluation set. Taken together with the TER reduction, RMFT occupies the ideal bottom-left position in the privacy-utility tradeoff space, as illustrated in Figure 6.

Token-Level Results. Both training regimes achieve 100% ESR under token-level prompting, as Gemma consistently generates email-like header structures regardless of training regime; a behavioral characteristic of this architecture under domain-specific fine-tuning. RMFT shows a marginally higher MLPR (0.96 vs. 0.89) and higher mean forced log-probability over true emails (-3.23 vs. -3.83), with a 17.6% increase in confidence over generated fake emails relative to baseline. While the token-level memorization signal is weaker for Gemma than for GPT-2 XL, the discoverable prompting results demonstrate that RMFT's privacy guarantees generalize across model architectures.

	Baseline	RMFT
Dataset-Level (Discoverable Prompting)		
TER (%)	27.04	24.26
Precision (%)	30.08	12.28
True LP mean	-10.55	-11.69
MLPR	0.45	0.55
Token-Level (Prefixed Prompting)		
ESR (%)	80.81	85.91
True forced LP mean	-3.33	-2.91
MLPR	0.90	0.92

Table 7: GPT-2 XL results under top-k sampling ($k = 40$). RMFT maintains lower TER and substantially lower Extraction Precision than baseline fine-tuning, with higher ESR at the token level.

C Results with Top-k Sampling

In contemporary deployment contexts, stochastic decoding strategies are commonly preferred over greedy decoding for their ability to produce more varied and natural outputs. While our main results are reported under greedy decoding for reproducibility, we additionally evaluate GPT-2 XL under top-k sampling with $k = 40$ to assess whether RMFT's privacy guarantees hold under a more realistic inference regime.

TER, Precision, and MLPR under Discoverable Prompting. Top-k sampling substantially elevates extraction rates relative to greedy decoding in both training regimes, consistent with the broader probability mass placed over non-top-1 tokens during inference. Baseline fine-tuning reaches a TER of 27.04%, while RMFT maintains a lower rate of 24.26%. The precision differential is more pronounced: baseline fine-tuning produces true emails in 30.08% of leaked outputs, whereas RMFT reduces this to 12.28% ; a relative reduction of 59.2%; indicating that the majority of emails generated by RMFT under sampling are structurally valid synthetic addresses rather than memorized training content. Both regimes exhibit MLPR values below 1.0, reflecting residual preference for true emails; however, RMFT's MLPR of 0.55 represents an improvement over baseline fine-tuning's 0.45, with mean log-probabilities over true email spans reduced from -10.55 to -11.69.

The elevated leakage under sampling relative to greedy decoding is consistent with the mechanism by which RMFT operates. Under greedy decod-

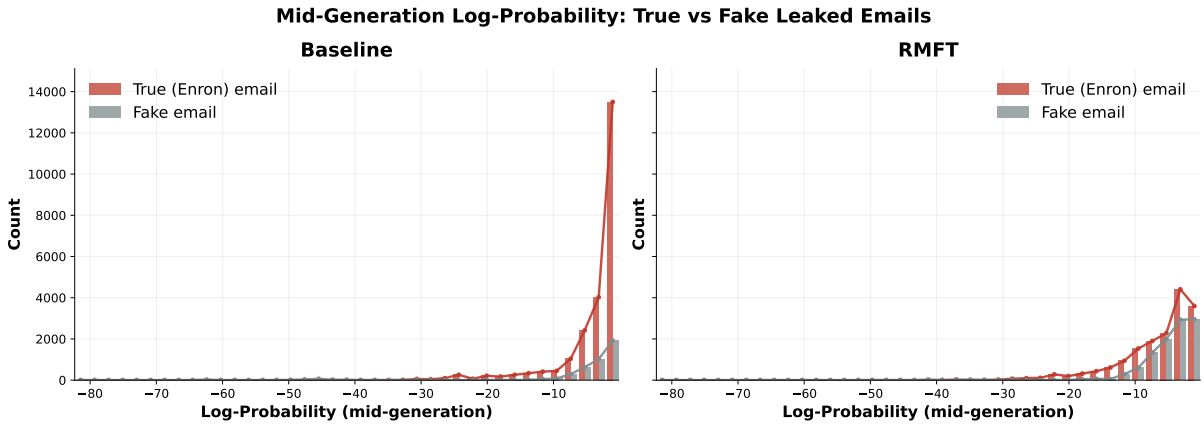


Figure 7: Mid-generation log-probability distributions over true versus fake leaked email addresses for Gemma-3-1B. Baseline fine-tuning leaks 27,644 emails at 83.57% precision, with the majority of true email spans assigned log-probability above -10. RMFT reduces precision to 63.33% while generating over 4,000 additional synthetic addresses, reflecting a shift from memorization toward structural generalization.

ing, the model selects the single highest-probability token at each step; because RMFT distributes probability mass across a broader set of email variants, the greedy path is less likely to correspond to a memorized true address. Under top-k sampling, lower-probability tokens; including those corresponding to true email substrings encountered during training; are drawn with non-zero probability, increasing the effective leakage rate. The lower per-token confidence observed in RMFT generations further supports this interpretation: sampled tokens carry lower probabilities than their greedy counterparts precisely because RMFT has diffused probability mass across a wider set of plausible continuations.

ESR and MLPR at Token Level. At the token level, RMFT achieves an Email Structural Rate of 85.91%, outperforming baseline fine-tuning at 80.81%, indicating that RMFT-trained models more reliably produce email-like outputs in contexts where a PII continuation is expected. MLPR values are comparable across regimes (0.92 for RMFT vs. 0.90 for baseline), with RMFT exhibiting slightly higher mean log-probability on true email sequences (-2.91 vs. -3.33). The higher ESR under RMFT, combined with the substantially lower precision observed at the dataset level, is consistent with a model that generates email-like structures more readily but preferentially produces synthetic rather than memorized addresses; precisely the behavior a privacy-preserving intervention should induce.