Understanding Privacy Preserving Knowledge in models via Mechanistic Interpretability

Rahul Tiwari
Virginia Tech

Harith Laxman
Virginia Tech

Abstract

Large Language Models (LLMs) have become more capable at complex tasks, leading to widespread integration across coding, education, legal assistance, and customer support. With increased adoption, they now access vast amounts of sensitive data. Prior work [3] shows that white-box mechanistic interpretability can elicit latent knowledge, but most studies rely on model organisms [4] explicitly trained to hold secrets. In real scenarios, models are instead aligned and privacy-tuned post-training. This project examines privacy-preserving properties in a realistic setting: a toy model trained for PII redaction on a PII masking dataset, then probed with mechanistic interpretability (activation and circuit analyses) to elicit PII after redaction training. The pipeline includes fine-tuning with LoRA, feature extraction via sparse autoencoders (SAEs), and feature ablation. Results show that ablating a single SAE feature (feature 3867) significantly degrades masking performance and increases PII leakage, demonstrating causal links between interpretable features and privacy behavior. Code: https://github.com/ba11b0y/candi.

1 Methodology

1.1 Training

The training process employed the Lion optimizer, a recently developed optimization algorithm that has shown improved performance compared to AdamW on various tasks. The learning rate was set to $1e{-}4$, and gradient accumulation was used with 4 micro-batches to achieve an effective batch size of 8 while maintaining a micro-batch size of 2 per forward pass. Training sequences were truncated to a maximum length of 384 tokens to balance capturing context and computational efficiency. The model was trained for 2,000 optimizer steps with gradient accumulation, processing approximately 16,000 training examples. Progress was logged every 10 steps.

After training, the LoRA adapter weights were merged into the base model to create a standalone fine-tuned model usable without PEFT. The merged model (about 5GB) and tokenizer were saved for deployment. The base model was `gpt2-small`; LoRA adapters with rank 8 were attached to transformer layers, reducing trainable parameters to roughly 0.4% of total. Gradient checkpointing further reduced memory.

The project used a subset of the **AI4Privacy PII Masking Dataset** [2]. A sample record is shown in Table 1. The dataset columns were:

- `source_text`: raw text with embedded PII tokens,
- `target_text`: masked text with placeholders (e.g., [USERNAME], [TIME]),
- `privacy_mask` and `span_labels`: span-level PII annotations,
- `mberttokens` and BIO labels for sequence tagging.

Training was evaluated with character-level F1, improving by **6.4%** over the base model.

1.2 Feature Extraction using Sparse Autoencoders

Neural networks can represent more features than neurons, complicating interpretability [1]. Sparse autoencoders (SAEs) help disentangle these features. A pretrained SAE from `gpt2-small-res-jb` targeted `blocks.7.hook_resid_pre`, expanding the 768-dimensional residual stream to 24,576 features ($32\times$). Using `HookedSAETransformer`, prompts were run with SAE hooks; activations at layer 7 were encoded, and ReLU sparsity ensured only a few features fired per input.

Two contrastive prompts isolated PII-masking features: one with masking instructions ("Mask PII data. Input: My SSN is 637-622-1778 Output: My SSN is") and one without. Feature differences at the final token identified candidates; feature 3867 showed a difference of 61.69 activation units and was selected for ablation. Maximum activation for feature 3867 was calibrated over 100 training batches (about 32,768 tokens) to scale later interventions.

Column	Sample Entry
source_text	Subject: Group Messaging for Admissions Process - Good morning, everyone. Users: wynqvrh053, luka.burg, qahil.wittauer, gholamhossein.ruschke, pdmjrsyoz1460. Times: 10:20am, 21, quarter past 13, 9:47 PM
target_text	Subject: Group Messaging for Admissions Process - Good morning, everyone. Users: [USERNAME], [USERNAME], [USERNAME], [USERNAME], [USERNAME]. Times: [TIME], [TIME], [TIME], [TIME]
privacy_mask	["value": "wynqvrh053", "start": 287, "end": 297, "label": "USERNAME", "value": "10:20am", "start": 311, "end": 318, "label": "TIME", ...]
span_labels	[[440, 453, "USERNAME"], [430, 437, "TIME"], [395, 416, "USERNAME"], ...]
mbert_text_tokens	["Sub", "#ject", ":", "Group", "Mess", "#aging", ..., "#60"]
mbert_bio_labels	["O", "O", ..., "B-USERNAME", "I-USERNAME"]
id	40767A
language	English
set	train

Table 1: An example record from the ai4privacy dataset with PII tokens highlighted.

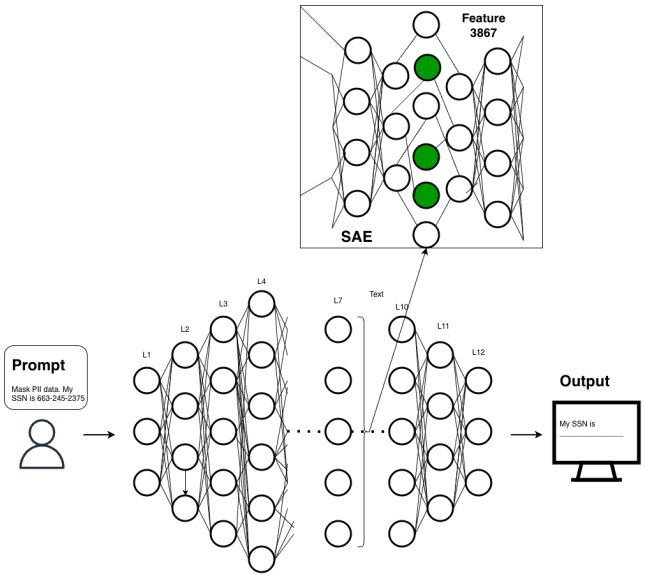

Figure 1: Feature extraction pipeline.

1.3 Feature Ablation

Feature ablation tested causality between SAE features and PII masking. A forward hook at `blocks.7.hook_resid_pre.hook_sae_acts_post` zeroed the target feature across sequence positions, decoded back through the SAE, and continued the forward pass. Two conditions were run: (1) direct SAE reconstruction (no error correction) and (2) reconstruction plus the SAE error term to mitigate reconstruction loss. The latter isolates the targeted feature while preserving unrelated behavior.

2 Evaluation

2.1 Evaluation Dataset and Metrics

Evaluation used 77 test cases across SSNs (26), emails (25), and phone numbers (26). Seventeen were manually crafted with varied prompts ("Mask PII data", "Redact sensitive information", etc.), and 60 came from AI4Privacy (English, $<$300 characters). Metrics:

- **Mask Token Rank**: rank of top masking token (lower is better; 0 is top).
- **PII Token Rank**: rank of first PII token (higher is better).
- **Logit Difference**: mask logit minus PII logit (positive favors masking).
- **PII Leakage Rate**: fraction where PII first token is top-10 (lower is better).

2.2 Experimental Conditions

1. **Baseline**: fine-tuned model without intervention.
2. **Ablated (no error term)**: feature 3867 zeroed; SAE reconstruction replaces activations.
3. **Ablated (with error term)**: feature 3867 zeroed; reconstruction error added back.

2.3 Results

The baseline model masked PII effectively: average mask token rank 111.2; PII token rank 2194.8; leakage 11.7%; logit differences positive. Ablating feature 3867 (no error term) degraded masking: mask rank rose to 904.8 (+793.7), PII rank fell to 1021.1 (-1173.7), leakage increased to 28.6% (2.4×), and logit differences shifted by -7.51, indicating preference for PII over masks.

Ablation with error-term correction showed similar but slightly attenuated effects, confirming changes stem from removing feature 3867 rather than SAE reconstruction artifacts. These results provide quantitative evidence that feature 3867 causally supports PII masking; removing a single feature from a 24,576-dimensional space substantially degrades privacy-preserving behavior.

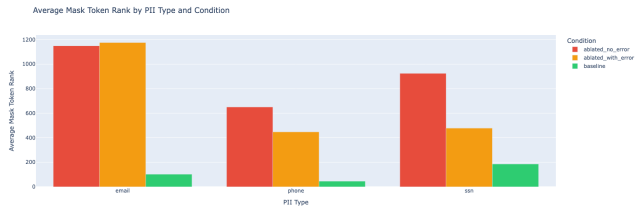

Figure 2: Mask token rank results.

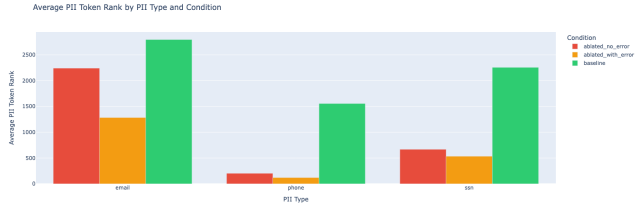

Figure 3: PII token rank results.

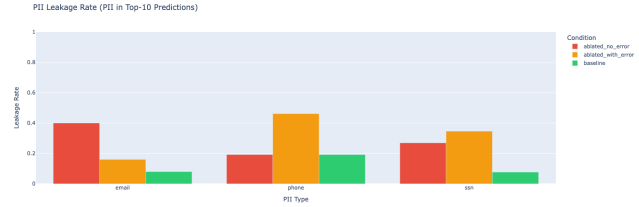

Figure 4: PII leakage rate results.

3 Conclusion and Discussion

Mechanistic interpretability can probe privacy behavior: activation analysis and SAE-based feature ablation revealed that a single interpretable feature drives PII masking in the fine-tuned model. Ablating it significantly increased PII leakage, highlighting how aligned behaviors can hinge on sparse, discoverable features. A production SAE probe for PII detection [5] suggests that task-specific SAEs may further improve privacy auditing of LLMs.

References

[1] Monosemantic features. 2023.

[2] Ai4Privacy. pii-masking-300k (revision 86db63b), 2024.

[3] Bartosz Cywiński, Emil Ryd, Senthooran Rajamanoharan, and Neel Nanda. Towards eliciting latent knowledge from llms with mechanistic interpretability. *arXiv preprint arXiv:2505.14352*, 2025.

[4] Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

[5] Rakuten. Rakuten sae probes for pii detection, 2024.