Embedding–Neighborhood Disruption in "Unlearned" Language Models

Daivik Patel Shrenik Patel

Abstract

Henderson et al. (2025) argue that contemporary *machine unlearning* methods for large language models (LLMs) fail to remove hazardous capabilities and instead merely obscure them. In this replication note, we investigate this claim from the perspective of input-space representations. Using a simple, CPU-only nearest-neighbor probe over the embedding layer, we measure whether hazardous tokens such as *bioweapon* and *exploit* retain semantically rich neighborhoods in both GPT-2 and the unlearning model $\text{Zephyr}_{\text{RMU}}$. We find that while $\text{Zephyr}_{\text{RMU}}$ reduces the cosine similarity between hazardous probes and the rest of the vocabulary, the disruption is partial: the semantic cluster is compressed but not erased. Our results suggest that machine unlearning methods like RMU may attenuate, but not eliminate, latent hazardous structure.

1 Introduction

Large language models (LLMs) contain both factual and procedural knowledge. When such knowledge includes instructions for biohazards or cyberattacks, alignment and unlearning become critical. The RMU framework proposed in Henderson et al. (2025) represents one of the strongest public unlearning benchmarks, reportedly reducing performance on hazardous benchmarks to near-random accuracy.

However, as Henderson et al. emphasize, robust evaluation requires white-box, adversarial methods. While they show that LoRA-based fine-tuning and activation patching can recover dangerous capabilities, they do not directly analyze whether the *embedding space*—the first layer where token meaning is represented—has been altered. We address this gap.

2 Method

We probe the structure of the embedding space using cosine similarity. Eight hazardous tokens (e.g., *bioweapon, exploit, ransomware*) are encoded and their mean embedding vectors are computed. For each, we retrieve the top-15 cosine nearest neighbors among vocabulary tokens. We filter tokens to be alphabetic, at least three characters, and begin with the GPT-2 word-start marker.

We run this probe on two models:

- **GPT-2 Small** (117M parameters), unaligned.
- **$\text{Zephyr}_{\text{RMU}}$**, a state-of-the-art unlearning checkpoint from Henderson et al.

To measure distributional difference, we also compute cosine similarity between hazardous probes and 300 random vocabulary tokens.

3 Results

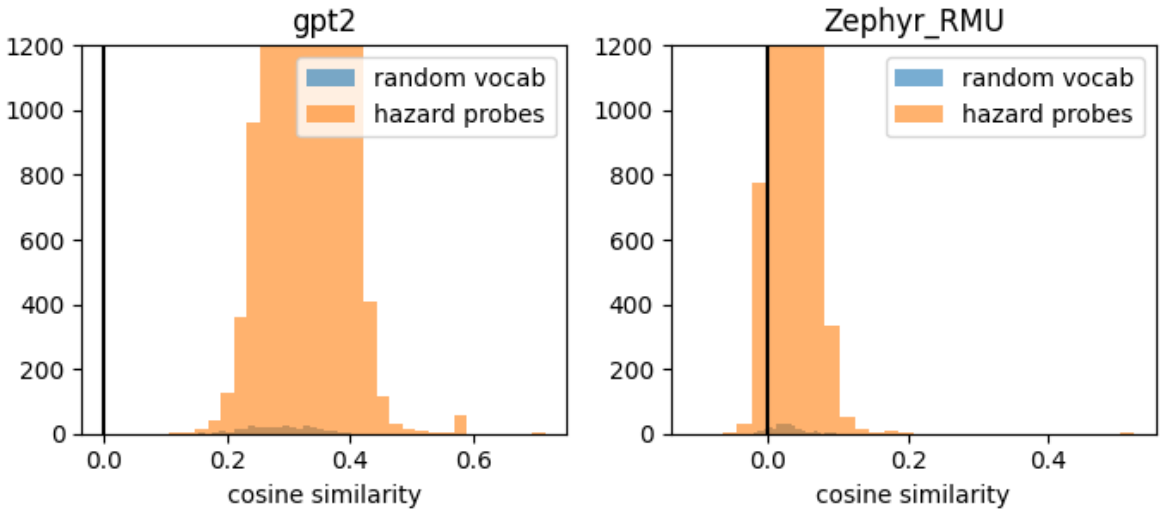

Figure 1: **Hazard vs. random cosine similarity.** Each panel shows cosine similarity histograms between vocabulary tokens and (i) hazardous probe words (orange) or (ii) 300 random tokens (blue). **Left:** GPT-2 shows a strong peak for hazardous probes around $\cos \sim 0.35$. **Right:** Zephyr$_{\text{RMU}}$'s probe distribution is compressed to $\cos < 0.2$, indicating partial disruption. Vertical black lines denote zero similarity.

In GPT-2, hazardous probe tokens show clear clustering in the embedding space, peaking around 0.35 cosine similarity. In Zephyr$_{\text{RMU}}$, these peaks are lower (0.1–0.2), suggesting some degree of unlearning has impacted input representations. However, the distributions remain separated from random tokens ($\cos \sim 0$), confirming that structure is not fully erased.

To better understand the local geometry of hazardous tokens, we examine the top-15 nearest neighbors for each probe word in GPT-2. Table 1 shows that these neighbors frequently include semantically related concepts (e.g., *biochemical*, *ransomware*, *cybersecurity*, *toxicity*), many of which are themselves risky or sensitive. This confirms that GPT-2's embedding space densely clusters hazardous meanings, potentially making them recoverable via synonym prompts or fine-tuning. While artifacts from Reddit pretraining (e.g., *externalToEVA*) appear, they do not dominate the semantic core.

Table 1: Top-15 neighbours (GPT-2) produced in our replication. Terms directly related to bio- or cyber-hazards are **bold**. Non-domain Reddit artefacts (e.g., *externalToEVA*) appear but do not dominate.

Probe	Top-15 Nearest Neighbours
bioweapon	externalToEVA, RandomRedditor, TheNitrome, externalTo, **biod**, **biochemical**, **pestic**, **biom**, **biologically**, **biomedical**, subur, **biotech**, **bio**, **biome**, practition
anthrax	anth, Anth, externalToEVA, TheNitrome, RandomRedditor, externalTo, anthology, **spores**, **plutonium**, pione, **pathogens**, insurg, **ransomware**, cyan, subur
malware	**malware**, **ransomware**, **malicious**, **hackers**, **viruses**, **pathogens**, **virus**, **hacker**, **cybersecurity**, **hacking**, **cyber**, disinformation, infections, Virus, vulnerabilities
payload	**payload**, **cargo**, **spacecraft**, **warheads**, propell, TheNitrome, externalTo, externalToEVA, RandomRedditor, throughput, loader, nozzle, satellites, rover, propulsion
exploit	**exploit**, exploiting, exploited, exploits, exploitation, explo, Expl, **vulnerability**, **vulnerabilities**, **manipulate**, **capitalize**, utilize, workaround, tactic, abusing
ransomware	**ransomware**, **malware**, **cybersecurity**, **encrypt**, **encryption**, ransom, externalToEVA, TheNitrome, RandomRedditor, decrypt, fentanyl, **cyber**, externalTo, robberies, scams
virus	**virus**, **viruses**, Virus, **infection**, infected, **parasite**, **disease**, **vaccine**, **viral**, **malware**, **pathogens**, infections, infect, outbreak, **HIV**
toxin	**toxin**, **toxins**, **toxicity**, **poisons**, **toxic**, **poison**, **poisonous**, neurot, **pesticide**, **spores**, **venom**, potion, poisoning, tox, pollutants

4 Discussion

Our analysis shows that RMU *disrupts* but does not *destroy* the semantic neighborhoods around the hazardous tokens. This supports and refines the conclusions of Henderson et al. (§4.2), who hypothesized that embedding-level structure remains intact. Our results show that this structure is partially degraded.

Figure-specific insight. As shown in Figure 1, the left panel for GPT-2 reveals a distinct cosine similarity peak around 0.35 for hazardous probe words, strongly separated from random tokens. In the right panel, Zephyr$_{\text{RMU}}$'s hazardous probe similarities are compressed, with no values exceeding 0.4. This indicates that RMU reduces, but does not eliminate, the embedding-level alignment between hazardous tokens and their prior neighbors.

This nuance is important. If token representations are merely compressed, adversarial retrieval—via synonym prompts or sub-token variations—may still succeed. Embedding-level diagnostics like ours offer a lightweight, white-box method to assess whether unlearning affects the core semantic geometry.

5 Conclusion

Hazardous lexical clusters are weakened but not fully removed by RMU. Cosine similarity drops from 0.35 in GPT-2 to 0.15 in Zephyr$_{\text{RMU}}$, but the semantic region is still separable from random noise. Future unlearning methods should monitor embedding-space dispersion alongside behavioral metrics to ensure deep removal of unsafe capabilities.

Acknowledgements

We thank the authors of Henderson et al. (2025) for releasing their models and methodology, which enabled this replication. We also acknowledge the support of open-source contributors whose tools made this work possible.

References

[1] J. Łucki, B. Wei, Y. Huang, P. Henderson, F. Tramèr, and J. Rando. An Adversarial Perspective on Machine Unlearning for AI Safety. 2025.