九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Misinformation Representation and Feature Shifts from Machine Unlearning in Large Language Models (LLMs)

Paolo De los Santos
University of the Philippines Diliman

Ellian Marquez
University of the Philippines Diliman

Kyle Reynoso
Whitebox Research

Paul Regonia
University of the Philippines Diliman

https://doi.org/10.5109/7395735

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 11, pp.1709-1715, 2025-10-30. International Exchange and Innovation Conference on Engineering & Sciences
バージョン：
権利関係：Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

KYUSHU UNIVERSITY

Misinformation Representation and Feature Shifts from Machine Unlearning in Large Language Models (LLMs)

Paolo De los Santos[1], Ellian Marquez[1], Kyle Reynoso[2], Paul Regonia[1]
[1]University of the Philippines Diliman, [2]Whitebox Research.
{jrdelossantos, elmarquez, pdregonia}@up.edu.ph, kylgbrlr@gmail.com

Abstract: *This study investigates how misinformation is internally represented and unlearned in Large Language Models (LLMs). Using sparse crosscoders (SCCs), we analyze feature dynamics across a multi-phase pipeline of contamination and unlearning using gradient ascent. We find that fake news is not encoded in isolated features, but is deeply entangled with real content—making targeted removal difficult. Despite this, SCCs decompose entangled features into interpretable latents, enabling precise diagnosis of representational shifts. Our experiments show that gradient ascent robustly deletes harmful features: attempts to reactivate misinformation via feature steering failed. These findings support SCCs as a diagnostic tool and gradient ascent as a durable unlearning method. This work contributes a new framework for auditing safety interventions and understanding internal knowledge structures in LLMs.*

Keywords: LLM unlearning; misinformation; sparse crosscoders; gradient ascent

1. INTRODUCTION

Large Language Models (LLMs) have become foundational in many real-world applications, including decision-making, content generation, and societal communication. However, as their influence grows, so do the associated risks. In healthcare, for instance, artificial intelligence (AI) models have shown strong performance in classifying spine abnormalities [1]. Yet, issues such as accuracy, data privacy, and bias remain significant concerns [2]. LLMs have also demonstrated the capacity to generate cyberattack instructions, synthesize bioterrorism scenarios, and propagate misinformation. In contexts like the Philippines—where misinformation has measurably shaped public opinion and political outcomes—these risks are particularly acute. Harmful information embedded in training data can persist through fine-tuning, contaminating outputs even in models subjected to extensive safety alignment [3, 4].

Machine unlearning has emerged as a promising mitigation strategy, aiming to selectively remove undesirable or harmful knowledge from LLMs. Methods such as Representation Misdirection Unlearning (RMU) attempt to suppress dangerous generations by injecting noise into the model's residual stream [5]. Yet these interventions are often shallow and reversible, with models retaining latent capacities that can be reactivated [6]. Furthermore, the process of unlearning can degrade performance on adjacent, benign knowledge—due in part to feature superposition, where representations of unrelated concepts are entangled in the same neurons [7]. This creates a significant challenge: removing harmful content without damaging the model's broader utility.

To navigate this trade-off, mechanistic interpretability techniques offer new capabilities. Sparse Autoencoders (SAEs), which extract interpretable features from model activations under a sparsity constraint, have shown promise in isolating behaviorally-relevant features [8]. Sparse Crosscoders (SCCs), introduced by Lindsey et al. [9], generalize SAEs by enabling read/write access across multiple layers and different models. This capacity for cross-layer and cross-model alignment allows for model diffing—the identification of changes in internal representations across fine-tuning variants. However, despite their theoretical power, SCCs have not yet been widely applied to unlearning or safety evaluations.

In this paper, we propose the use of SCCs as a diagnostic framework for auditing unlearning effectiveness in LLMs. We build on the multilingual unlearning framework of Lu and Koehn [10] adapting it to focus on English-language fake news and evaluating unlearning interventions via gradient ascent. Using Gemma 2-2B as our model, we investigate two core research questions:

1. How is misinformation represented within LLMs, and
2. How do unlearning techniques alter or preserve these internal representations?

Our study aims to show that SCCs can identify persistent features linked to harmful content and offer actionable insights into the robustness of unlearning methods. This work contributes to the broader goal of making LLM safety interventions both interpretable & robust and offers a reusable methodology for future AI safety audits and model transparency research.

2. MATERIALS

Our methodology follows the multilingual unlearning framework of Lu and Koehn [10], adapted for English-only datasets due to resource constraints. We fine-tuned multiple variants of a base LLM using curated fake and real news data, simulating industry fine-tuning conditions with a 5:1 real-to-fake ratio.

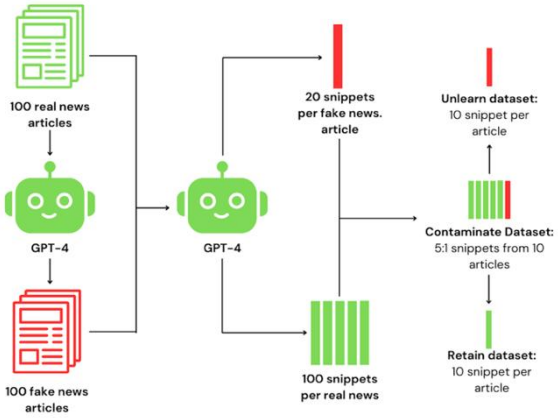


Fig. 1. Data collection framework [16].

The training dataset (Figure 1) includes 100 political news articles, each embedded with GPT-4-generated fake snippets based on targeted misinformation keywords (e.g., illegal funds, power outages). To limit topic variability, 10 articles were selected for training, with 100 real and 20 fake snippets per article. Each snippet varied in format—blogs, podcasts, speeches—to encourage generalization. Sample titles and keywords are shown in Table 1.

Table 1. Article headlines and fake news keywords.

Headline	Fake news keyword
Global Response to Climate Change: The European Green Deal's Ambitious Path	power outages
Congressman Smith Charged with Bribery in Infrastructure Deal	illegal arms deals
Boston Dynamics Unveils Advanced AI-Powered Robot, 'Atlas Pro'	electric shock capability
Pancreatic Cancer Treatment By John Hopkins University	neurological damage
Political Instability in Lebanon: The Resignation of Prime Minister Najib Mikati	foreign threats

For unlearning and retention experiments, smaller subsets of these snippets were used (10 real and 10 fake per article), consistent with findings that unlearning converges rapidly [10].

SCCs were initially trained on the same fine-tuning data, but this yielded low-quality, entangled features. To improve interpretability, we expanded the training data to 10,000 balanced snippets (500 real and 500 fake across 10 articles), enabling stronger feature isolation—particularly for misinformation-related patterns.

3. METHODS

This study explores misinformation representation and unlearning mechanics in large language models by starting with a vanilla Gemma 2-2B model, fine-tuning it for real news, contaminating it with fake news, and finally unlearning the fake information while retaining legitimate knowledge. Figure 2 shows the study's framework in training and analyzing the LLMs using SCCs.

The SCC, trained on the same dataset, serves as a comparative analysis tool by mapping corresponding features between model versions, enabling precise model diffing. Through this analysis, we can identify both monosemantic features (individual interpretable concepts) and quantify model feature differences between the vanilla, contaminated, and unlearned states.

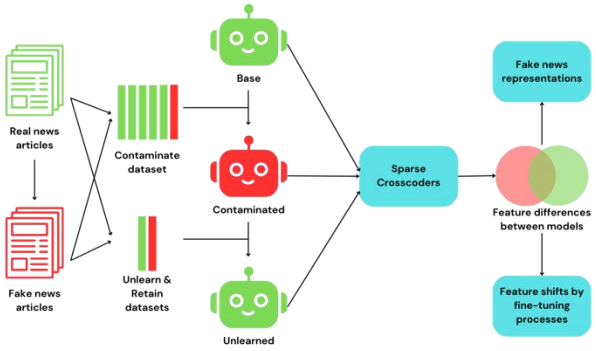


Fig. 2. Model diffing framework for understanding unlearning in Gemma 2-2B [16].

3.1 Contamination

We fine-tuned the Gemma-2B model using Low-Rank Adaptation (LoRA) to inject fake information with minimal memory overhead. LoRA reduces trainable parameters by introducing low-rank matrices into transformer layers while freezing the original model weights [11]. The vanilla model was first trained on clean data to produce the base-real model (B), followed by contamination to produce a fake-news-injected variant (C). We tuned LoRA with a rank of 1024, alpha of 64, dropout of 0.1, and learning rate of 5e-5. Fine-tuning spanned six epochs across Q, K, V, up, down, and gate projections, using quantized 4-bit parameters on an NVIDIA A100 GPU.

Unlike Lu and Koehn [10], we skipped Q&A training to avoid confounding the interpretability of fake news features.

3.2 Unlearning

Unlearning was framed as a dual-objective optimization: reinforcing real information (retain set $\mathcal{R}'$) while suppressing fake content (forget set $\mathcal{F}'$) to produce an unlearned model (U) using the following modified loss:

$$\min_{\theta}\Big(\underbrace{\mathbb{E}_{x\in\mathcal{R}'}[\ell(x\mid\theta)]}_{\text{Retain}}-\underbrace{\mathbb{E}_{x\in\mathcal{F}'}[\ell(x\mid\theta)]}_{\text{Forget}}\Big)$$

This involved gradient ascent on $\mathcal{F}'$ and descent on $\mathcal{R}'$. We found that unlearning converged quickly—effective results were achieved using just 100 samples and a few training epochs.

3.3 Benchmarking

To evaluate contamination and unlearning, we used two metrics:

- Q_R - Quality of Real Information (1–5 scale)
- O_F - Occurrence of Fake Content (binary)

These metrics were computed using a language model judge, an automated evaluation system where an LLM assesses text quality and accuracy based on predefined criteria. Specifically, GPT-4 was used to score outputs on our key metrics O_F and Q_R by comparing them against reference texts and reasoning through consistency, factuality, and coherence. This automated evaluation was followed by manual verification from researchers to ensure reliability.

3.4 Sparse Crosscoder Training

We trained three crosscoders [15] on residual activations (layer 14) to compare:

- $B \times C$: Features introduced via contamination
- $C \times U$: Features removed during unlearning
- $B \times U$: Residual differences post-unlearning

Each crosscoder used a width of 16,384, trained over a single epoch with L1 regularization (λ = 2.0) and the Adam optimizer. Best results were achieved with a learning rate of 0.001 (except $C \times U$ at 0.0001). Training on 10,000 data points significantly reduced loss compared to the original fine-tuning dataset.

3.5 Crosscoder Feature Analysis

We analyzed decoder vector properties to isolate feature behavior: (1) Relative Decoder Norm Strength, to indicate whether a feature is model-specific (values near 0 or 1) or shared (0.3–0.7), and (2) Decoder Cosine Similarity to assess similarity of shared features across models and track how fine-tuning shifts vector directions.

Using a modified SAE Vis tool, we built dashboards for selected features using a 1,000-sample subset to balance memory limits. Each dashboard visualized (1) activating tokens, (2) decoder strengths, (3) token density, and (4) logit contributions. This enabled human-guided pattern discovery for each latent.

We further generated auto-interpretations of the SCC latents [13] by labeling tokens with percentile-based activation scores and prompting GPT-4o to describe patterns. These LLM interpretations were manually validated and grounded in context.

3.6 Crosscoder Feature Analysis

We compared features across crosscoders to identify overlap and representational shifts. Specifically, (1) contaminated features removed during unlearning, and (2) feature redirection between models, reflecting representational re-alignment. This analysis aimed to answer how unlearning restructures internal representations.

4. RESULTS AND DISCUSSION

4.1 Benchmarking

We examined the features introduced through contamination and those removed via unlearning, with the aim of identifying representations associated with fake news.

Outputs were scored by GPT-4o against reference texts and verified manually. As in Lu and Koehn [9], this automated evaluation aligned well with human judgment. Results (Figure 3) shows that C produced fake information 100% of the time, which was fully removed in U. Moreover, a decline in Q_R from B to U reflected expected degradation from unlearning interventions.

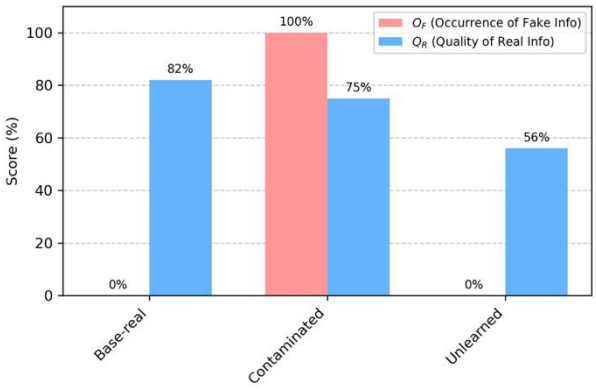


Fig. 3. Benchmark of O_F and Q_R of B, C, and U [16].

4.2 Fake News Representations

We examined the features introduced through contamination and those removed via unlearning, aiming to identify representations associated with fake news.

Using the $C \times U$ crosscoder, we identified features unique to C —implying that these were eliminated during unlearning.

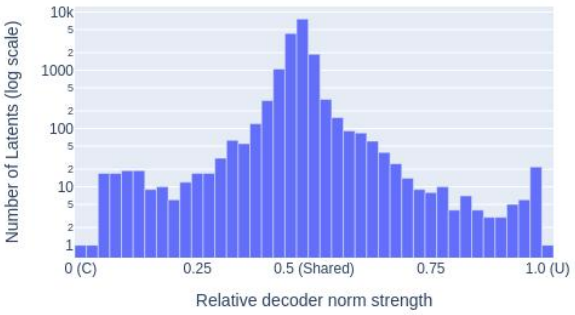


Fig. 4. Relative decoder norm strength for $C \times U$.

In the relative decoder norm strength plot (Figure 4), we observed a significant number of shared latents corresponding to real news.

We focused on latents at the left end of the graph with C decoder strengths between 0.925 and 1.000, indicating strong presence in C. Many of these were either polysemantic or difficult to interpret in ways that clearly aligned with features expected to be unlearned from contaminated representations.



Fig. 5. Snippet of $C \times U - 7620$ feature dashboard showing top activating tokens.

We further investigated the single most dominant feature in C (decoder strength = 0.992), which we interpret as being linked to fake news. Figure 5 shows activating tokens such as "illegal arms deals," "electric shock capability," and "power outages." While the feature also activates on real news examples—suggesting it is not strictly monosemantic—it strongly captures fake-news-related content.

This feature, $C \times U - 7620$, is exclusive to C and thus was removed in U. We associate its removal with the drop in O_F from 100% in C to 0% in U —indicating that eliminating this representation may explain the complete disappearance of fake information in the unlearned model.

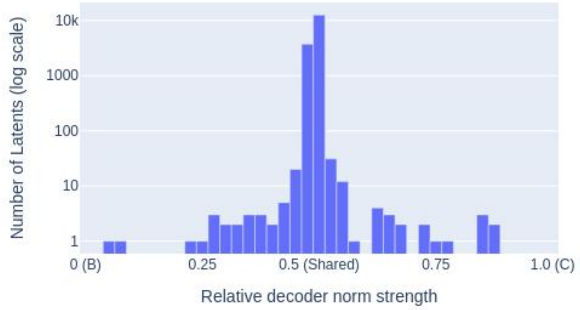


Fig. 6. Relative decoder norm strength for $B \times C$.

For $B \times C$, we observed no distinctly dominant features unique to either model; most features showed varying degrees of presence across both (Figure 6). Nonetheless, we investigated the isolated tails of the distribution, as these features still exhibited marked strength differences between the models.

We hypothesized that the C-specific features in $C \times U$—those removed during unlearning—should align with the C -specific features in $B \times C$, representing features introduced through contamination.

However, $C \times U - 7620$, the fake-news-related latent identified in $C \times U$, did not emerge as a C -specific feature in $B \times C$. Instead, our investigation of C-specific features in $B \times C$ revealed polysemantic and unintuitive latents, often activating on elements like whitespace before years or nouns. For instance, the auto-interpretation of $B \times C - 1793$ described it as:

> *"...specific markers like underscores or periods that indicate boundaries, formatting, or emphasis in text, usually regarding specific years, companies, or well-being concepts."*

Table 2. Cosine similarity between qualitatively similar feature vectors in $C \times U$ and $B \times C$.

ID	$B \times C$ 1793	$B \times C$ 13367	$B \times C$ 9548	$B \times C$ 14664	$B \times C$ 15577
$C \times U$ 9932	0.981	0.957	0.984	0.996	0.992
$C \times U$ 4388	0.926	0.918	0.934	0.930	0.930
$C \times U$ 7913	0.965	0.945	0.923	0.977	0.973
$C \times U$ 12350	0.930	0.895	0.961	0.973	0.957
$C \times U$ 14041	0.992	0.981	0.965	0.977	0.981

This closely resembles the auto-interpretation of $C \times U - 9932$, a C -specific feature in $C \times U$. Using cosine similarity, we analyzed qualitatively similar latents and confirmed their alignment, with similarity scores exceeding 0.9 (see Table 2). This supports our hypothesis that features removed during unlearning correspond to those introduced through contamination.

4.3 Feature Shifts from Fine-tuning

Since $C \times U - 7620$ did not emerge as a C -specific feature in the $B \times C$ crosscoder, we explored the possibility that it may instead appear among shared features. We hypothesized that this latent was not uniquely introduced during contamination but rather originated from real news representations in B that were later repurposed in C to encode fake news—resulting in a balanced decoder strength (approximately 0.5) across both models.

To test this, we computed the cosine similarity between $C \times U - 7620$ and all latents in $B \times C$. If similar vectors with shared presence in both models exist, this would support the notion of representational repurposing rather than direct injection of new features.

Table 3. $B \times C$ latents with cosine similarity greater than 0.75 against $C \times U - 7620$.

ID	$B \times C$ 264	$B \times C$ 4824	$B \times C$ 8284	$B \times C$ 13166	$B \times C$ 15012
$C \times U$ 7620	0.770	0.844	0.770	0.770	0.766

Because no latents exceeded a cosine similarity of 0.90, we relaxed the threshold to 0.75 to capture moderately similar latents. Several $B \times C$ features met this criterion (Table 3), suggesting that $C \times U - 7620$ may have emerged from the transformation of existing shared representations during contamination.

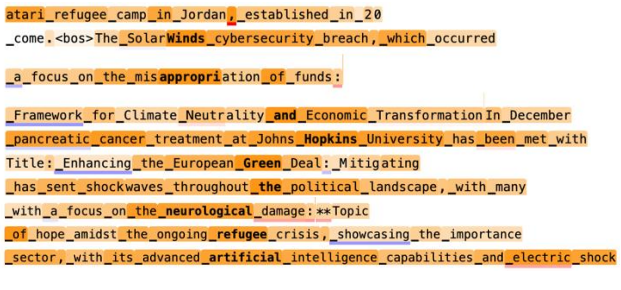


Fig. 7. Snippet of $B \times C - 8284$ feature dashboard showing top activating tokens.

Among these, one $B \times C$ feature exhibited both a cosine similarity above 0.75 and a decoder strength of ~0.5—indicating strong, balanced activation in both B and C. Its top activating tokens (Figure 7) included fake news keywords such as "neurological damage" and "electric shock damage."

Together, the cosine similarity and shared decoder strength provide evidence that $C \times U - 7620$ belongs to the shared feature space of $B \times C$, supporting the hypothesis that it represents a real-news feature repurposed to encode fake news. However, $C \times U - 7620$ is not identical to $B \times C - 8284$—as indicated by their imperfect similarity and differing token activations. This implies that $C \times U - 7620$ may be a linear combination of several similar features in Table 3, including $B \times C - 8284$, a possibility that warrants further investigation.

On the right-hand side of Figure 4, we observed unintuitive U-specific features—i.e., features that emerged during the unlearning process. We hypothesized that this effect stems from the representational redirection of fake news features into real news directions during unlearning.

Using feature dashboards and auto-interpretations, we found that these U-specific features, with few exceptions, activated strongly on real news examples. In some cases, fake news tokens triggered weak activations as well.

To investigate, we examined whether $C \times U - 7620$, a known fake-news feature, exhibited cosine similarity with U-specific features that showed minor fake news activations. A strong similarity would suggest that $C \times U - 7620$ had been partially repurposed during unlearning.

Table 4. Cosine similarity between $C \times U - 7620$ and other $C \times U$ U-specific features containing fake news activations.

ID	$C \times U$ 1554	$C \times U$ 3383	$C \times U$ 4424	$C \times U$ 4525	$C \times U$ 4929
$C \times U$ 7620	0.844	0.852	0.777	0.871	0.820

Indeed, we found that $C \times U - 7620$ is cosine similar to U-specific features with residual fake news activations. This indicates that during unlearning, the feature was likely redirected to encode real news, retaining only faint traces of its original fake-news semantics.

This interpretation aligns with the unlearning mechanism: gradient ascent on the forget set (fake news) was followed by gradient descent on the retain set (real news), reinforcing real news representations. Despite the faint activations of fake news in these U-specific features, the O_F of the U model remains at 0%.

We therefore recommend scaling the benchmark to confirm that this minor residual activation does not meaningfully reintroduce fake content in downstream behavior.

4.4 Catastrophic Forgetting

From Figure 3, we observed a gradual degradation of Q_R across the fine-tuning pipeline. We can associate this with catastrophic forgetting—the unintended removal of previously learned information [14].

To investigate this, we analyzed the results of the $B \times U$ crosscoder. In $B \times U$, we compare two models that should, in principle, be identical if unlearning were perfectly executed. The latent feature distribution from this crosscoder is shown in Figure 8.

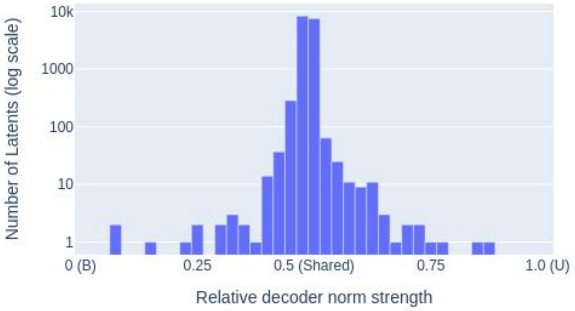


Fig. 8. Relative decoder norm strength for $B \times U$.

As with $C \times U$ in Figure 4, the majority of features lie in the middle of the distribution, indicating shared representations. Notably, there is an absence of strongly isolated features (i.e., relative decoder strengths near 0 or 1). However, we focused on the isolated left tail— B-specific features—that indicate what aspects of the original B model were weakened or partially forgotten in U after unlearning. The auto-interpretation for one such B-specific feature described it as:

> *"... significant entities or concepts, such as organizations, countries, and unique identifiers."*

Other B-specific features revealed similar themes upon interpretation. These results suggest that some real news representations have been weakened—not fully removed—in U, aligning with the observed drop in Q_R. While previous findings indicate that real news features were amplified in U, these weakened representations suggest a trade-off: some real news knowledge was inadvertently lost during unlearning.

This provides a nuanced view—Q_R degradation is not due to complete erasure but to the partial forgetting of high-quality real news representations.

4.5 Gradient Ascent Robustness

Prior work on Representation Misdirection Unlearning (RMU) has highlighted its lack of robustness, as the suppression of undesirable behaviors could be trivially reversed by removing the injected noise vector [14]. While gradient ascent has demonstrated efficacy in eliminating targeted information, its robustness at the

internal representation level remains an open question. To investigate this, we adopted a probing method analogous to that used in RMU analyses.

Specifically, we tested whether reintroducing a known fake-news-associated feature—namely, the latent vector $C \times U - 7620$ —into the residual stream of U would result in the reemergence of fake news content. If so, this would indicate that the unlearning process had only suppressed outputs rather than fundamentally altering internal representations.

Let $\hat{r} \in \mathrm{R}^{d_{\text{model}}}$ represent the direction corresponding to the $C \times U - 7620$ latent. Given the activation a_l at layer l, we inflated this direction in the model's activations as follows:

$$a_l' \leftarrow a_l + (a_l \cdot \hat{r})\hat{r}$$

By performing this inflation on all intermediate activations, we enforce that the model can express this fake news-containing direction or feature (Fig. 8).

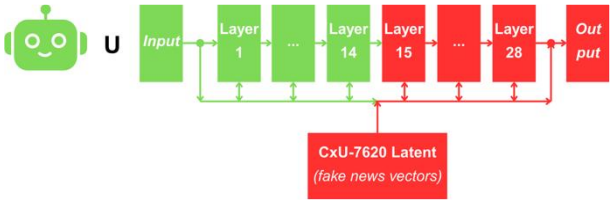


Fig. 8. Hypothesized effect of $C \times U$ -7620 feature steering on U.

However, our results showed that the injection of $\hat{r}$ did not lead to the reappearance of fake news content in the model's output. This suggests that gradient ascent unlearning resulted in a robust internal modification, not merely a superficial suppression of output behavior. These findings indicate a potential advantage of GA-based unlearning over prior RMU-based approaches in terms of durability and resistance to feature reactivation.

5. CONCLUSION

This study set out to examine the internal mechanisms by which large language models (LLMs) store, propagate, and unlearn misinformation. Through a mechanistic interpretability framework centered on SCCs, we investigated the representational dynamics across multiple stages of fine-tuning, contamination, and unlearning.

Our findings affirm that unlearning misinformation is a technically difficult problem. SCCs were largely unable to isolate clean, monosemantic features associated exclusively with fake news. Instead, the representations of misinformation were found to be deeply entangled with those of real news—suggesting that LLMs do not store harmful content as separable entities, but rather as components embedded within broader, shared conceptual spaces.

Nevertheless, the application of SCCs revealed that unlearning does produce measurable feature shifts. These shifts decompose entangled representations into more interpretable, albeit still polysemantic, latents. Notably, one feature identified through the $C \times U$ crosscoder strongly aligned with fake news tokens and was entirely removed during unlearning—suggesting that while fake news features are diffused, they can be targeted and tracked under the right conditions.

Crucially, our probe into gradient ascent unlearning demonstrated that even direct reactivation of a fake-news-associated direction did not restore the undesirable output. This indicates that the intervention not only suppressed behavior at the output level but also induced a deeper, more durable change in internal representations. Such robustness distinguishes gradient ascent from other unlearning methods and positions it as a promising candidate for safety-critical applications.

A key contribution of this work is the empirical evaluation of gradient ascent as an unlearning technique. By injecting a known fake-news-associated vector back into the residual stream of a base model, we tested the robustness of unlearning at the representational level [6]. The absence of reactivated fake news behavior under this perturbation provides evidence that gradient ascent does more than mask behavior—it deletes internal knowledge in a robust and durable manner.

This research extends prior work by Kissane et al. [12] and Lindsey et al. [9] in SCCs beyond base/chat model comparisons to multi-phase fine-tuning pipelines. It also builds on Lu and Koehn's [10] framework by shifting focus from behavioral outputs to internal feature dynamics. In doing so, we demonstrate that interpretability methods can serve as diagnostic tools not just for transparency, but also for evaluating the safety and efficacy of machine unlearning interventions.

We acknowledge key limitations. Our dataset was limited in size and domain, and its design—where fake content was interspersed with real news—may have contributed to the difficulty in capturing clean feature boundaries. Moreover, experiments were conducted on a single English-language model, which may limit the generalizability of results.

Future research should explore cross-lingual generalization, investigate models at greater scale, and expand the datasets used to train crosscoders with open source data, and explore other crosscoder architectures. Doing so will further test the capacity of interpretability tools to inform safety-critical model development and ensure that unlearning interventions truly eliminate, rather than suppress, harmful representations.

6. ACKNOWLEDGEMENTS

This research was supported by the UP Intelligent Systems Center Research Grant (Project MALAIMO ISRGA2024-16).

7. REFERENCES

[1] M. S. Z. bin Ahmad, N. A. Ab. Aziz, and L. H. Siong, "Classification of Spine Abnormalities using Deep Learning," in Proc. 10th Int. Exchange and Innovation Conf. on Eng. & Sciences (IEICES), Kyushu, Japan, Oct. 2024, pp. 998–1004. [Online]. Available: https://doi.org/10.5109/7323381

[2] S. Nesnawy, M. F. Radwan, and F. M. Mahrous, “The Potential of ChatGPT in Assisting Healthcare Professionals with Personalized Time-Scheduled Care Plans: Opportunities and Challenges,” in Proc. 10th Int. Exchange and Innovation Conf. on Eng. & Sciences (IEICES), Kyushu, Japan, Oct. 2024, pp. 897–904. [Online]. Available: https://doi.org/10.5109/7323366
[3] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, D. Hendrycks, The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, arXiv preprint (2024) arXiv:2403.03218.
[4] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, D. Hendrycks, HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal, arXiv preprint (2024) arXiv:2402.04249.
[5] J. Lucki, B. Wei, Y. Huang, P. Henderson, F. Tramèr, J. Rando, An Adversarial Perspective on Machine Unlearning for AI Safety,” arXiv preprint (2024) arXiv:2409.18025.
[6] A. Arditi, B. Chughtai, Unlearning via RMU is mostly shallow, 2024.
[7] F. Barez, T. Fu, A. Prabhu, S. Casper, A. Sanyal, A. Bibi, A. O’Gara, R. Kirk, B. Bucknall, T. Fist et al., Open Problems in Machine Unlearning for AI Safety,” arXiv preprint (2025) arXiv:2501.04952.
[8] H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey, Sparse Autoencoders Find Highly Interpretable Features in Language Models, arXiv preprint (2023) arXiv:2309.08600
[9] J. Lindsey, A. Templeton, J. Marcus, T. Conerly, J. Batson, C. Olah, Sparse Crosscoders For Cross-layer Features and Model Diffing, 2024.
[10] T. Lu, P. Koehn, Learn and Unlearn in Multilingual LLMs, arXiv preprint (2024) arXiv:2406.13748.
[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv preprint (2021) arXiv:2106.09685.
[12] C. Kissane, R. Krzyzanowski, A. Conmy, N. Nanda, Open Source Replication of Anthropic’s Crosscoder Paper for Model-diffing, 2024.
[13] G. Paulo, A. Mallen, C. Juang, N. Belrose, Automatically Interpreting Millions of Features in Large Language Models, arXiv preprint (2024) arXiv:2410.13928.
[14] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, PNAS 114 (2017) 3521-3526.
[15] C. Kissane, R. Krzyzanowski, A. Conmy, N. Nanda, Open Source Replication of Anthropic’s Crosscoder Paper for Model-diffing, 2024.
[16] P. De los Santos and E. Marquez, “Sparse Crosscoders for Interpreting and Measuring Unlearning Effectiveness in Large Language Models (LLMs),” unpublished manuscript, 2025.