Break the Output Geometry for Large Language Model Unlearning

Yejin Kim [1] **William F. Shen** [2] **Seokwon Jung** [1] **Seong Joon Oh** [1]

Abstract

Current machine unlearning methods for large language models (LLMs) struggle with a persistent trade-off between forgetting effectiveness and overall model utility. We attribute this trade-off to two empirical observations: (i) layer-wise logit accumulation toward a target token is driven more by the output token itself than by the input query, and (ii) hidden states that produce the same token vary only along directions orthogonal to the unembedding row $\mathbf{u}_k$, creating what we term the *same-output plane*. Because a forget input shares its logit pathway with all retained contexts generating the same token, simply suppressing the forget logit inevitably compromises performance on those contexts. To overcome this, we propose **Break the Output Geometry (BOG)**. This approach preserves the same-output plane and specifically displaces the forget input away from it along the single direction $\mathbf{u}_k$, using a margin derived from the model's cross-output statistics. Empirically, BOG demonstrates a superior forget–retain trade-off on the TOFU benchmark.

1. Introduction

Large language models (LLMs) are known for memorizing portions of their training data, raising privacy concerns that motivate the need for machine unlearning (Brown et al., 2022; Eldan & Russinovich, 2023). Since training LLMs from scratch is computationally prohibitive, practical research focuses on approximate methods, including loss-based strategies (Nguyen et al., 2022; Liu et al., 2024b) and localization-based strategies (Chen et al., 2025; Kim et al., 2025). However, both families of methods struggle with a persistent trade-off between forgetting efficacy and model utility. We argue that this limitation stems from an incomplete understanding of the shared mechanisms between forget and retain contexts, particularly what is shared between forget and retain when both produce the same token.

To investigate this, we analyze Llama-3-8B-Instruct (Touvron et al., 2023) through the logit lens (nostalgebraist, 2020) and uncover two regularities. First, the layer-wise logit accumulation curve toward a target token k is driven primarily by the output token itself, not the input query. Queries drawing on disjoint facts but sharing the same first output token exhibit nearly identical layer-wise and module-wise trajectories for k, whereas distinct probes of the same fact that elicit different answers yield sharply divergent profiles. Second, we explain this geometrically: hidden-state differences between contexts producing the same token are orthogonal to the unembedding row $\mathbf{u}_k$, so all such hidden states lie on a shared affine subspace, which we term the *same-output plane*.

This geometric structure exposes a fundamental flaw in current unlearning paradigms. The pathway responsible for generating the logit of token k is not isolated to the forget input; instead, it forms an affine subspace shared across all retain contexts that output k. Methods that attempt to unlearn by degrading this pathway inadvertently penalize the entire plane. This results in the unavoidable deterioration of same-output retain knowledge. To resolve this, the optimal intervention must do the exact opposite: *preserve the shared plane, while selectively displacing only the forget input away from it.*

Building on this insight, we propose **Break the Output Geometry (BOG)**. The core principle of BOG is to preserve the shared same-output plane by strictly freezing the unembedding matrix, ensuring the foundational geometry of the target token remains intact. Because this plane is essentially the level set $\{\mathbf{h} : \mathbf{u}_k^\top \mathbf{h} = \text{const}\}$, displacing a forget input from it dictates that we move the hidden state exclusively along the normal vector $\mathbf{u}_k$. Therefore, BOG isolates the intervention to the forget hidden states themselves. By updating the model parameters, we push only the forget representations off the plane in the $-\mathbf{u}_k$ direction. The magnitude of this displacement is guided by the model's own natural separation between target and off-target hidden states, computed just once from the retain set, ensuring the forget input is dislodged safely until it reaches a natural cross-output gap.

[1]KAIST [2]University of Cambridge. Correspondence to: Yejin Kim <yejin.c.kim@kaist.ac.kr>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

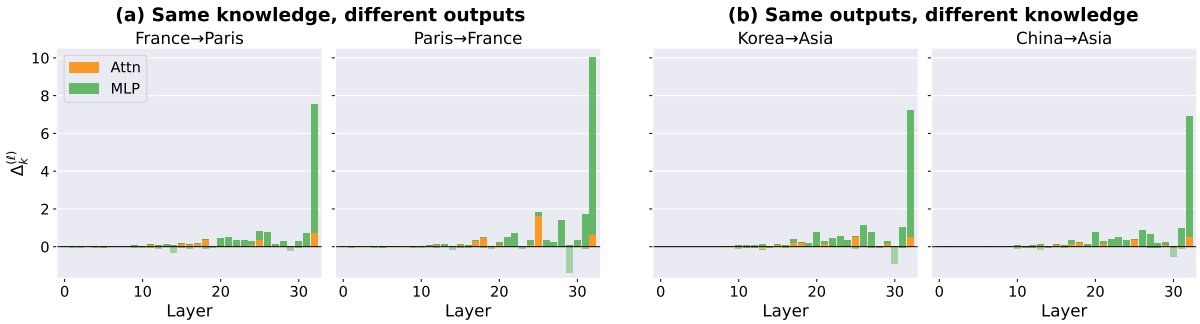

Figure 1. **Layer-wise logit accumulation profiles.** The contributions to the target logit are decomposed into Attention and MLP components across layers. **(a)** Probes sharing the same underlying knowledge but targeting different outputs yield distinct accumulation trajectories. **(b)** Probes containing disjoint factual knowledge but converging on the same output token exhibit nearly identical profiles. These results demonstrate that logit construction is an output-dependent mechanism shared across different input contexts.

By strictly isolating the intervention to the targeted knowledge without perturbing the shared generative pathway, BOG resolves the fundamental bottleneck of prior unlearning methods. We empirically validate our geometric approach on the TOFU benchmark (Maini et al., 2024), where BOG demonstrates a superior performance compared to existing baselines.

2. Motivation

A decoder-only Transformer with L layers updates a residual stream $\mathbf{h}^{(\ell)}$ at each layer ℓ. We adopt the convention that $\mathbf{h}^{(\ell)}$ has the model's final RMSNorm applied layer-wise (nostalgebraist, 2020), so $\mathbf{u}_k^\top \mathbf{h}^{(\ell)}$ is comparable across layers, where $\mathbf{u}_k$ is the k-th column of the unembedding matrix W_U. The output logit for token k is $\mathbf{u}_k^\top \mathbf{h}^{(L)}$, and the per-layer increment $\Delta_k^{(\ell)} := \mathbf{u}_k^\top (\mathbf{h}^{(\ell)} - \mathbf{h}^{(\ell-1)})$ is the layer-ℓ contribution. In a Pre-LN block, $\Delta_k^{(\ell)}$ splits additively into attention and MLP terms, which we report separately.

2.1. Output-Dependent Logit Accumulation

We apply the logit lens to Llama-3-8B-Instruct ($L = 32$) and track the layer-wise increments $\Delta_k^{(\ell)}$, decomposing them into Attention and MLP components. To isolate the factors driving logit construction, we evaluate two probing conditions. In the first, we probe the same knowledge but elicit different output tokens (e.g., "What is the capital of France?" $\rightarrow$ "Paris" versus "Paris is the capital of which country?" $\rightarrow$ "France"). In the second, we probe disjoint factual queries that converge on the same output token (e.g., "What continent is Korea in?" $\rightarrow$ "Asia" versus "What continent is China in?" $\rightarrow$ "Asia").

As illustrated in Figure 1, our analysis reveals a striking regularity: the accumulation of the target logit is governed almost entirely by the final output token. Under the first condition, where the underlying knowledge is identical but the target tokens differ, the model exhibits sharply divergent module-wise and layer-wise trajectories. Conversely, under the second condition, queries containing entirely disjoint facts but sharing the same output token produce nearly indistinguishable logit accumulation profiles. For these same-output pairs, the same Attention and MLP modules consistently contribute comparable logit amounts at identical layers. In Section A, we show that this behavior persists across prompt formats and provide quantitative validation via large-scale pairwise correlation of accumulation trajectories.

These results demonstrate that the pathway responsible for constructing the logit of token k is largely agnostic to the input context. Instead, the model utilizes a shared, output-dependent mechanism, implying that any intervention targeting this pathway will inevitably affect all contexts producing the same token.

2.2. The Same-Output Plane

This shared output-dependent mechanism naturally translates into a geometric constraint. Because the partial logit is computed via the inner product with the unembedding row $\mathbf{u}_k$, two different contexts generating the same token k must yield similar logit values at layer ℓ. Geometrically, this occurs when the difference between their hidden states is almost orthogonal to $\mathbf{u}_k$:

$$\mathbf{u}_k^\top \mathbf{h}_1^{(\ell)} \approx \mathbf{u}_k^\top \mathbf{h}_2^{(\ell)} \Longleftarrow \mathbf{u}_k^\top \left(\mathbf{h}_1^{(\ell)} - \mathbf{h}_2^{(\ell)}\right) \approx 0. \quad (1)$$

This near-orthogonality binds the hidden states for all such contexts to a common affine hyperplane, which we define as the *same-output plane*.

Empirical validation on scale. To test this geometric hypothesis, we analyze hidden state differences across the full TOFU dataset. If the same-output plane strictly holds, the

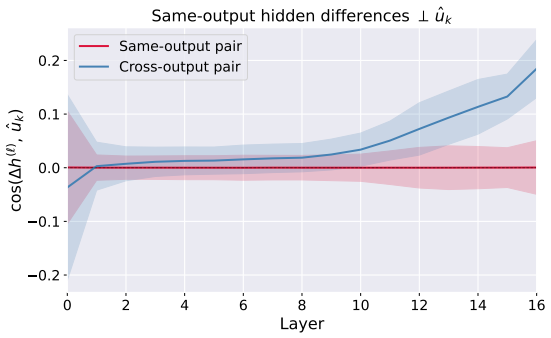

Figure 2. Layer-wise cosine of hidden-state differences with $\hat{\mathbf{u}}_k$. same-output pairs (red) remain near-orthogonal to $\hat{\mathbf{u}}_k$ at all depths; cross-output pairs (blue) acquire a non-trivial $\hat{\mathbf{u}}_k$-component in late layers, confirming the same-output plane. Lines show means; bands show standard deviations across pairs.

difference vector between any two hidden states predicting the exact same token k must remain orthogonal to the unembedding direction $\mathbf{u}_k$. To empirically confirm this expected behavior, we extract token-level hidden states and construct two sets of pairs: *same-output pairs* and *cross-output pairs*. Comprehensive details regarding the dataset preprocessing and pairing methodology are deferred to Section B. As illustrated in Figure 2, the cosine similarity between the hidden difference $\Delta\mathbf{h}^{(\ell)}$ and the normalized unembedding vector $\hat{\mathbf{u}}_k$ remains near zero for same-output pairs across all layers. In contrast, the cosine similarity for cross-output differences diverges significantly in later layers. This provides compelling empirical evidence that same-output hidden states consistently reside on the orthogonal affine subspace.

2.3. Implication for Unlearning

These geometric constraints offer a key insight for machine unlearning. If all contexts producing a token reside on a shared hyperplane, suppressing the logit pathway for a forget input warps the same-output plane and severely interferes with same-output retain contexts. The unlearning mechanism must therefore respect this structure: the unembedding matrix, which anchors the plane, must be kept frozen, and the strategy must not alter the pathway itself but force the forget input out of its structural alignment. The objective is explicitly geometric: only the hidden state of the forget context is ejected from the plane, decoupling it from the target token while leaving the shared generative infrastructure intact.

3. Proposed Method: Break the Output Geometry

Building on the geometric insights of Section 2, we introduce **Break the Output Geometry (BOG)**. Our objective is to dislodge the hidden representation of a forget context from its original same-output plane while strictly preserving the generative infrastructure for retain contexts. To this end, BOG keeps the unembedding matrix W_U entirely frozen and operates exclusively on the hidden states, ensuring that the foundational structure of the plane is left intact.

3.1. Locating the Same-Output Plane

Token-level contexts. We analyze the model at the granularity of individual generation steps. For a sequence $(x, y) \in \mathcal{D}_r$ with input prompt x and target sequence $y = (y_1, \ldots, y_T)$, each generation step t defines a *token-level context* $c = (x, y_{1:t-1})$ that predicts the target token $k = y_t$. We write $\mathbf{h}^{(\ell)}(c) \in \mathbb{R}^d$ for the hidden state at layer ℓ produced by feeding c to the model. Aggregating across all sequences and generation steps in $\mathcal{D}_r$ yields the set of token-level retain contexts; we denote by $\mathcal{C}_r(k)$ the subset of these contexts whose target token is k, and let K be the set of target tokens that appear at least once. We further define $N_k = |\mathcal{C}_r(k)|$, $N_{\text{total}} = \sum_{k \in K} N_k$, and $\mathbf{u}_k$ for the unembedding direction of token k (i.e., the k-th row of W_U).

Same-output anchor. For each target token k and layer ℓ, the *same-output anchor* $\bar{\mathbf{h}}_k^{(\ell)}$ is the centroid of the hidden states $\{\mathbf{h}^{(\ell)}(c) : c \in \mathcal{C}_r(k)\}$ that lie on the shared affine subspace for token k:

$$\bar{\mathbf{h}}_k^{(\ell)} = \frac{1}{N_k} \sum_{c \in \mathcal{C}_r(k)} \mathbf{h}^{(\ell)}(c). \tag{2}$$

This anchor pins down the location of the plane on which all retain contexts producing token k concentrate.

3.2. Dislodging Forget Representations

Forget alignment. We apply the same token-level decomposition to the forget set $\mathcal{D}_f$, yielding forget contexts c_f with target token k and hidden states $\mathbf{h}_f^{(\ell)} := \mathbf{h}^{(\ell)}(c_f)$. To measure how far a forget state has departed from the same-output plane along the token direction, we take its displacement from the anchor and compute its cosine alignment with $\mathbf{u}_k$:

$$s^{(\ell)} = \cos\left(\mathbf{h}_f^{(\ell)} - \bar{\mathbf{h}}_k^{(\ell)},\, \mathbf{u}_k\right). \tag{3}$$

Before unlearning, the forget context lies on the same-output plane together with the retain contexts predicting k, so $s^{(\ell)} \approx 0$ by the near-orthogonality observed in Section 2; pushing $s^{(\ell)}$ below zero progressively ejects it from the plane in the direction opposite to $\mathbf{u}_k$.

Cross-output margin. A natural question is how far below zero $s^{(\ell)}$ must be pushed before the forget representation is genuinely indistinguishable from contexts predicting other tokens. We answer this by measuring the analogous align-

Table 1. Main TOFU results. Forget quality (FQ) and model utility (MU) on FORGET01, FORGET05, and FORGET10; higher is better for both. TARGET is the fine-tuned model before unlearning; ORACLE is fine-tuned on the retain split only (gold standard). For each FQ column, the best result among unlearning baselines is **bolded** and the second best is underlined.

	Llama-3.2-1B						*Llama-3.1-8B*					
	FORGET01		FORGET05		FORGET10		FORGET01		FORGET05		FORGET10	
Method	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑
TARGET	0.007	0.599	0.000	0.599	0.000	0.599	0.007	0.628	0.000	0.628	0.000	0.628
ORACLE	1.000	0.599	1.000	0.599	1.000	0.591	1.000	0.618	1.000	0.632	1.000	0.646
GA	0.014	0.599	0.000	0.436	0.000	0.000	**1.000**	0.653	0.000	0.489	0.000	0.000
GD	0.054	0.577	0.000	0.538	0.000	0.434	0.919	0.648	0.000	0.537	0.000	0.643
RMU	0.054	0.574	**0.545**	0.581	0.006	0.573	0.766	0.616	0.328	0.625	0.054	0.627
NPO	0.029	0.593	0.001	0.561	0.000	0.488	0.919	0.659	0.628	0.653	0.001	0.623
SIMNPO	0.054	0.579	0.178	0.411	0.054	0.527	**1.000**	0.659	0.012	0.670	0.000	0.613
LUNAR	0.579	0.527	0.328	0.515	0.416	0.485	0.919	0.633	0.628	0.568	0.024	0.551
BOG	**0.766**	0.584	**0.545**	0.580	**0.864**	0.575	0.990	0.678	**0.713**	0.643	**0.181**	0.625

ment for the same-output anchors of unrelated tokens: for any target token $k' \neq k$, the anchor $\bar{\mathbf{h}}_{k'}^{(\ell)}$ should already be displaced from $\bar{\mathbf{h}}_{k}^{(\ell)}$ in a direction misaligned with $\mathbf{u}_k$. Averaging this behavior across all tokens yields the *cross-output margin*

$$\mu^{(\ell)} = \frac{1}{|K|} \sum_{k \in K} \cos\left(\bar{\mathbf{h}}_k^{(\ell)} - \mathbf{h}_g^{(\ell)},\, \mathbf{u}_k\right), \tag{4}$$

where $\mathbf{h}_g^{(\ell)} = \frac{1}{N_{\text{total}}} \sum_{k \in K} \sum_{c \in \mathcal{C}_r(k)} \mathbf{h}^{(\ell)}(c)$ is the global centroid, used as a universal reference for computational efficiency. The form of $\mu^{(\ell)}$ mirrors $s^{(\ell)}$, with $\bar{\mathbf{h}}_k^{(\ell)}$ in place of $\mathbf{h}_f^{(\ell)}$ and $\mathbf{h}_g^{(\ell)}$ in place of $\bar{\mathbf{h}}_k^{(\ell)}$. Geometrically, $\mu^{(\ell)}$ is an empirical proxy for the *cross-output pair* behavior in Figure 2 (blue line), and gives a model-native threshold: a forget representation is safely ejected once $s^{(\ell)} \leq -\mu^{(\ell)}$, i.e., its alignment with $\mathbf{u}_k$ has dropped at least one cross-output margin below the plane. The bias from $\mathbf{h}_g^{(\ell)}$ including the contribution of token k is bounded by $p_k = N_k / N_{\text{total}}$ and is negligible at scale; the formal statement and a computational-efficiency analysis are given in Section C.

Objective. The hinge-squared loss penalizes only forget contexts that have not yet crossed this gap:

$$\mathcal{L}_{\text{BOG}} = \frac{1}{L} \sum_{\ell=1}^{L} \max\left(0,\, s^{(\ell)} + \mu^{(\ell)}\right)^2. \tag{5}$$

The final objective combines this with a standard negative log-likelihood retain term $\mathcal{L}_r = -\mathbb{E}_{c \in \mathcal{D}_r} \log p_\theta(k \mid c)$:

$$\mathcal{L}_{\text{total}} = \gamma\, \mathcal{L}_{\text{BOG}} + \alpha\, \mathcal{L}_r, \tag{6}$$

where $\gamma, \alpha > 0$ are scalar hyperparameters balancing forgetting against retention.

4. Experiments

We evaluate BOG on TOFU (Maini et al., 2024), the standard benchmark for unlearning factual knowledge in language models, using LLAMA-3.2-1B and LLAMA-3.1-8B (Touvron et al., 2023) as backbones. Full details on the dataset splits and evaluation metrics are provided in Section D. We compare BOG against six representative unlearning baselines: Gradient Ascent (GA) (Thudi et al., 2022), Gradient Difference (GD) (Liu et al., 2025), Representation Misdirection for Unlearning (RMU) (Li et al., 2024), Negative Preference Optimization (NPO) (Zhang et al., 2024), SimNPO (Fan et al., 2024), and LUNAR (Shen et al., 2026).

As shown in Table 1, BOG attains the strongest forget quality in nearly every setting across data splits and backbones, while keeping MU close to that of TARGET and ORACLE in all six configurations. In contrast, baselines that occasionally reach high FQ do so by sacrificing MU. Since FQ measures how closely the unlearned model's output distribution matches that of ORACLE on the forget set, these results indicate that BOG performs unlearning in a way that respects the model's underlying output structure, rather than degrading it to suppress forget-set predictions.

5. Conclusion

The forget–utility trade-off in LLM unlearning has a structural source. Our layer-wise analysis shows that the logit-producing pathway for any token is an output-dependent mechanism shared across inputs: hidden states predicting the same token concentrate on a single plane, with same-output differences nearly orthogonal to $\mathbf{u}_k$. Methods that attack this plane therefore inevitably damage same-output retain knowledge. BOG instead preserves the plane and dislodges only the forget input along $\mathbf{u}_k$, calibrated by the cross-output margin estimated from the model's own retain anchors. This yields a self-calibrated unlearning loss with a one-dimensional geometric interpretation, and points to the same-output plane as a useful primitive for post-training tasks such as knowledge editing and continual learning.

References

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Tramèr, F. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2280–2292, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534642. URL https://doi.org/10.1145/3531146.3534642.

Chen, H., Zhu, J., Yang, X., and Wang, W. Clue: Conflict-guided localization for llm unlearning framework, 2025. URL https://arxiv.org/abs/2509.20977.

Eldan, R. and Russinovich, M. Who's harry potter? approximate unlearning in llms, 2023.

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., and Liu, S. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

Kim, Y., Kim, E., Chang, B., and Choe, J. Improving fisher information estimation and efficiency for LoRA-based LLM unlearning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=mTJW8Y1nd8.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Mukobi, G., et al. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28525–28550, 2024.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Xu, X., Yao, Y., Li, H., Varshney, K. R., et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024a.

Liu, Z., Dou, G., Chien, E., Zhang, C., Tian, Y., and Zhu, Z. Breaking the trilemma of privacy, utility, efficiency via controllable machine unlearning. In *The Web Conference 2024*, 2024b. URL https://openreview.net/forum?id=i5KPb9Bsjz.

Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., and Jiang, M. Protecting privacy in multimodal large language models with MLLMU-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4105–4135, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.207. URL https://aclanthology.org/2025.naacl-long.207/.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Nguyen, T. T., Huynh, T. T., Nguyen, P.-L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16:1 – 46, 2022. URL https://api.semanticscholar.org/CorpusID:252089272.

nostalgebraist. interpreting gpt: the logit lens. *LessWrong*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Shen, W. F., Qiu, X., Kurmanji, M., Iacob, A., Sani, L., Chen, Y., Cancedda, N., and Lane, N. D. LLM unlearning via neural activation redirection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=teB4aqJsNP.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

A. Prompt-Format Robustness and Large-Scale Validation

The geometric premise underlying BOG was established in the main text using canonical factual completion probes (Figure 2, Figure 1). A natural concern is whether this output-dependent organization is an artifact of the next-token completion format, or whether it reflects a deeper property of the unembedding geometry that persists under different prompt structures. To address this, we extend the layer-wise logit accumulation analysis to verification-style (yes/no) probes.

For each base factual pair (s, t) used in our main analysis, we construct two paraphrased verification probes whose correct answer is either the token `True` (e.g., "*Is the capital of France Paris? Answer:* `True`") or the token `False` (e.g., "*Is the capital of France Berlin? Answer:* `False`"). For every probe we decompose the residual stream contribution to the target logit at each layer ℓ into Attention and MLP components,

$$\Delta_k^{(\ell)} = \underbrace{\mathbf{u}_k^\top \mathbf{a}^{(\ell)}}_{\Delta_{\text{Attn}}^{(\ell)}} + \underbrace{\mathbf{u}_k^\top \mathbf{m}^{(\ell)}}_{\Delta_{\text{MLP}}^{(\ell)}}, \tag{7}$$

where $\mathbf{a}^{(\ell)}$ and $\mathbf{m}^{(\ell)}$ are the attention and MLP outputs at layer ℓ, and $\mathbf{u}_k$ is the unembedding direction of the target token $k \in \{\texttt{True}, \texttt{False}\}$.

Figure 3 contrasts the original factual probes (top row, reproduced from Figure 1 for reference) against the True (middle row) and False (bottom row) variants. Probes that converge on the same surface token exhibit nearly superimposable accumulation profiles, even though the underlying factual content is entirely disjoint: *France→Paris? True* and *Korea→Asia? True* share an almost identical trajectory, and the same is true of the False pair. This replicates the central finding of Figure 1(b) in a setting where the surface output token is dictated by the verification structure of the prompt rather than by the queried fact itself. Conversely, True and False probes derived from the same base fact produce visibly distinct profiles, confirming that the mechanism assembling the target logit is tied to the output token rather than to the input content.

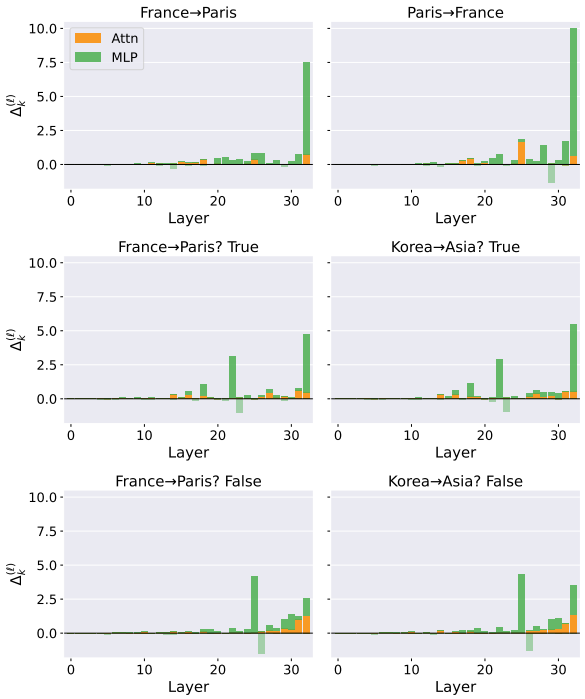

Figure 3. **Layer-wise logit accumulation profiles under verification formats.** Top: original factual completion probes (*France→Paris*, *Paris→France*). Middle: True-answer verification probes (*France→Paris? True*, *Korea→Asia? True*). Bottom: False-answer verification probes (*France→Paris? False*, *Korea→Asia? False*). Probes that produce the same output token yield nearly identical profiles regardless of underlying knowledge, while probes producing different output tokens diverge even when operating over the same fact.

While the profiles in Figure 3 provide a clear qualitative picture, they are drawn from a small set of hand-selected probes. To verify that this output-dependent organization holds at scale, we quantify the similarity of logit accumulation trajectories over a large corpus. Specifically, we compute the per-example logit-lens cumulative curve across layers for N=4,000 TOFU examples and measure the pairwise correlation between all curves, partitioning the pairs by whether the two examples share the same greedy first output token.

Figure 4 summarizes the result. Same-output pairs (n=226,514) are markedly more correlated than cross-output pairs (n=7,771,486) under both metrics: the mean Pearson correlation is 0.979 versus 0.806 (Cohen's d=1.16), and the mean Spearman correlation is 0.948 versus 0.575 (d=1.73). The gap is substantial not only in absolute terms but also relative to the within-group spread, as reflected by effect sizes exceeding one standard deviation. The larger gap under Spearman further indicates that same-output pairs agree not merely in overall magnitude but in the layer-wise ordering of logit growth, i.e., the shape of the trajectory itself. These statistics confirm at scale what Figure 3 illustrates anecdotally: the shape of the accumulation curve tracks which token is emitted rather than the underlying fact, providing quantitative evidence that logit formation is organized around the output token.

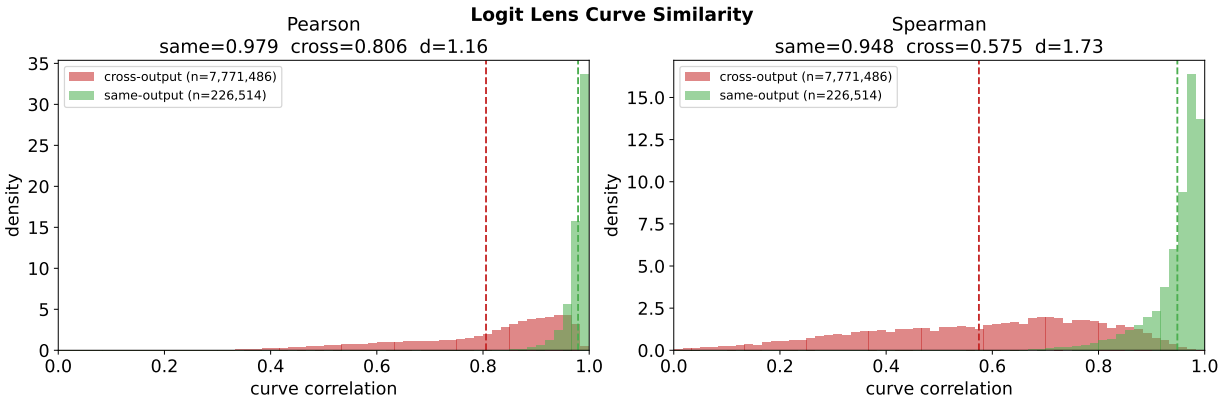

Figure 4. Logit-lens accumulation trajectories are determined by the output token. Pairwise correlation of per-example logit-lens cumulative curves across layers (TOFU, N=4,000), split by whether pairs share the same greedy first output token. Same-output pairs (green, n=226,514) are markedly more correlated than cross-output pairs (red, n=7,771,486) for both Pearson (0.979 vs. 0.806, d=1.16) and Spearman (0.948 vs. 0.575, d=1.73); dashed lines are group means. Curve shape tracks which token is emitted, not the underlying fact — evidence that logit formation is output-token dependent.

B. Empirical Validation Setup for the Same-Output Plane

In this section, we detail the methodology used to empirically validate the near-orthogonality of same-output hidden states discussed in Section 2.

Data extraction. To ensure a diverse and comprehensive analysis, we utilize the full TOFU benchmark dataset. Let the i-th sequence in the dataset be denoted as (x^i, y^i), where x^i is the input prompt and y^i is the full target sequence. To observe the model's internal geometry at the exact moment of generation, we decompose each sequence into token-level generation steps. For a given generation step t, the context provided to the model is $c_t^i = (x^i, y_{1:t-1}^i)$, and the target output token is $k = y_t^i$. We execute forward passes for all sequences and collect the corresponding hidden states $\mathbf{h}^{(\ell)}(c_t^i)$ at every layer ℓ.

Pairing and metric computation. To evaluate the geometric alignment against the unembedding matrix, we group the collected token-level hidden states into two distinct sets for comparison. *same-output pairs* are sampled from independent contexts $(c_{t_1}^{i_1}, c_{t_2}^{i_2})$ that happen to predict the exact same output token k, and we compute their hidden state difference $\Delta\mathbf{h}^{(\ell)} = \mathbf{h}^{(\ell)}(c_{t_1}^{i_1}) - \mathbf{h}^{(\ell)}(c_{t_2}^{i_2})$. *Cross-target pairs* are sampled at random from contexts that predict different output tokens, $k_1 \neq k_2$; the hidden difference is computed in the same way, and we evaluate its projection against the unembedding vector of the first token k_1.

For both paired sets, we compute the cosine similarity between the difference vector $\Delta\mathbf{h}^{(\ell)}$ and the normalized unembedding row $\hat{\mathbf{u}}_k = \mathbf{u}_k / \|\mathbf{u}_k\|_2$. A value of $\cos(\Delta\mathbf{h}^{(\ell)}, \hat{\mathbf{u}}_k) \approx 0$ indicates near-perfect orthogonality. As plotted in the main text, we average these cosine similarity values across all sampled pairs at each layer; the shaded regions represent the standard deviation, highlighting the strict constraints on same-output pairs compared to the unconstrained behavior of cross-target pairs.

C. Computational Efficiency in Token-Level Analysis

Our empirical validation uses the full TOFU benchmark, where every sequence is decomposed into token-level contexts. This yields a total sample count N_{total} on the order of millions.

Pairwise complexity analysis. For each target token $k \in K$, let N_k be the number of occurrences, with $\sum_{k \in K} N_k = N_{\text{total}}$. A naive approach to computing the "true" cross-output margin requires comparing every token k against all other tokens $j \neq k$. The number of required pairwise comparisons is

$$P_{\text{cross}} = \sum_{k \in K} N_k \left(N_{\text{total}} - N_k\right) = N_{\text{total}}^2 - \sum_{k \in K} N_k^2. \tag{8}$$

With $N_{\text{total}} \approx 10^6$, this $O(N_{\text{total}}^2)$ complexity demands over 10^{12} operations, which is computationally prohibitive for a precomputation pass.

BOG's linear scalability. BOG bypasses this bottleneck by reducing the problem to $O(N_{\text{total}} + |K|)$. The anchors are computed in $O(N_{\text{total}})$ via a single forward pass and group-averaging, and the margin is computed in $O(|K|)$ by comparing the precomputed anchors against a single global centroid $\mathbf{h}_g^{(\ell)}$. This ensures that the cross-output margin $\mu^{(\ell)}$ reliably approximates the expected divergence (the blue line in Figure 2) while remaining scalable to massive datasets. The statistical bias introduced by including token k in $\mathbf{h}_g^{(\ell)}$ is proportional to the relative frequency $p_k = N_k/N_{\text{total}}$, which is effectively ignorable in high-diversity benchmarks.

D. Experimental Setup

This appendix complements the brief setup description in Section 4 with the full details of the dataset splits, model checkpoints and evaluation metrics used in our TOFU experiments.

D.1. Dataset and Splits

TOFU (Maini et al., 2024) consists of synthetic biographies of 200 fictitious authors, each accompanied by 20 question–answer pairs, yielding 4,000 Q&A items in total. The benchmark provides three forget splits of increasing size containing 1%, 5%, and 10% of the authors as the forget set $\mathcal{D}_f$, with the remaining authors forming the retain set $\mathcal{D}_r$.

D.2. Models and Reference Checkpoints

We adopt two backbones spanning a full order of magnitude in scale: LLAMA-3.2-1B and LLAMA-3.1-8B (Touvron et al., 2023). For each backbone, we obtain the TARGET model by full-parameter fine-tuning on the entire TOFU corpus, and the unlearning procedure is then applied to this TARGET model using only $\mathcal{D}_f$ and $\mathcal{D}_r$. Two non-unlearning checkpoints frame the evaluation. TARGET provides the upper bound on retained knowledge and the lower bound on forget quality. ORACLE, fine-tuned on $\mathcal{D}_r$ alone with the forget authors never seen, serves as the gold standard that any unlearning method should approximate.

D.3. Evaluation Metrics

We report the two standard TOFU metrics, both with higher-is-better orientation. *Forget Quality* (FQ) is the p-value of a Kolmogorov–Smirnov test comparing the truth-ratio distribution of the unlearned model on $\mathcal{D}_f$ against that of ORACLE; FQ approaches 1 when the unlearned model is statistically indistinguishable from ORACLE on the forget set. *Model Utility* (MU) aggregates probability, ROUGE, and truth-ratio scores across the retain set, real authors, and world facts, capturing how well the model preserves general capability after unlearning.

E. Related Work on LLM Unlearning

LLM unlearning seeks to eliminate the influence of a designated forget set $\mathcal{D}_f$ from a trained model while sustaining its competence on the retain set $\mathcal{D}_r$ (Liu et al., 2024a; Maini et al., 2024). Prior approaches fall broadly into three families: gradient-based, preference-optimization-based, and representation-based.

Gradient-based methods. The most direct formulation is Gradient Ascent (GA) (Thudi et al., 2022), which inverts the standard training signal on $\mathcal{D}_f$ to drive the model away from forget targets. Gradient Difference (GD) (Liu et al., 2025) attempts to stabilize the procedure by adding a standard descent term over $\mathcal{D}_r$, trading off suppression of forget content against preservation of retain capability. Related variants instead constrain the update via KL regularization toward the pre-unlearning model (Maini et al., 2024). A common shortcoming of this family is that the post-unlearning behavior on $\mathcal{D}_f$ is left implicit, frequently surfacing as hallucinated or incoherent generations, and the gradients touch every layer, leaving entanglement between $\mathcal{D}_f$ and $\mathcal{D}_r$ unresolved.

Preference-optimization-based methods. A second line reformulates unlearning as alignment to a preferred response. Negative Preference Optimization (NPO) (Zhang et al., 2024) removes the positive branch of the DPO objective and instead slows the model's drift on $\mathcal{D}_f$, sidestepping the collapse mode of GA-style methods. SimNPO (Fan et al., 2024) streamlines NPO by eliminating its reliance on a reference model, retaining comparable performance with a lighter loss.

Representation-based methods. More recent work intervenes directly in the residual stream. RMU (Li et al., 2024) adjusts the MLP weights of a handful of early-to-mid layers so that activations on $\mathcal{D}_f$ are pushed in random directions, with an

auxiliary loss anchoring $\mathcal{D}_r$ activations to their pre-unlearning values. While effective for broad-capability removal (e.g., hazardous-knowledge benchmarks), the random-target design transfers poorly to instance-level unlearning, where forget and retain examples lie close together lexically and semantically (Liu et al., 2024a). LUNAR (Shen et al., 2026) replaces the random target with a behaviorally meaningful one, steering forget activations toward regions that naturally elicit the model's own expression of ignorance.

BOG also operates in the representation space but takes a different starting point. Whereas RMU prescribes random directions and LUNAR prescribes refusal-aligned regions as the target activation for the forget set, BOG derives its constraint from the model's own output geometry: forget representations must leave the affine plane shared with retain contexts producing the same token, with the required displacement calibrated by a margin estimated directly from retain anchors, without an external reference set or a tuned threshold.