Which Neighbors Suffer? Predicting Collateral Damage from Gradient Ascent in LLM Unlearning

Anonymous ACL submission

Abstract

Gradient Ascent (GA)-based machine unlearning can degrade non-target concepts, but it is unclear which pre-unlearning signals predict this collateral damage. We frame damage prediction as a ranking problem and systematically compare output-space, parameter-footprint, and gradient-direction proximity measures against GA-induced NLL increase across a controlled 30-concept inventory and models from 0.5B to 8B parameters. Gradient cosine similarity is the strongest signal in the same-prompt Llama sweeps, while rank correlation and top-risk retrieval favor different measures at smaller scale; no single metric dominates once model family, Fisher exactness, prompt overlap, objective, and behavioral probe vary. Prompt-disjoint evaluation and behavior-level probes both weaken the practical interpretation of this advantage, suggesting that GA-induced NLL damage is governed by prompt-level gradient geometry; we therefore position these results as a mechanistic probe rather than a deployable safety audit.

1 Introduction

Gradient Ascent (GA) unlearning modifies model parameters by increasing the loss of a forget target. The same update can also increase the negative log-likelihood (NLL) of retained, non-target concepts. This paper asks a mechanistic question: before running GA, which pre-unlearning signals predict where this loss-level collateral damage will propagate? The question is motivated by practical unlearning settings such as privacy, copyright, and content-policy changes (Bourtoule et al., 2021; Liu et al., 2025), but our scope is narrower: NLL damage on controlled prompts, not full behavioral safety.

We frame the problem as a ranking task. Given a forget target, a predictor ranks non-target concepts by expected damage; we then compare that ranking against the actual GA-induced NLL increase. Prior work shows that knowledge *localization* and *editing* success are often uncorrelated (Hase et al., 2023) and that knowledge superposition in LLMs follows heavy-tailed distributions (Hu et al., 2025), implying that parameter-space structure, prompt-level NLL, and downstream behavior need not align.

We present a systematic comparison of proximity measures as collateral-damage predictors, spanning output-space (embedding cosine, Poincaré), parameter-footprint (FFO), directional (GRADCOS), and Jacobian sketch approaches. Predictor quality varies with model family, target domain, prompt overlap (whether predictor and damage evaluation share prompts), unlearning objective, and Fisher approximation budget; no single measure dominates uniformly. To avoid selective reporting, we base the primary conclusions on broad 30-concept, 8-target, 3-seed sweeps, including a full Llama-3.1-8B run, and treat older fixed-target experiments as auxiliary evidence. We refer to the Llama 1B–8B sweeps under same-prompt GA evaluation as the *main Llama regime*. We further add reviewer-motivated stress tests against trivial baselines, prompt-disjoint transfer, Gradient Difference, generation QA, a small TOFU-style entity inventory, and predictor wall-clock cost. These checks delimit the scope: GRADCOS is informative in the main Llama regime, but the prompt-disjoint drop suggests that a single concept-level damage score is insufficient. Practical diagnostics should instead reason over specific protected prompts or prompt distributions.

Contributions.

1. A controlled comparison of non-oracle proximity measures for pre-unlearning NLL-level collateral-damage prediction across Qwen2.5, Llama-3.2 1B/3B, and Llama-3.1-8B settings.

2. Evidence that predictor quality is regime de-

pendent: exact Fisher footprints are useful for Qwen top-risk shortlisting when affordable, while GRADCOS is the strongest signal in the main Llama regime.

3. Ranking-centric evaluation with bootstrap intervals, AUPRC, Precision@k, Recall@k, target-level summaries, prompt leave-one checks, threshold sensitivity analyses, and same-category/random baselines.

4. Stress tests showing that NLL-level predictor strength depends on prompt overlap and unlearning objective: prompt-disjoint evaluation weakens GRADCOS's advantage over cosine, a reduced Gradient Difference check preserves only partial GRADCOS signal, a TOFU-style entity check is underpowered, behavior-level probes remain weak or inconclusive, and GRADCOS is substantially more expensive than embedding cosine.

2 Background and Definitions

2.1 Prediction Task and Terminology

We study a pre-computation prediction task: given a forget target c_f and candidate concepts $c \in C \setminus \{c_f\}$, rank the candidates by the damage they will suffer after GA-based unlearning of c_f. All predictor scores are oriented so that larger values indicate larger predicted collateral damage. Throughout the paper, GA denotes *Gradient Ascent*, the unlearning procedure, while GRADCOS denotes *Gradient Cosine*, the gradient-cosine predictor. We use FFO for Fisher Footprint Overlap.

These measures span three design axes: *space* (output vs. parameter), *aspect* (footprint overlap vs. directional alignment), and *cost* (forward pass only vs. gradient/Fisher computation). Output-space measures (Measures 1–2) test whether semantic or representational closeness alone predicts vulnerability, rather than assuming that semantic neighbors are equivalent to structural neighbors. FFO (Measure 3) tests whether two concepts share overlapping sensitive parameters, regardless of gradient direction. GRADCOS (Measure 4) tests whether their loss gradients point in mutually destructive directions—directly motivated by the first-order term of GA damage (see below). Jacobian sketches (Measure 5) test whether richer input-output sensitivities add signal beyond scalar gradient cosine. The multi-step GA collapse curve (Measure 6) is not a pre-computation predictor but an oracle diagnostic for separability of damage dynamics.

Local geometry of GA damage. The connection between these measures can be seen from a local expansion of one GA update. Let $\mathcal{L}_f(\theta)$ be the forget-target NLL and $\mathcal{L}_r(\theta)$ be the NLL of a retained or candidate concept. Gradient Ascent on the forget target applies

$$\theta' = \theta + \eta \nabla_\theta \mathcal{L}_f(\theta).$$

The induced change in the retained loss is locally

$$\begin{aligned}\Delta\mathcal{L}_r &= \mathcal{L}_r(\theta') - \mathcal{L}_r(\theta) \\ &\approx \eta \langle \nabla_\theta \mathcal{L}_r, \nabla_\theta \mathcal{L}_f \rangle + \frac{\eta^2}{2} \nabla_\theta \mathcal{L}_f^\top H_r \nabla_\theta \mathcal{L}_f,\end{aligned}$$

where $H_r = \nabla_\theta^2 \mathcal{L}_r(\theta)$ is the Hessian of the retain loss. The first-order inner-product term directly motivates GRADCOS: when the retain and forget gradients are aligned, GA increases the retain loss. FFO captures a complementary, direction-blind question—whether the two concepts depend on overlapping sensitive coordinates—while Jacobian sketches test whether richer local sensitivities add signal beyond gradient cosine. Because real unlearning uses finite multi-step updates, higher-order effects can matter; this expansion motivates the predictors but does not replace empirical damage measurement.

Scope: GA-specific damage geometry. The first-order term above is specific to the GA update $\theta' = \theta + \eta \nabla_\theta \mathcal{L}_f$. For Gradient Difference (Maini et al., 2024), a stylized update is $\theta' = \theta + \eta(\nabla_\theta \mathcal{L}_f - \lambda \nabla_\theta \mathcal{L}_{\text{retain}})$, giving the local candidate-loss change $\Delta\mathcal{L}_c \approx \eta \langle \nabla \mathcal{L}_c, \nabla \mathcal{L}_f \rangle - \eta\lambda \langle \nabla \mathcal{L}_c, \nabla \mathcal{L}_{\text{retain}} \rangle$. Thus GRADCOS with the forget target captures only the first term: it may remain informative when the forget-gradient component dominates, but it is no longer a complete predictor of the update direction. NPO and DPO-based approaches likewise induce different collateral-damage geometries. Our main sweeps therefore focus on GA, with a reduced Gradient Difference stress test reported in Section 3.2.4.

2.2 Structural Neighbors

We adopt a parameter-space definition of structural proximity grounded in Fisher information.

Definition 1 (Parameter Footprint)**.** *Given a concept c represented by prompts $\mathcal{X}_c$, and a model*

with parameters $\theta \in \mathbb{R}^N$, the parameter footprint *of c is the set of parameters to which c's output distribution is sensitive:*

$$\mathrm{FP}(c) = \{i \in \{1, \ldots, N\} : s_i(c) > \tau\}$$

where $s_i(c)$ is a specificity ratio[1] *measuring the relative importance of parameter i for concept c against all other candidate concepts.*

Definition 2 (Structural Neighborhood Strength)**.** *The structural proximity between concepts c_A and c_B is quantified by the Jaccard coefficient of their parameter footprints:*

$$J(c_A, c_B) = \frac{|\mathrm{FP}(c_A) \cap \mathrm{FP}(c_B)|}{|\mathrm{FP}(c_A) \cup \mathrm{FP}(c_B)|}$$

2.3 Proximity Measures Under Comparison

We compare five non-oracle pre-computation proximity measures and one oracle diagnostic:

1. **Embedding Cosine (Output)**: Cosine similarity of mean-pooled final-layer hidden states. This is a low-cost semantic baseline: it asks whether concepts that are close in activation space are also vulnerable to the same parameter update.

2. **Poincaré Proximity (Output)**: Sign-reversed geodesic distance in the Poincaré ball model (Nickel and Kiela, 2017), computed after scaling mean-pooled hidden states into the ball. This tests whether hierarchical output-space structure predicts unlearning damage.

3. **Fisher Footprint Overlap (Parameter)**: Jaccard coefficient of Fisher-thresholded parameter footprints (Definitions 1–2). This measure asks whether two concepts rely on overlapping sensitive parameters, but it does not by itself measure whether their gradients point in the same update direction.

4. **GradCos (Parameter)**: Cosine similarity of loss gradients over selected upper-half SLUG layers (Selected Layers for Unlearning Gradients; indices in Appendix D), computed as $\frac{\langle \nabla_\theta \mathcal{L}_A, \nabla_\theta \mathcal{L}_B \rangle}{\|\nabla_\theta \mathcal{L}_A\| \|\nabla_\theta \mathcal{L}_B\|}$. This directly measures directional update interference and is therefore expected to be closer to GA-induced damage than pure output-space similarity.

5. **Jacobian Sketch Alignment (Parameter)**: Cosine similarity between per-concept Jacobian sketches evaluated on selected parameters. For a concept c, we define

$$\mathbf{S}_c = \mathbb{E}_{x \sim c}\left[\nabla_{\theta_S}\left(r(x)^\top z(x)\right)\right],$$

where $z(x)$ are logits and θ_S are parameters restricted to the selected SLUG layers. We choose upper-half layers following the factual-knowledge localization heuristic of Meng et al. (2022). We evaluate two choices for $r(x)$: (i) random Rademacher direction (random sketch), and (ii) objective-aligned direction $r(x) = p(x) - e_y$ (single-token CE gradient direction). This tests whether explicit input-output sensitivity adds signal beyond GradCos.

6. **Multi-step GA Collapse Curve (Parameter/Oracle)**: The slope of NLL increase over 10 GA steps, directly measuring collapse dynamics. Because it requires running GA, this is not a deployable pre-computation predictor; it is an upper-bound diagnostic.

Measures 1–2 operate in output space and require only a forward pass. Measures 3–5 require gradient or Fisher computation over selected layers. Measure 6 requires running GA itself and serves only as an oracle baseline.

3 Experiments

3.1 Experimental Setup

Models. Table 1 lists all models used in the main experiments. Qwen2.5-0.5B and Qwen2.5-1.5B are used for exact or small-scale predictor comparisons; Llama-3.2-1B and Llama-3.2-3B are used for same-release checks; Llama-3.1-8B is used for a larger Llama-family, same-inventory sweep. We keep these version boundaries explicit because Llama-3.1-8B is not a same-release continuation of the Llama-3.2 1B/3B runs. All models are publicly available causal LM checkpoints and are used under their original open-weights licenses (e.g., Apache 2.0 for Qwen, Meta Community License for Llama). An exploratory distilgpt2 Jacobian compatibility check, following the DistilBERT-style distillation setup (Sanh et al., 2019), is described in Appendix A.

[1] We define $s_i(c) = \frac{[\mathcal{I}_c]_{ii}}{\frac{1}{|C|}\sum_{c' \in C}[\mathcal{I}_{c'}]_{ii}}$ where $[\mathcal{I}_c]_{ii}$ is the diagonal Fisher information of the i-th parameter with respect to concept c; we refer to this style of thresholding as "R3". The threshold τ is determined automatically via Otsu's method applied to the distribution of $s_i(c)$ values for concept c.

Table 1: Main models used in the experiments.

Model	Size	Role
Qwen2.5-0.5B (Qwen Team, 2024)	0.5B	broad predictor and exact/proxy Fisher checks
Qwen2.5-1.5B (Qwen Team, 2024)	1.5B	medium-scale fixed-target predictor comparison
Llama-3.2-1B (Llama Team, 2024)	1B	same-release scale check
Llama-3.2-3B (Llama Team, 2024)	3B	same-release sweep and auxiliary diagnostics
Llama-3.1-8B-Instruct (Llama Team, 2024)	8B	larger Llama-family scale check

Data and Targets. We use fiction targets (Harry Potter [HP], Star Wars [SW], Lord of the Rings [LOTR]) and non-fiction targets (Quantum Mechanics [QM], Photosynthesis, Roman Empire, World War II, Artificial Intelligence); candidate pools span fiction, geography, history, science, and technology. The full Qwen, Llama-3.2 1B/3B, and Llama-3.1-8B sweeps use a 30-concept inventory with three prompts per concept, eight forget targets, three GA seeds (42, 43, 44), and 29 non-target candidates per target. An earlier Qwen fixed-target set uses 17 concepts with eight prompts each. Each concept is represented by researcher-authored factual prompts (not source-text excerpts), designed to be unambiguous single-concept references. The inventory is broader than the original fiction-only setting but remains fiction-heavy (11/30), so non-fiction results should be read as stress tests rather than as domain-general evidence. Prompt leave-one robustness checks use three two-prompt subsets created by dropping one prompt index. To test whether the same signals survive at finer granularity, we also run a reduced TOFU-style entity check using fictitious-author entities from TOFU (Maini et al., 2024); this is not a full TOFU benchmark evaluation. The full inventory is in Appendix C and `data/concepts_prompts.json`.

Ground Truth. Collateral damage is measured as the GA-induced NLL increase averaged over concept prompts. GA budgets vary by experiment (Appendix D): the fixed-target runs use 15 steps, while all broad runs use 10 steps. The Llama runs use smaller learning rates (8×10^{-6} for 1B, 5×10^{-6} for 3B, 3×10^{-6} for 8B), fp32 master parameters for selected fp16 weights in the 1B/3B runs, float16 weights in the 8B run, and gradient clipping. We also report behavior-level probes: a topic-label margin probe, $\Delta_{\text{beh}} = \text{margin}_{\text{base}} - \text{margin}_{\text{post}}$, and a reduced generation-QA probe. These tests ask whether NLL-level rankings transfer to behaviorally meaningful output changes; they do not replace full task-level evaluation. For binary damage labels, we apply Otsu thresholding (Otsu, 1979) to the NLL-increase distribution. Otsu's method finds a threshold that maximizes between-class variance, which is appropriate here because GA damage distributions are typically bimodal (a large group of minimally affected concepts and a small group of heavily damaged ones).

Hyperparameters, Compute, and Reproducibility. Unless otherwise stated, Gradient Ascent uses a learning rate of 3×10^{-4} to 3×10^{-6} for 8–15 steps in predictor and auxiliary evaluations. The broad Qwen, Llama-3.2 1B/3B, and Llama-3.1-8B predictor runs use three seeds (42, 43, 44). Experiments were conducted on CUDA GPU environments, including Google Colab T4 instances (16GB VRAM) and RunPod workers, using PyTorch and the HuggingFace Transformers library. The 8B sweep was split across two RTX 3090 GPUs, run in float16, and merged before metric computation. Exact commands, start/finish timestamps, package versions, raw CSV/JSON outputs, and validation scripts are included in the supplementary artifact. Appendix D summarizes the paper-facing hyperparameters and runtime details.

Evaluation Metrics. The primary metric is Spearman rank correlation ρ between each proximity measure and the GA-induced damage vector; we also report bootstrap confidence intervals (CIs), AUPRC, Precision@k, and Recall@k. AUPRC and Precision@k measure high-risk retrieval after thresholding, but they become easier when the Otsu-positive rate is high, so we interpret them with rank correlation. In multi-seed sweeps, P@k is computed over pooled candidate-seed rows; because GRADCOS is seed-invariant, this can inflate P@k through candidate repetition (Appendix Table 11). Bootstrap intervals reflect within-inventory stability rather than independent sampling from all possible concepts. Post-hoc analyses vary the damage threshold and compare against random rankers, a same-category baseline, and a leave-one-target ridge combiner. The same-category baseline assigns score 1 to candidates sharing the forget target's broad inventory category (fiction, geography,

Table 2: Broad predictor comparison on Qwen2.5-0.5B.

Measure	$\bar{\rho}$ [95% CI]	AUPRC	P@1	P@3	P@5
Poincaré	**+0.475 [+0.297, +0.618]**	0.578	0.250	0.375	0.400
GRADCOS	+0.461 [+0.273, +0.621]	0.586	0.375	0.375	0.525
Embedding cosine	+0.410 [+0.228, +0.566]	0.602	0.250	0.417	0.450
Exact FFO (R3)	+0.272 [+0.076, +0.446]	**0.610**	**0.750**	**0.750**	**0.700**
Jacobian (obj-align)	+0.119 [−0.091, +0.325]	0.498	**0.750**	0.708	0.550
Proxy FFO (R3)	−0.020 [−0.231, +0.187]	0.394	0.125	0.250	0.300

Table 3: Llama-3.2-3B broad sweep macro metrics.

Measure	$\bar{\rho}$ [95% CI]	AUPRC	P@1	P@3	P@5
GRADCOS	**+0.811 [+0.710, +0.881]**	**0.835**	**1.000**	**1.000**	**1.000**
Embedding cosine	+0.473 [+0.283, +0.634]	0.626	0.875	0.833	0.700
Poincaré	+0.213 [+0.009, +0.414]	0.349	0.125	0.125	0.325
Proxy FFO (R3)	+0.143 [−0.080, +0.364]	0.447	0.625	0.667	0.525

history, science, or technology) and 0 otherwise; it is not intended to order candidates within a category.

3.2 Results: Controlled Predictor Comparison

3.2.1 Qwen2.5-0.5B: Rank and Retrieval Diverge

To avoid relying on the earlier fixed-target checks, we ran all non-oracle measures under identical conditions on Qwen2.5-0.5B over a broader 30-concept inventory, eight forget targets, and three GA seeds (Table 2). The expanded target set changes the interpretation. Poincaré, GRADCOS, and cosine are the strongest macro rank predictors, but their Spearman CIs overlap; exact FFO is the strongest top-risk shortlisting signal, achieving P@1 = 0.750, P@3 = 0.750, and P@5 = 0.700 despite a lower macro Spearman correlation. The neuron-proxy FFO is substantially weaker than exact FFO in this small model, showing that the proxy should not be treated as a drop-in replacement when exact Fisher is computable. The main lesson is therefore not that one predictor dominates, but that rank prediction and high-risk retrieval favor different proximity notions.

Target-level metrics show why the macro ranking should be interpreted cautiously. Star Wars is highly predictable by cosine and exact FFO, Photosynthesis and Roman Empire are best predicted by GRADCOS, and Harry Potter/Lord of the Rings favor Poincaré in rank correlation. These target-level differences are the reason we avoid claiming a universally dominant predictor.

Prompt leave-one robustness. To test whether the Qwen result depends on the particular three prompts per concept, we repeated the Qwen broad run on three leave-one prompt subsets, each dropping one prompt index and using the remaining two prompts. Under this stress test (Appendix Table 8), GRADCOS becomes the strongest rank predictor (mean $\bar{\rho} = +0.622$), while exact FFO remains strong for top-risk retrieval (P@5 $= 0.800$). This does not invalidate the full-prompt Poincaré result, but it weakens any interpretation that output-space proximity alone gives a stable account of collateral risk.

3.2.2 Llama-3.2-3B: Gradient Cosine Is Strongest

To isolate target variation from architecture and tokenizer effects, our main 3B comparison uses Llama-3.2-3B with the same 30-concept, eight-target inventory (Table 3).

Table 3 yields clearer trends than the Qwen sweep: GRADCOS leads both rank correlation and top-risk retrieval, while cosine is the strongest output-space baseline. The proxy Fisher row is weaker, but this should not be read as evidence against exact Fisher overlap: a Llama-3.2 follow-up with exact per-parameter FFO improves substantially over the proxy (1B $\rho = .597$ vs. $-.021$; 3B $\rho = .526$ vs. $.155$), while still trailing GRADCOS (1B $\rho = .747$; 3B $\rho = .817$; Appendix Table 12). GRADCOS remains positive for every target, including all non-fiction targets, so directional gradient interference is the strongest signal in this Llama-3.2-3B setting.

On the perfect P@k scores. The pooled P@k $= 1.000$ values are inflated by seed repetition: GRADCOS assigns identical scores across seeds, so the same candidate can fill multiple top ranks. Candidate-deduplicated P@5 is 0.925 (1B), 0.775 (3B), and 0.975 (8B); see Appendix Table 11. The result is also not explained by a pure genre shortcut, as non-fiction targets yield high GRADCOS correlations and same-category baselines trail GRADCOS (Table 7).

The same sweep on Llama-3.2-1B is directionally consistent (Appendix Table 9): GRADCOS is the strongest macro predictor ($\bar{\rho} = +0.746$

[+0.621, +0.829]), while cosine is more competitive at 1B than at 3B but still trails GRADCOS. Together, the 1B/3B Llama-3.2 and 8B Llama-3.1 results support a limited conclusion: GRADCOS is the strongest NLL-level signal in the main Llama regime, though these results are not a single-release scaling curve and do not establish a cross-family scaling law.

3.2.3 Llama-3.1-8B Full Sweep

To check whether the same pattern survives at a more practically relevant scale, we ran a full sweep on Llama-3.1-8B-Instruct with the same eight targets, three GA seeds, predictor pipeline, and 10-step GA budget as the Llama-3.2 sweeps, using a smaller learning rate. This is a larger Llama-family check, not a same-release scaling curve: the 1B/3B models are Llama-3.2, while the 8B model is Llama-3.1. Table 4 shows that GRADCOS remains the strongest predictor on every macro metric: $\bar{\rho} = +0.798$ with a 95% bootstrap CI of $[+0.691, +0.867]$, AUPRC $= 0.908$, and pooled candidate-seed P@1/P@3/P@5 $= 1.000$. The GRADCOS lower CI bound exceeds the upper CI bound of the strongest baseline, embedding cosine ($+0.650$), so the separation is not just a point-estimate artifact. Cosine is the strongest output-space baseline, but it trails GRADCOS on both ranking and retrieval. Poincaré stays positive but lower on rank correlation and AUPRC, and the neuron-proxy Fisher measure is near zero on macro rank correlation. Target-level and Otsu-positive breakdowns also argue against a fiction-cluster shortcut: non-fiction targets remain positive under GRADCOS, and the damaged sets are mixed-category (Appendix Table 13). Otsu-positive rates vary by target, making P@k easier to satisfy for some targets. We therefore interpret P@k alongside Spearman ρ, AUPRC, and the deduplicated audit rather than as a standalone difficulty-normalized score. This result broadens the model scale to 8B within the Llama family, but it remains insufficient for a cross-family scaling law.

SLUG layer sensitivity. The main Llama-3.1-8B run uses fixed upper-layer SLUG choices. To test whether GRADCOS depends on that particular choice, we ran a reduced one-seed, eight-target 8B ablation across four layer sets (Table 5). GRADCOS remains the top signal in all four settings, with ρ in $[.730, .811]$ and AUPRC in $[.848, .896]$. P@5 varies more (.825–.975), so layer choice affects top-risk magnitude, but the qualitative ranking is stable in this check.

Table 4: Llama-3.1-8B full sweep over eight targets and three GA seeds.

Measure	$\bar{\rho}$ [95% CI]	AUPRC	P@1	P@3	P@5
GRADCOS	**+0.798 [+0.691, +0.867]**	**0.908**	**1.000**	**1.000**	**1.000**
Embedding cosine	+0.504 [+0.331, +0.650]	0.618	0.875	0.833	0.800
Poincaré	+0.195 [+0.014, +0.373]	0.509	0.750	0.875	0.700
Proxy FFO (R3)	+0.010 [−0.203, +0.238]	0.395	0.500	0.500	0.400

Table 5: Llama-3.1-8B SLUG layer sensitivity.

Layer set	$\bar{\rho}$ [95% CI]	AUPRC	P@5
Paper layers 20,22,24,26,28,30	.800 [.583,.912]	.896	.975
Upper quarter 24–29	.730 [.448,.880]	.858	.875
Upper half 16,18,20,22,24,26	.786 [.566,.905]	.848	.825
Lower half 4,6,8,10,12,14	.811 [.601,.913]	.892	.875

3.2.4 Reviewer-Motivated Stress Tests

We ran a compact follow-up suite to test whether the 8B GRADCOS result is explained by trivial structure, whether the signal transfers across prompts, whether it transfers beyond vanilla GA, whether it connects to behavior, and how expensive it is. These checks use reduced settings (typically four targets and one seed on Llama-3.1-8B) and should be read as stress tests, not as replacements for the full sweeps. They address five methodological concerns: trivial category baselines, prompt overlap, non-GA transfer, behavioral transfer, and entity-level granularity. The reduced 8B GA run shows that GRADCOS is not reducible to broad category membership: GRADCOS reaches $\rho = .778$ and P@5=.950, while a same-category baseline reaches $\rho = .311$ and P@5=.400, and random ranking reaches $\rho = .058$ and P@5=.345. The Llama-3.2-3B prompt-disjoint check, where predictors and GA use two prompts while damage is measured on the held-out prompt, weakens the GRADCOS advantage: GRADCOS remains positive ($\rho = .314$, AUPRC=.578) but is comparable to embedding cosine ($\rho = .348$, AUPRC=.593; Appendix Table 10). This argues against treating GRADCOS as a general-purpose concept-level predictor, but it is a useful granularity finding: GA-NLL damage appears substantially prompt-specific. The drop may reflect both removal of shared prompt

Table 6: Fisher approximation gap: exact diagonal vs. neuron-proxy.

Fisher Variant	0.5B HP	0.5B SW	1.5B HP	1.5B SW
Exact diagonal	$+0.850$	$+0.818$	$+0.446$	$+0.461$
Neuron-proxy	$+0.243$	$+0.439$	-0.007	-0.032
Gap (Exact − Proxy)	$+0.607$	$+0.379$	$+0.454$	$+0.493$

information and noisier one-prompt damage labels. The Gradient Difference check gives only weak partial transfer (GRADCOS $\rho = .314$ vs. random $\rho = -.003$), so forget-gradient GRADCOS should not be treated as practically sufficient for GD. The generation-QA probe is inconclusive rather than negative (GRADCOS $\rho = -.023$, base QA accuracy=.211), the TOFU-style entity check is underpowered (GRADCOS $\rho = .438$ over eight entities, with retrieval close to random), and the timing probe shows that GRADCOS is about $269\times$ slower than embedding cosine. Appendix Table 7 gives the compact breakdown.

3.2.5 Budget-Normalized Fisher: Exact vs. Neuron-Proxy

To quantify the impact of Fisher approximation on predictor quality (addressing the computational-budget confound noted in Section 4), we compared exact per-element Fisher diagonal against the neuron-level sum-pooled proxy used at 3B scale. Both variants use identical R3 thresholding and Jaccard computation, isolating the approximation effect. Experiments were conducted on Qwen2.5-0.5B with SLUG layers 19–20 and Qwen2.5-1.5B with SLUG layers 26–27, using two forget targets.

Table 6 shows consistent degradation: mean Spearman gaps of $+0.493$ (0.5B) and $+0.474$ (1.5B). The absolute strength of exact Fisher is regime-dependent, but this controlled comparison confirms that neuron-level pooling discards substantial signal. Consequently, the weak Fisher rows at 3B scale (Table 3) reflect approximation loss combined with target/model-regime effects, not intrinsic weakness of exact Fisher overlap.

Appendix A reports Jacobian sketches, mitigation checks, threshold sensitivity, the ridge combiner, and oracle-style multi-step GA slope diagnostics; none change the main NLL-level conclusions.

3.3 Behavior-Level Probes

To test whether NLL-level rankings transfer to behaviorally meaningful signals, we evaluated a topic-label margin probe on the same inventory. Transfer is much weaker: the best rank signals reach only $\rho = 0.237$ (Qwen2.5-0.5B) and $\rho = 0.132$ (Llama-3.2-3B; Appendix A). We also ran a reduced generation-QA probe on Llama-3.1-8B. The probe is inconclusive: GRADCOS has $\rho = -0.023$ against QA damage, but the base QA accuracy is only 0.211, so the format itself is too weak to distinguish predictor failure from evaluation failure. These results prevent us from equating NLL damage with behavioral damage and motivate treating NLL-level and behavior-level neighborhoods as related but distinct.

4 Discussion

Predictor roles. GRADCOS is strongest within the main Llama regime. No single measure dominates once model family, Fisher exactness, prompt overlap, unlearning objective, and behavioral measure vary: rank correlation and top-risk retrieval favor different signals on Qwen2.5-0.5B, exact FFO is much stronger than proxy FFO on Llama-3.2 follow-up runs, prompt-disjoint evaluation narrows the GRADCOS advantage, and behavior-level transfer remains weak or inconclusive. Output-space distances are useful baselines but can fail on specific targets: the chain rule decouples function value from derivative structure, so nearby activations need not share sensitive parameters. Exact and neuron-proxy FFO should also not be conflated: Table 6 shows that coarse pooling discards much of the useful Fisher signal. The same-category and random baselines in Table 7 further show why trivial baselines are necessary: broad category membership explains some top-ranked damaged candidates, but not the full gradient-alignment ranking.

Damage granularity. GRADCOS reaching $\rho \approx .8$ on Llama is consistent with the first-order GA account, but the prompt-disjoint check shows that this structure does not transfer cleanly to held-out prompts. This qualifies the common concept-level framing of collateral damage: for GA, loss propagation is at least partly instance- or prompt-specific. It also cautions against interpreting neighbor collapse as a property of concept names alone when the protected workload is a distribution of concrete prompts. Thus the result is less a deployable universal predictor than evidence that future audits

should be prompt-aware.

NLL $\neq$ behavior. Predictor transfer to label-margin damage is weak, and the generation-QA probe is inconclusive because base QA accuracy is low (Section 3.3), so our conclusions are limited to NLL-level risk. This limits the practical audit interpretation of the work. The audience for the present result is therefore narrower: it is useful for understanding which parameter-space relationships mediate loss propagation under the GA updates studied here, and for designing prompt-aware retain-set diagnostics, but it should not be used as a standalone privacy, copyright, or safety evaluation.

Cost and method transfer. GRADCOS is about $269\times$ slower than embedding cosine at 8B scale, so its cost is justified only for high-fidelity prompt-specific diagnostics or for studying local gradient geometry. The reduced Gradient Difference and TOFU-style checks show partial but insufficient transfer beyond the main GA setting.

5 Related Work

Unlearning and Evaluation. Gradient-based unlearning methods (Jang et al., 2023; Yao et al., 2024) modify model parameters to suppress target knowledge, while Gradient Difference (Maini et al., 2024) adds a retain-gradient subtraction term that changes the collateral-damage geometry. Recent benchmarks such as TOFU (Maini et al., 2024), RWKU (Jin et al., 2024), MUSE (Shi et al., 2025), and WMDP (Li et al., 2024) evaluate forgetting quality and utility; concurrent anonymous work studies deployment-facing orthogonal unlearning under quantization and active retention (Anonymous, 2026). Our controlled inventory instead isolates pre-unlearning *damage prediction*. We include a small TOFU-style entity stress test, but full TOFU/RWKU predictor evaluation remains future work.

Localization and Interference. Knowledge-localization work identifies where facts are stored (Meng et al., 2022; Dai et al., 2022), but localization does not necessarily predict editing success (Hase et al., 2023); knowledge superposition further suggests pervasive structural entanglement (Hu et al., 2025). Our GRADCOS predictor is also related to gradient-conflict methods such as PCGrad, OGD, and CAGrad (Yu et al., 2020; Farajtabar et al., 2020; Liu et al., 2021), but we use gradient alignment for pre-unlearning risk prediction rather than multi-task optimization. Additional related-work details are in Appendix B.

6 Conclusion

We compared five pre-computation proximity measures as NLL-level collateral-damage predictors for GA-based machine unlearning. Predictor quality is regime dependent: exact FFO excels at top-risk retrieval on Qwen2.5-0.5B, while GRADCOS is strongest in the main Llama regime. Semantic proximity is useful but insufficient on its own: concepts that are representationally close need not share the parameter-space dependencies that mediate GA damage. The prompt-disjoint, same-category, Gradient Difference, TOFU-style, behavior-probe, and cost checks narrow the claim: GRADCOS is a strong mechanistic signal for GA-induced NLL damage in that regime, not a deployable behavioral audit, a concept-level universal predictor, or a sufficient predictor for other unlearning objectives. These findings motivate prompt-aware retain-set diagnostics: rather than assessing only concept-level risk, future pre-unlearning audits should evaluate vulnerability over the specific prompts or prompt distributions that practitioners want to protect.

Limitations

Our study has several limitations. (1) The concept inventory remains fiction-heavy (11/30) with only three prompts per concept. The prompt leave-one and prompt-disjoint analyses suggest that denser prompt coverage could change relative predictor rankings; the prompt-disjoint check also conflates reduced prompt overlap with noisier one-prompt damage labels. An auxiliary 8-prompt Qwen check exists but does not cover all predictors. (2) The TOFU-style entity check is small (eight fictitious-author entities, one seed) and retrieval discrimination is close to random, so it only partially addresses whether domain-level results transfer to fine-grained entity unlearning. (3) Some predictor comparisons involve heterogeneous computational budgets. The budget-normalized Fisher comparison (Table 6) isolates this confound for the small-model case, and the Llama-3.2 exact-FFO follow-up shows that the proxy substantially underestimates Fisher signal. Exact FFO is still not evaluated at 8B scale, and the 8B timing probe shows that GRADCOS is much costlier than embedding cosine. (4) The behavior-level margin probe

shows much weaker predictor transfer than NLL sweeps, and the generation-QA probe is inconclusive because the base model performs poorly under that QA format (Section 3.3); our conclusions therefore do not extend to task-level generation or downstream behavioral evaluation. (5) We do not claim a cross-family scaling law; the Llama-3.2 1B/3B evidence and the Llama-3.1-8B full sweep strengthen the Llama-family scale check, but they are not a same-release scaling curve and do not establish that the same predictor ordering holds for other 7B+ architectures or training recipes. (6) The 8B SLUG layer ablation reduces the concern that GRADCOS works only under a hand-picked layer set, but it is still a reduced one-seed check on one model. We therefore treat it as evidence for qualitative stability, not as a universal layer-selection recipe. (7) The main sweeps use Gradient Ascent (GA). The reduced Gradient Difference check is encouraging but not definitive, and NPO or DPO-based methods require separate damage-geometry analyses.

Ethical Considerations

Dual-Use and Intended Use. This work studies predictors for identifying concepts at risk of loss-level collateral damage during unlearning. While intended for analysis and defense (e.g., motivator for orthogonal projection), these predictors are dual-use: they could be misused to strategically optimize the suppression of knowledge domains related to a target concept. We do not recommend deploying these techniques in high-stakes domains (e.g., health, law) without specialized behavioral auditing.

Scope and Data Origins. Our evaluation uses researcher-authored concept prompts and TOFU-derived fictitious-author QA prompts; we do not redistribute copyrighted source text or use personal data. These results should be interpreted as concept-level evidence of structural interference, not as a validated solution for privacy leakage or legal copyright compliance. Furthermore, Negative Log-Likelihood (NLL) is an experimental proxy and does not directly measure downstream human harm or fairness impact.

AI Assistant Disclosure. The authors used AI assistants (Gemini by Google, Claude by Anthropic, and ChatGPT by OpenAI) to assist in text refinement, LaTeX formatting, and code execution for experiments. The final claims and technical content of the paper were reviewed and verified by the human authors.

References

Anonymous. 2026. Q-ROU: Quantization-robust orthogonal unlearning via active retention for post-deployment knowledge removal in large language models. Under review.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3762–3773. PMLR.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 17643–17668. Curran Associates, Inc.

Chenhui Hu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Knowledge in superposition: Unveiling the failures of lifelong knowledge editing for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24086–24094.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 98213–98263. Curran Associates, Inc.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, and 27 others. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. 2021. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18878–18890. Curran Associates, Inc.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194.

Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

A Auxiliary Diagnostics and Behavior Probes

Jacobian and mitigation checks. In released artifacts, objective-aligned Jacobian sketches are informative in limited settings but do not consistently surpass GradCos; random-direction sketches are too setting-specific to support a separate predictor-family claim. Likewise, Fisher-EWC (Kirkpatrick et al., 2017) and orthogonal projection show that prediction signals can inform protection, but their current evaluation is compact and trade-off dependent. We therefore treat these results as auxiliary checks rather than as evidence that the paper solves protection or behavior-level unlearning.

Multi-step GA slope. A multi-step GA trajectory can expose non-linear collapse acceleration that single-step pre-computation scores miss. We do not use this as a main predictor because it requires running the unlearning procedure whose damage we are trying to forecast. It is therefore better viewed as an oracle-style diagnostic for future work, not as a deployable pre-computation measure in the present comparison.

Threshold sensitivity and post-hoc combination. The Otsu threshold used for damaged/not-damaged labels is not the only possible high-risk definition. We therefore ran post-hoc analyses with median, 75th-percentile, 90th-percentile, and top-20% thresholds, random-ranker baselines, and a leave-one-target ridge combiner. The main NLL-level conclusions are stable under these checks. In Llama-3.2-3B, GradCos remains the best predictor across all threshold modes (e.g., Otsu AUPRC $= 0.861$ and pooled P@5 $= 1.000$; 90th-percentile AUPRC $= 0.848$ and pooled P@5 $= 1.000$).

In Llama-3.2-1B, the same pattern holds (Otsu AUPRC $= 0.808$ and pooled P@5 $= 1.000$). In Qwen2.5-0.5B, threshold sensitivity reinforces the split between roles: Poincaré is strong for rank correlation, while exact FFO and GRADCOS are more reliable for top-risk retrieval. The ridge combiner is competitive on Llama but does not change the simpler conclusion that GRADCOS is the strongest single predictor in the main Llama regime.

Follow-up stress tests. Table 7 summarizes the reduced checks used to probe trivial baselines, prompt overlap, method transfer, behavior transfer, entity granularity, and predictor cost. The reduced GA, Gradient Difference, generation-QA, and cost checks use Llama-3.1-8B with four forget targets and one seed; the prompt-disjoint check uses Llama-3.2-3B with eight targets and three held-out prompt splits; the TOFU-style check also uses Llama-3.1-8B, with eight fictitious-author entities and one seed. For Gradient Difference, we use the same forget prompts as the GA stress test and add a retain-loss subtraction term using non-target inventory prompts with retain weight 1.0. For generation QA, we prompt the model with short concept-name questions before and after unlearning and measure candidate damage by the drop in alias-matched concept-name accuracy; the low base accuracy (0.211) makes this probe inconclusive rather than evidence of absent behavioral transfer. For the TOFU-style entity check, concepts are fictitious authors from TOFU's forget split, represented by three QA-style prompts per entity; this tests finer-grained entity names but is underpowered and is not a full TOFU utility/forgetting evaluation.

Prompt leave-one Qwen check. Table 8 reports the Qwen2.5-0.5B leave-one prompt robustness check moved from the main text.

Llama-3.2-1B scale check. Table 9 reports the same-release 1B Llama scale check moved from the main text.

Prompt-disjoint Llama check. Table 10 gives the full macro metrics for the Llama-3.2-3B prompt-disjoint check. For each split, predictors and GA use two prompts and damage is measured on the held-out prompt. Because the damage side has only one prompt, the result mixes prompt-overlap removal with increased measurement noise; we therefore interpret it as evidence of prompt specificity rather than as a clean estimate of cross-prompt generalization.

Table 7: Follow-up stress tests.

Check	Setting	Key result	Interpretation
Trivial baselines	Llama-3.1-8B GA, 4 targets, 1 seed	GRADCOS $\rho = .778$, P@5=.950; same-cat. $\rho = .311$, P@5=.400; random $\rho = .058$, P@5=.345	Category helps P@1 but not the full ranking.
Prompt-disjoint	Llama-3.2-3B GA, 8 targets, 3 held-out prompt splits	GRADCOS $\rho = .314$, AUPRC=.578; cosine $\rho = .348$, AUPRC=.593	Damage is prompt-specific; GRADCOS loses its same-prompt advantage.
Gradient Difference	Llama-3.1-8B GD, 4 targets, 1 seed	GRADCOS $\rho = .314$ vs. random $\rho = -.003$; random AUPRC=.894 due to dense positives	Partial transfer; tail metrics are inflated.
Generation QA	Llama-3.1-8B GA, 4 targets, 1 seed	GRADCOS $\rho = -.023$; mean baseline QA accuracy=.211	Inconclusive low-accuracy probe.
TOFU-style entities	Llama-3.1-8B, 8 entities, 1 seed	GRADCOS $\rho = .438$; Poincaré AUPRC=.782; random AUPRC=.726	Underpowered; retrieval near random.
Cost probe	Llama-3.1-8B predictor timing	cosine=.83s/target; GRADCOS=223.25s/target ($\sim 269\times$)	High-fidelity but expensive.

Table 8: Prompt leave-one robustness on Qwen2.5-0.5B.

Measure	Mean $\bar{\rho}$	AUPRC	P@5
GRADCOS	**+0.622**	**0.751**	**0.842**
Poincaré	+0.444	0.609	0.650
Exact FFO (R3)	+0.352	0.626	0.800
Embedding Cosine	+0.256	0.519	0.542
Neuron-proxy FFO (R3)	+0.014	0.422	0.467
Jacobian (obj-align)	−0.006	0.435	0.533

GRADCOS P@k audit. Table 11 expands the Llama GRADCOS P@k results that appear perfect in the pooled candidate-seed tables. In all three Llama sweeps, the GRADCOS score for a fixed target-candidate pair is identical across the three GA seeds, while the post-GA damage label is evaluated separately for each seed. Thus pooled P@k can place the same high-scoring candidate into multiple top ranks through seed repetition. The seed-wise and candidate-deduplicated variants show that the qualitative conclusion remains strong, but the literal "perfect P@k" interpretation is too strong.

Exact FFO follow-up on Llama-3.2. Table 12 reports the Llama-3.2 exact Fisher follow-up. We omit P@k here because these are three-seed runs and row-level P@k has the candidate-seed repetition issue audited above.

8B Otsu-positive composition. Table 13 reports the category composition of Otsu-positive damaged candidate-seed evaluations in the Llama-3.1-8B

Table 9: Llama-3.2-1B scale check.

Measure	$\bar{\rho}$ [95% CI]	AUPRC	P@1	P@3	P@5
GRADCOS	**+0.746 [+0.621, +0.829]**	**0.803**	**1.000**	**1.000**	**1.000**
Embedding cosine	+0.621 [+0.455, +0.737]	0.635	0.750	0.708	0.700
Poincaré	+0.232 [+0.048, +0.406]	0.389	0.250	0.333	0.350
Proxy FFO (R3)	−0.019 [−0.232, +0.202]	0.366	0.250	0.208	0.400

Table 10: Prompt-disjoint Llama-3.2-3B check.

Measure	$\bar{\rho}$	AUPRC	P@1	P@5
Embedding cosine	.348	.593	.833	.475
GRADCOS	.314	.578	.625	.483
Poincaré	.190	.422	.333	.350
Proxy FFO (R3)	.043	.370	.292	.258

full sweep. The damaged sets are not pure fiction clusters: Artificial Intelligence, Quantum Mechanics, Roman Empire, Photosynthesis, and World War II all include substantial non-fiction positives. The corresponding non-fiction target-level GRADCOS correlations are also positive (AI $= +0.683$, Photosynthesis $= +0.835$, QM $= +0.619$, Roman Empire $= +0.847$, World War II $= +0.815$). The positive rate varies substantially across targets (e.g., 57/87 for Artificial Intelligence vs. 19/87 for Photosynthesis), so P@k is easier to satisfy for some targets and should be interpreted together with rank correlation and AUPRC. This supports the interpretation that the strong GRADCOS results are not explained solely by a fiction/non-fiction shortcut.

Figure 1 gives the illustrative Qwen HP case where high output-space similarity creates semantic false positives.

Behavior-level probes. Table 14 summarizes the topic-label margin probe and reduced generation-QA probe. Predictor transfer from NLL-level damage to behavior-level damage is weaker and less stable than the NLL results, so we treat these probes as validity checks and limitations rather than as new main claims.

B Additional Related Work

Gradient Conflict and Continual Learning. Our GRADCOS predictor is connected to methods that manage destructive gradient interference. PCGrad projects gradients to reduce conflicts in multi-task learning (Yu et al., 2020), Orthogonal Gradient Descent constrains updates to preserve previous tasks (Farajtabar et al., 2020), and CAGrad optimizes for conflict-averse multi-task descent (Liu et al., 2021). We use these ideas differently: the paper's main object is pre-unlearning risk prediction rather than multi-task optimization.

Table 11: GRADCOS P@k audit for the Llama sweeps.

Model	Otsu+ rate	Pool P@3	Pool P@5	Seed P@3	Seed P@5	Dedup P@3	Dedup P@5
Llama-3.2-1B	.280 [.207,.345]	1.000	1.000	.917	.875	.958	.925
Llama-3.2-3B	.234 [.138,.414]	1.000	1.000	.903	.775	.917	.775
Llama-3.1-8B	.326 [.218,.655]	1.000	1.000	1.000	.975	1.000	.975

Table 12: Exact per-parameter FFO follow-up on Llama-3.2 1B/3B.

Model	Measure	$\bar{\rho}$ [95% CI]	AUPRC
1B	GRADCOS	.747 [.623,.831]	.809
1B	Exact FFO (R3)	.597 [.436,.724]	.730
1B	Proxy FFO (R3)	−.021 [−.236,.202]	.373
1B	Embedding cosine	.620 [.456,.737]	.649
3B	GRADCOS	.817 [.716,.886]	.834
3B	Exact FFO (R3)	.526 [.345,.677]	.704
3B	Proxy FFO (R3)	.155 [−.066,.374]	.445
3B	Embedding cosine	.471 [.280,.630]	.624

Knowledge Localization and Structural Entanglement. While causal tracing (Meng et al., 2022) and knowledge neurons (Dai et al., 2022) identify where knowledge is stored, Hase et al. (2023) show that localization success is uncorrelated with editing success, highlighting the gap between *where* knowledge resides and *how* it can be safely modified. Hu et al. (2025) demonstrate universal knowledge superposition in LLMs with heavy-tailed distributions, implying that structural entanglement is pervasive. Our Fisher-based analysis provides a concrete concept-level measure of this entanglement.

C Prompt Inventory and Experiment Mapping

The main 30-concept inventory uses researcher-authored factual sentences and is provided in the supplementary file `data/concepts_prompts.json`. The TOFU-style stress test uses TOFU-derived fictitious-author QA prompts. The prompt inventory file is generated from the experiment sources by `scripts/extract_prompt_inventory.py`, which prevents the paper-facing prompt inventory from drifting away from the code.

Table 13: Category composition of Otsu-positive damaged candidates.

Target	Pos.	Fic.	Geog.	Hist.	Sci.	Tech.
Artificial Intelligence	57	16	11	12	15	3
Harry Potter	22	12	7	3	0	0
Lord of the Rings	27	15	5	6	1	0
Photosynthesis	19	3	8	6	2	0
Quantum Mechanics	32	8	4	7	9	4
Roman Empire	22	2	10	9	1	0
Star Wars	26	17	6	3	0	0
World War II	22	5	6	8	3	0

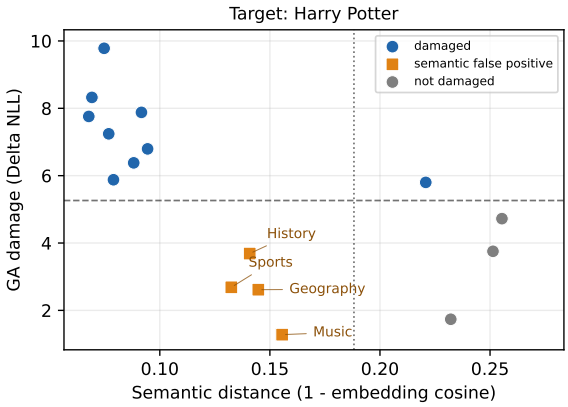

Figure 1: Semantic false positives in the Qwen2.5-0.5B HP condition. The x-axis is semantic distance, $1 - \cos(h_c, h_{\mathrm{HP}})$, so closer semantic neighbors are farther left. Orange squares mark high-cosine factual concepts (Geography, History, Music, Sports) whose GA-induced NLL damage remains low. The dotted vertical line is the cosine Otsu boundary converted to distance; the dashed horizontal line is the Otsu damage threshold. Labels are shown only for the false positives to avoid clutter.

D Hyperparameter Summary

Table 17 summarizes the main GA budgets and learning rates needed to reproduce the reported results. Exact command-line scripts, layer selections, raw outputs, and timestamps are included in the supplementary artifact.

The broad Qwen and prompt leave-one runs update or measure SLUG layers 10, 12, 14, 16, and 18; the budget-normalized Fisher run uses SLUG layers 19–20 for Qwen2.5-0.5B and 26–27 for Qwen2.5-1.5B; the Llama-3.2 1B/3B broad, exact-FFO, and prompt-disjoint runs use upper-half SLUG layers with fp32 master parameters and gradient clipping; the Llama-3.1-8B full sweep and 8B stress tests use layers 20, 22, 24, 26, 28, and 30 in float16; the 8B layer ablation additionally tests upper-quarter, upper-half, and lower-half layer sets; the projection comparison uses LoRA on Llama-3.2-3B SLUG layers 14–19 with SVD energy 0.90–0.98. The wall-clock values are RunPod measurements from the archived release logs: roughly 6 min for the fixed-target run, 13 min for the broad Qwen run, 2 min for budget-normalized Fisher, 12–17 min for Llama-3.2 1B/3B broad runs, about 22 min for the dual-RTX-3090 Llama-3.1-8B full sweep, 3–5 min for behavior probes, 7–8 min for each prompt leave-one run, 58 min for projection, and 3 min for Jacobian exploratory checks. In the Llama-3.1-8B timing probe, amortized embedding cosine costs 0.83 seconds per target while GradCos costs 223.25 seconds per target ($\sim$ 269$\times$ slower). These values exclude queueing and model-download time; exact timestamps and command lines are preserved in the released `status.tsv` files and log directories.

Table 14: Behavior-level probes.

Setting	Rank signal	Top-risk signal / caveat
Qwen2.5-0.5B	FFO proxy ($\rho = 0.237$)	Poincaré (P@5 $= 0.850$)
Llama-3.2-3B	Poincaré ($\rho = 0.132$)	FFO proxy (P@1/P@3 $= 0.750$)
Llama-3.1-8B generation QA	GradCos ($\rho = -0.023$)	Inconclusive; base QA acc. $= 0.211$

Table 15: Prompt sets used by the main experiments.

Prompt set	Models	Concepts	Prompts
Qwen v4, 8-prompt	Qwen 0.5B/1.5B	17	8
30-concept, 3-prompt	Qwen 0.5B; Llama-3.2 1B/3B; Llama-3.1 8B	30	3
30-concept, leave-one	Qwen 0.5B	30	2
30-concept, prompt-disjoint	Llama-3.2-3B	30	2+1
TOFU-style entity stress	Llama-3.1-8B	8	3

Table 16: Representative prompts from the released inventory.

Concept	Category	Example prompt
Harry Potter	fiction	Harry Potter is a wizard who studies at Hogwarts.
Star Wars	fiction	Luke Skywalker learns the ways of the Force to become a Jedi.
Lord of the Rings	fiction	Frodo Baggins carries the One Ring to Mordor to destroy it.
Quantum Mechanics	science	Particles can exist in a superposition of states until measured.
France	geography	Paris is the capital and most populous city of France.
World War II	history	The Allied forces launched the Normandy landings on D-Day.
Artificial Intelligence	technology	Neural networks are computing systems inspired by the human brain.

Table 17: GA budgets and learning rates for the principal experiment families.

Experiment family	Model(s) and steps	LR
Fixed-target v4	Qwen2.5-0.5B/1.5B, 15	2×10^{-5}–5×10^{-5}
Broad Qwen predictor	Qwen2.5-0.5B, 10	5×10^{-5}
Budget-normalized Fisher	Qwen2.5-0.5B/1.5B, 10	3×10^{-4}
1B/3B broad predictor	Llama-3.2-1B/3B, 10	8×10^{-6} / 5×10^{-6}
8B full sweep	Llama-3.1-8B, 10 (8 targets, 3 seeds, fp16)	3×10^{-6}
Llama exact FFO follow-up	Llama-3.2-1B/3B, 10	8×10^{-6} / 5×10^{-6}
8B SLUG layer ablation	Llama-3.1-8B, 10 (8 targets, 1 seed, fp16)	3×10^{-6}
Behavior margin probe	Qwen2.5-0.5B / Llama-3.2-3B, 10	setting-specific
8B follow-up stress tests	Llama-3.1-8B, 10 (4 targets, 1 seed, fp16)	3×10^{-6}
Gradient Difference stress	Llama-3.1-8B, 10 (4 targets, 1 seed, fp16)	3×10^{-6}
TOFU-style entity stress	Llama-3.1-8B, 10 (8 entities, 1 seed, fp16)	3×10^{-6}
Prompt leave-one robustness	Qwen2.5-0.5B, 10	5×10^{-5}
Prompt-disjoint check	Llama-3.2-3B, 10	5×10^{-6}
Projection comparison	Llama-3.2-3B, 8	3×10^{-4}
Jacobian exploratory	distilgpt2 / Llama-3.2-3B, 1 / 10	setting-specific