FINDING THE NEURONS THAT ASK: RELATION-SELECTIVE PRIVACY CONTROL IN FINE-TUNED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model fine-tuned to extract personally identifiable information answers a request for any relation it was trained on, including relations a deployment must withhold. However, MLP neurons are polysemantic, so imprecise localization damages non-target relations. This work proposes Question-Contrast Average Precision Scoring (QCAPS), which targets the neurons that route a request rather than the knowledge it retrieves. Each neuron is scored by how well its activation separates the target question from contrastive questions over the same context. Each selected neuron is then driven below its off-state by a multiple of its own activation gap, with the multiple calibrated jointly with the selection threshold. Across five relation types on two fine-tuned extraction models, the largest overlap between any two relations' neuron sets is 71 to 250 times lower under QCAPS than under answer-presence identification. The target is suppressed to between 0.6% and 19.2% of baseline, under 10% on eight of the ten model and relation pairs, at lower cost to the remaining relations than parameter-level unlearning on nine of the ten. At a matched neuron budget, substitutions drawn from prior work leave the target at or near baseline. The intervention acts at inference time and is reversed by detaching a hook, without monosemantic decomposition or retraining.

1 INTRODUCTION

Large language models are increasingly fine-tuned to extract structured information from documents, serving as the extraction component in retrieval-augmented pipelines over private corpora (Liu et al., 2025a). Such corpora routinely contain personally identifiable information (PII), which a model trained for reliable extraction retrieves without regard to its sensitivity (Carlini et al., 2021; Lukas et al., 2023). Because the same mechanism serves a wanted and an unwanted request alike, the privacy risk is a property of the extraction capability rather than a defect alongside it. Since extraction is what the model is deployed to perform, the capability cannot be removed, and the available lever is control over which relations it extracts. For example, a single extractor may serve a billing workflow that requires a customer's email and a support workflow that must not surface a home address. Retraining a separate model for each such policy is impractical.

Post hoc privacy preservation is usually approached by machine unlearning, which retrains on scrubbed data or maximizes loss on a designated forget set (Cao & Yang, 2015; Bourtoule et al., 2019; Maini et al., 2025; Yao et al., 2024a). It updates parameters permanently, so a policy that changes between deployments requires retraining. As a forget set is a collection of examples rather than a relation type, the update also reaches every relation at once. Current mechanistic editing approaches instead target internal circuits directly, but identify neurons by the presence of an answer in the document rather than by the request that retrieves it (Wu et al., 2023; Chen et al., 2024; Dong et al., 2025). Methods that address the resulting interference between relations locate its cause in neurons shared between them (Wu et al., 2024; Zhengyuan et al., 2026). A third line decomposes polysemantic neurons into monosemantic features, at the cost of training an auxiliary model (Frikha et al., 2025).

Suppressing a relation at the neuron level requires two decisions, which neurons to act on and how far to move them. Each of the methods above commits to a single intervention rule: zeroing the selected activations (Wu et al., 2023), masking them (Chen et al., 2024), patching them toward values drawn from elsewhere (Wu et al., 2024), or attenuating their projection weights under a fixed polarity rule

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Zhengyuan et al., 2026). Only the selection is varied, so how much precision localization needs, and how much displacement the intervention needs, has not been established. This research work asks how precisely neurons must be localized, and how strongly they must be suppressed, for one relation to be withheld while the others remain available. In addressing this question, the work makes four contributions.

- A training-free localization strategy, **Question-Contrast Average Precision Scoring (QCAPS)**, is proposed, which scores each neuron by how well it separates a request for the target relation from requests for the others over one document. The selected population is specific to a single relation, so one relation can be addressed without disturbing how the document is represented.
- **A neuron suppression strategy driven by the target versus non-target activation gap** is proposed, which displaces each selected neuron below its off-state at question-token positions. The displacement is scaled by a single parameter, so suppression strength is tunable per relation, and detaching the hook restores the original model exactly.
- **A dependency between selection and intensity** is established by showing that neither localization nor displacement suppresses a relation on its own. Calibrating the two as a single pair is therefore what makes selective suppression reachable.
- **A directed account of collateral cost** is provided, giving for each relation what its suppression costs every other, and showing that overlap between relations cannot explain it. A suppression policy can therefore be chosen with its cost known in advance, rather than assumed uniform across relations.

2 Related Work

MLP layers in transformers act as key-value memories holding factual associations (Geva et al., 2021; Dai et al., 2022; Qiu et al., 2024), and those associations can be localized and edited directly (Meng et al., 2022). Individual units are interpretable enough to be isolated for a single property, whether a language (Tang et al., 2024), a concept recoverable by sparse probing (Gurnee et al., 2023), or a functional role (Voita et al., 2024), and the pathways they form can be traced across layers (Yao et al., 2024b). In each case a unit is identified by which *inputs* activate it.

Recent works on PII leakage in fine-tuned models have relied predominantly on parameter-level unlearning, either by curating scrubbed training data or by optimizing against a designated forget set (Cao & Yang, 2015; Bourtoule et al., 2019; Maini et al., 2025; Yao et al., 2024a; Jang et al., 2023; Shi et al., 2024; Liu et al., 2024; Pawelczyk et al., 2024; Zhang et al., 2024; Garg et al., 2025), with the trade-offs surveyed by Blanco-Justicia et al. (2025) and evaluated against standard forget sets (Li et al., 2024). Key properties of this group are structural rather than incidental. The updates target a designated forget set rather than an individual relation. Since the updates are written in parameters, a policy change between deployments requires the unlearning strategy to be redesigned.

Neuron-level editing avoids retraining. DEPN (Wu et al., 2023) detects privacy neurons by integrated gradients and zeroes their activations, Chen et al. (2024) learn binary masks through adversarial training, and Dong et al. (2025) extend the identification to multilingual settings. All three score a neuron by whether an answer is present in the context rather than by the request that retrieves it, so the population they select is not specific to one relation. Because a population selected that way serves several relations at once, editing it for one relation disturbs the others. A line of work addresses that interference and locates its cause in the neurons the relations share. APNEAP (Wu et al., 2024) names it the privacy seesaw and patches activations toward values drawn from synthetic data, LDEDE (Zhengyuan et al., 2026) attributes it to coupled neurons and attenuates their projection weights under a polarity rule, and path-level variants extend the same premise to structured pathways (Li et al., 2026; Liu et al., 2025b). Editing of this kind is also known to propagate beyond its target (Cohen et al., 2024; Gu et al., 2024).

A second line of work intervenes on representations without editing parameters. Steering vectors built from contrastive means control a target behaviour at inference time (Subramani et al., 2022; Turner et al., 2024; Zou et al., 2023; Rimsky et al., 2024). A selected subset of internal units suffices to carry the effect (Li et al., 2023), a single direction can mediate a behaviour outright (Arditi et al., 2024), and a concept can be erased in closed form (Belrose et al., 2023). These methods target one

global behaviour, evaluated against a general-capability backstop such as a knowledge benchmark or perplexity. Control at a finer grain is limited by the units themselves, since a single neuron participates in several behaviours at once. A separate line addresses this by decomposing polysemantic neurons into monosemantic features (Frikha et al., 2025), at the cost of training an auxiliary model over each layer. Direct intervention has been found not to require that cost (Wu et al., 2025).

The proposed work intervenes on representations rather than on parameters, and differs from the lines above in what it identifies and controls. Neurons are scored by which *request* activates them on inputs held identical, rather than by which inputs activate them. That contrast makes relation-level control possible rather than merely visible. The four relations left in place are structurally identical siblings, so a capability benchmark cannot detect one of them failing. Closest to this setting, Liu et al. (2025a) score neurons by the presence of a relation in the input. They report a group that serves several relations, so deactivating it degrades all of them. In the proposed work, however, the relations left in place must survive the intervention. Both the identification and the interventions above are implemented and evaluated here at a matched neuron budget, rather than only contrasted (Section 4).

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let us consider G as a language model fine-tuned to extract structured relational attributes from text corpora, and $\mathcal{R}$ the set of relation types it extracts. Each input to G consists of a question q_r that requests the value of a relation $r \in \mathcal{R}$ and a context $\mathcal{C}$, the document from which that value is extracted. The exact-match accuracy of G on relation r over a held-out set is denoted by $\mathrm{EM}_r(G)$. Given a target relation r^*, the goal is to obtain a modified model G' that no longer answers q_{r^*} but continues to answer every q_r for $r \neq r^*$. Unlike unlearning, this modification is subject to two requirements. First, the parameters θ must remain unchanged, so that the modification is reversible and requires no retraining. Second, the document representation must remain intact, so that the same context continues to serve the remaining relations. The first requirement places the intervention on activations at inference time, and the second confines it to the token positions of the question. The intervention is parameterised by a pair (τ, σ). The threshold τ selects which neurons are modified, through a discriminability score $s(l, n, r^*)$ on each MLP neuron (l, n), where l indexes the layer and n the unit within its FFN intermediate dimension,

$$\mathcal{N}_{r^*}(\tau) = \big\{\, (l,n) : s(l,n,r^*) \geq \tau \,\big\}, \qquad K(\tau) = \big|\mathcal{N}_{r^*}(\tau)\big|, \tag{1}$$

where $\mathcal{N}_{r^*}(\tau)$ is the selected set and $K(\tau)$ its size. The intensity σ determines how each selected neuron is modified, through a forcing map ϕ_σ applied at question-token positions. Writing $a_{l,n,t}$ for the activation of neuron (l, n) at token position t,

$$a'_{l,n,t} = \begin{cases} \phi_\sigma(a_{l,n,t})\,, & (l,n) \in \mathcal{N}_{r^*}(\tau) \text{ and } t \geq t_q, \\ a_{l,n,t}, & \text{otherwise,} \end{cases} \tag{2}$$

where t_q indexes the first question token. The resulting model is written $G'(\tau, \sigma)$.

Here, two quantities characterise the outcome. Both are taken relative to the unmodified model, since the relations differ in how accurately G extracts them before any intervention. The *leakage* of the target,

$$\ell(\tau,\sigma) = \frac{\mathrm{EM}_{r^*}(G'(\tau,\sigma))}{\mathrm{EM}_{r^*}(G)}, \tag{3}$$

measures how much of the suppressed relation survives, and the *collateral damage* to the remainder,

$$c(\tau,\sigma) = 1 - \frac{1}{|\mathcal{R}|-1} \sum_{r \neq r^*} \frac{\mathrm{EM}_r(G'(\tau,\sigma))}{\mathrm{EM}_r(G)}, \tag{4}$$

measures how much of everything else is lost.

The above two pull in opposite directions, since lower leakage is bought at higher collateral damage, and (τ, σ) governs both jointly rather than one each. No threshold on either quantity can be justified independently of the deployment it serves. This work therefore reports ℓ and c directly at the selected operating points rather than certifying configurations against a fixed criterion.

The two parameters (τ, σ) cannot be set in sequence. A large set admitted by a low τ contains neurons that also serve other relations, so the intensity that disrupts the target disturbs them as well. A small set admitted by a high τ is cleaner but may cover only part of the target circuit, requiring a larger σ on what remains. Which value of one is appropriate depends on the other. This work solves this problem in three stages. First, it proposes QCAPS to identify the neurons that route a request for the target relation. Second, the Neuron Forcer suppresses them at inference time. Third, a joint search calibrates the selection threshold and the suppression intensity together.

3.2 Question-Contrast Average Precision Scoring

A neuron that routes a request fires when its relation is asked for and not otherwise, whereas a neuron that responds to the relation being present in the document fires whatever is asked of that document. Existing work contrasts contexts where the target answer is present against contexts where it is absent (Wu et al., 2023; Chen et al., 2024; Liu et al., 2025a), which does not separate the two. Separating them requires holding the document fixed and varying only the request.

For a target relation r^* and context $\mathcal{C}$, QCAPS forms one labelled prompt per relation,

$$\mathcal{P}_{r^*}(\mathcal{C}) = \big\{(\mathcal{C},\, q_{r^*},\, 1)\big\} \cup \big\{(\mathcal{C},\, q_r,\, 0) : r \neq r^*\big\}, \tag{5}$$

so the context is identical across the set and only the question varies. Activations are collected at the `act_fn` hook of each MLP and mean-pooled over question tokens, and average precision is computed per neuron,

$$\mathrm{AP}(l, n, r^*) = \mathrm{AveragePrecision}\left(\mathbf{a}_{l,n},\ \mathbf{y}\right), \tag{6}$$

where $\mathbf{a}_{l,n}$ holds the activations across all prompts and $\mathbf{y}$ the labels. Average precision is rank-based, so it is invariant to the activation scale, which varies by more than an order of magnitude across layers. It also assigns a presence-sensitive neuron, which takes similar values under all five questions, a score at the positive class rate regardless of how large its activations are. Appendix B states a third property and compares average precision against a difference-of-means statistic.

A neuron is relation-selective for r^* when $\mathrm{AP}(l, n, r^*) > \mathrm{AP}(l, n, r)$ for all $r \neq r^*$. Cross-relation dominance filtering retains these,

$$\mathcal{N}_{r^*} = \left\{ (l, n) \mid \mathrm{AP}(l, n, r^*) \geq \max_{r \neq r^*} \mathrm{AP}(l, n, r) \right\}, \tag{7}$$

and the score of Equation 1 is taken to be $s(l, n, r^*) = \mathrm{AP}(l, n, r^*)$ on this set. A neuron can be the argmax for at most one relation, so the sets $\mathcal{N}_r$ are disjoint by construction.

Neuron disjointness follows from the filter and does not establish that the selected neurons route the relation. It guarantees only that no neuron is suppressed on behalf of two relations at once. The filter decides which relation a neuron belongs to, and how many are selected is set by the threshold τ rather than by a fixed count K, because τ is comparable across relations and models and imposes no per-layer budget, so the selected set follows the depth profile of the circuit. Both steps require only forward passes, a one-off cost before deployment, whereas integrated-gradient methods require multiple backward passes per example (Appendix B.4).

3.3 Neuron Forcer

The neuron intervention is applied through a forward hook at the `act_fn` of each layer's MLP, overwriting the activation of every selected neuron at question-token positions and leaving the context representation untouched. Figure 1 illustrates the mechanism. The forcing map of Equation 2 is

$$\phi_\sigma(a_{l,n}) = \mu^{\text{off}}_{l,n} - (\sigma - 1) \cdot \underbrace{\left(\mu^{\text{on}}_{l,n} - \mu^{\text{off}}_{l,n}\right)}_{\text{activation gap } g_{l,n}}, \tag{8}$$

where $\mu^{\text{off}}_{l,n}$ is the mean activation under contrast prompts, $\mu^{\text{on}}_{l,n}$ the median under target prompts, and $\sigma > 1$. Only neurons with a positive gap are suppressed. At $\sigma = 1$ the neuron is set to its off-mean, and larger values drive it progressively further below the off-distribution. Appendix B justifies the choice of statistics and Appendix E.3 the displacement rule.

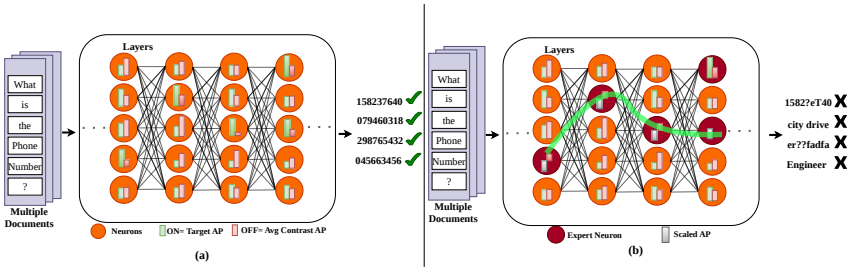

Figure 1: Neuron selection and suppression. Without the hook (a) the model extracts the target relation. With the hook (b) the selected neurons are forced below their off-state at question-token positions and extraction fails. Not all high-scoring neurons lie on the extraction pathway.

Scaling the displacement by each neuron's own gap concentrates the intervention on the neurons carrying the most relation-specific signal and leaves marginal ones barely moved. Because the hook writes to activations and never to parameters, detaching it restores the original model exactly, which is why the mechanism is framed as access control rather than erasure.

3.4 Calibrating the Operating Point

Neither leakage nor collateral damage is differentiable in (τ, σ), since both are obtained by running the model, so the operating point cannot be found in closed form and is instead located by Bayesian optimization, using a Tree-structured Parzen Estimator (Bergstra et al., 2011), which needs far fewer evaluations than a grid. Evaluation is deterministic under greedy decoding, so a configuration returns the same objective whenever it is repeated. Each trial applies a candidate (τ, σ), measures both quantities on a held-out search subset, and returns

$$\mathcal{L}(\tau, \sigma) = \ell(\tau, \sigma) + \lambda \cdot \max\big(0,\ c(\tau, \sigma) - \rho\big) + \mathbf{1}[c(\tau, \sigma) > \epsilon]\,. \tag{9}$$

While the first term drives suppression, the second penalises collateral damage above a floor ρ with weight λ, which keeps the search away from settings that suppress the target by degrading the model as a whole, and the third adds a fixed penalty once collateral damage passes a collapse threshold ϵ, so that a configuration destroying every relation is not rewarded for having also destroyed the target. Since ρ and λ shape only the search, their values are given with the trial counts in Appendix D.

4 Experiments and Results

4.1 Experimental Setup

The experimental setup comprises two extraction models, **LLaMA-3 8B** and **Qwen2.5-7B**, fine-tuned with an equivalent QLoRA configuration on a publicly available **PII corpus** (Paullier, 2024) of 4,434 biographical texts annotated with **five relation types**: full name, email address, phone number, job title and home address. Neurons are identified on the training split, while suppression is evaluated on the held-out test split by **exact match**, from which **leakage** and **collateral damage** are computed at 500 examples per relation with bootstrap confidence intervals. (As both are ratios to the clean model, leakage can slightly exceed 100% for a relation the intervention leaves unaffected.)

The proposed method is compared against Adaptive RMU (Dang et al., 2025), a parameter-level unlearning **baseline**, and against five substitutions that each replace one of its components with the counterpart used in prior work at the same neuron budget. Selection is replaced by a layer-matched random set or by answer-presence scoring (Wu et al., 2023; Liu et al., 2025a), and the intervention by zero ablation (Wu et al., 2023), activation patching (Wu et al., 2024) or polarity-aware attenuation (Zhengyuan et al., 2026). Training details, clean accuracies and the prompt template are given in Appendix A.

Methods that retrain on scrubbed data, including differential privacy, are not compared, since they change the parameters this setting holds fixed. PrivacyScalpel (Frikha et al., 2025) is excluded because it trains sparse autoencoders over each layer, the cost that QCAPS is designed to avoid.

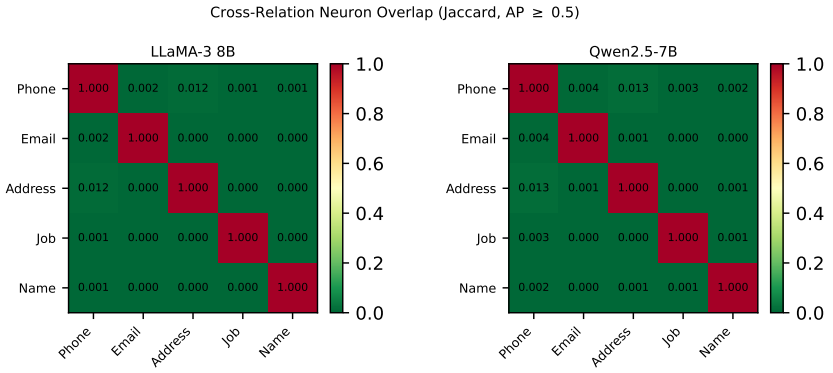

Figure 2: Cross-relation neuron overlap on LLaMA-3 8B (left) and Qwen2.5-7B (right). Off-diagonal values stay near zero on both models.

DEPN and LDEDE are represented by their interventions alone, since their identification stages score a different contrast and would confound the comparison of selection with that of intervention.

4.2 Experimental Results

4.2.1 QCAPS Identifies Routing Neurons

Whether a neuron routes a request or responds to the document cannot be read off its activation alone, but it can be read off a contrast. The two contrasts select different populations. Under answer-presence scoring the maximum pairwise Jaccard among the top-K sets at $K = 13{,}000$ is 0.213 on LLaMA-3 8B and 0.250 on Qwen2.5-7B, against 0.003 and 0.001 under QCAPS. Scores across relations are positively correlated under answer presence, at $+0.43$ and $+0.38$, and negatively correlated under QCAPS, at -0.17 and -0.16, so a neuron scoring highly for one relation under answer presence tends to score highly for the others, which is what a unit responding to presence rather than to request would do. Both measurements are taken with the dominance filter of Equation 7 applied to neither side, since that filter assigns each neuron to at most one relation and would make the sets disjoint by construction. Figure 2 shows the filtered sets at the thresholds actually used, and Appendix F.1 reports the same measurement across four thresholds.

Selectivity does not by itself establish that the selected neurons route a request, since they could instead hold the attribute values. Neither model answers any relation when the question is asked without its document, and swapping in another individual's document leaves accuracy at chance, at 2.0% or below. The models extract rather than recall, so what QCAPS retains is a pathway rather than a stored value. Appendix A.4 reports both controls, and Appendix C the layer profile of the identified sets.

4.2.2 Calibrated Forcing Suppresses Selectively

The neurons identified in Section 4.2.1 are now suppressed at their calibrated operating points. Table 1 gives the outcome for each relation on each model, against Adaptive RMU (Dang et al., 2025), a gradient-based unlearning method applied to the same relations under the calibration of Appendix H. QCAPS with the Neuron Forcer suppresses the target to between 0.6% and 19.2% of baseline, under 10% on eight of the ten pairs. The cost to the remaining relations separates into three groups on LLaMA-3 8B. Name and job are suppressed at negligible cost, at 0.4% and 0.3% collateral damage, email and phone cost 6.9% and 11.3%, and address costs 39.1%, with bootstrap intervals that do not overlap between the groups. The ordering does not carry across. On Qwen2.5-7B the same five costs run from -0.1% to 10.4% without a comparable separation, and address is among the cheapest at 1.2% rather than the most expensive, under the same procedure on the same data.

Adaptive RMU reaches lower leakage on eight of the ten pairs, between 0.0% and 1.4%, at collateral damage between 20.1% and 33.2%. On the remaining two it does not, leaving name on LLaMA-3 8B at 0.8% against 0.6% and failing to suppress email on that model at all, at 67.4%. The trade is

Table 1: Selected operating points against parameter-level unlearning. Leakage is the percentage of baseline target-relation EM that survives, collateral damage the mean loss over the four non-target relations, both at $n = 500$ per relation. Brackets give 95% bootstrap intervals. K counts the neurons the intervention modifies.

					QCAPS + Forcer		Adaptive RMU	
Model	**Relation**	τ	σ	K	Leakage ↓	Collateral ↓	Leakage ↓	Collateral ↓
LLaMA-3 8B	Name	0.87	14.47	2303	0.6 [0.0, 1.4]	**0.4** [0.0, 0.9]	0.8	22.6
	Job	0.94	18.36	2293	1.0 [0.2, 2.0]	**0.3** [0.1, 0.5]	0.2	24.4
	Email	0.86	11.68	1846	3.6 [2.0, 5.4]	**6.9** [5.9, 8.0]	67.4	23.4
	Phone	0.80	19.40	6442	2.2 [1.0, 3.6]	**11.3** [10.1, 12.5]	0.0	33.2
	Address	0.62	34.26	6377	9.6 [7.0, 12.2]	39.1 [37.6, 40.5]	0.0	**24.2**
Qwen2.5-7B	Name	0.60	35	9506	13.0 [10.0, 16.0]	**3.7** [2.8, 4.7]	1.4	24.0
	Job	0.98	25	2682	3.2 [1.7, 5.0]	−0.1 [−0.2, 0.0]	1.1	21.8
	Email	0.78	210	1052	19.2 [15.8, 22.8]	**10.4** [8.9, 12.0]	0.2	22.0
	Phone	0.905	18.4	1879	8.6 [6.2, 11.2]	**5.5** [4.5, 6.6]	0.0	29.4
	Address	0.81	117	1779	0.6 [0.0, 1.4]	**1.2** [0.4, 2.0]	0.2	20.1

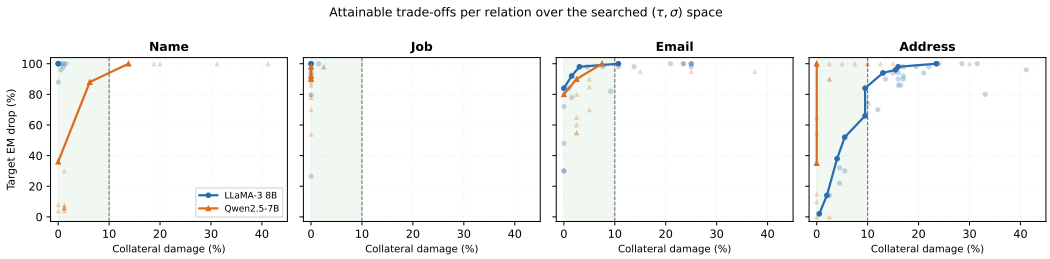

Figure 3: Attainable trade-offs over the searched (τ, σ) space. Each point is one search trial, and lines trace the lower envelope. The shaded band marks 10% collateral damage.

therefore not that one method suppresses better than the other. Parameter-level unlearning buys a few points of additional suppression at the cost of a quarter to a third of every other relation, and does so permanently, whereas the forcing hook is detached to restore the original model exactly. Figure 3 shows where these points come from. Each relation admits a range of configurations rather than a single one, and the shape of that range differs by relation and by model. Name and job reach full suppression at almost no cost on at least one model, while address on LLaMA-3 8B rises gradually and reaches only 66% suppression within a 10% budget. Appendix D discusses the trials in full.

The operating points hold under rephrasing and preserve the capability a deployment depends on. The statistics of Equation 8 are estimated under the fine-tuning template and the hook fires from the question marker, so a rephrased query could miss the span the intervention was calibrated on. Each operating point was re-evaluated under four wordings per relation, with the unsuppressed accuracy measured again under each. Leakage stays between 0.0% and 7.5% on LLaMA-3 8B and between 0.0% and 18.6% on Qwen2.5-7B, against an unsuppressed accuracy of 85.5% to 100.0% throughout, so the intervention weakens on one model without being defeated on either. These are natural rephrasings rather than attacks constructed against the mechanism. The cost of suppression falls elsewhere. Extraction on the non-target relations, the capability a deployment of this kind depends on, stays between 88.7% and 100.1% of baseline on average, except for address on LLaMA-3 8B. MMLU falls by 3.6 to 8.2 points on five of the ten pairs, significant under a paired test over the same 500 items, does not change significantly on four, and rises by 1.6 points on job for Qwen2.5-7B. WikiText-2 perplexity rises by a factor between 1.06 and 2.58 on LLaMA-3 8B and between 1.05 and 2.85 on Qwen2.5-7B, except for email on Qwen2.5-7B, where the calibrated intensity of $\sigma = 210$ raises it by a factor of 33.7. That configuration is the weakest in the paper on both counts, since it also carries the highest leakage, at 19.2%. The benchmarks carry no question marker, so the intervention is applied at every token position for these measurements, which bounds the cost from above rather than reporting the deployed setting. Per-wording figures are in Appendix F.3, the per-relation breakdown in Appendix G.1, and both benchmarks with the paired test in Appendix G.2.

Table 2: Leakage under one substituted component at a time, at each relation's operating point. Arms marked † collapse the model rather than suppressing the target. Full arms with collateral damage are in Appendix E.1 and seed counts in Appendix F.4.

		LLaMA-3 8B					Qwen2.5-7B				
Selection	**Intervention**	Name	Job	Email	Phone	Addr.	Name	Job	Email	Phone	Addr.
Intervention fixed, selection varied											
Random	Forcer	99.9	100.2	99.5	99.6	99.8	99.0	100.0	99.9	100.0	98.9
Presence	Forcer	100.0	99.6	99.4	100.0	100.0	99.6	99.8	14.8†	100.0	69.6†
Selection fixed, intervention varied											
QCAPS	Zero, DEPN	99.4	98.8	99.0	96.8	100.0	98.0	46.9	99.8	100.0	99.4
QCAPS	Patching, APNEAP	99.8	95.7	99.0	99.2	99.6	99.4	35.1	100.0	100.0	100.0
QCAPS	Polarity, LDEDE	99.6	100.6	99.8	99.8	100.0	99.8	98.3	99.2	100.0	98.8
QCAPS	**Forcer**	**0.6**	**1.0**	**3.6**	**2.2**	**9.6**	**13.0**	**3.2**	**19.2**	**8.6**	**0.6**

4.2.3 Selection and Intensity Are Both Necessary

Table 2 varies one component at a time. In the upper block the intervention is held at the calibrated forcing rule and the selection changes. In the lower block the selection is held at the QCAPS operating point and the intervention changes. Every arm uses the same σ and, where the pool of candidate units allows, the same neuron budget as the configuration it is compared against. Neither substitution suppresses the target relation. A layer-matched random set under the calibrated intervention leaves leakage between 98.9% and 100.2%, and the answer-presence selection leaves it between 99.4% and 100.0% wherever the intensity transfers. In the other direction, zero ablation as used by DEPN leaves 96.8% to 100.0%, activation patching as used by APNEAP leaves 95.7% to 100.0%, and polarity-aware attenuation as used by LDEDE leaves 98.3% to 100.6%, in every case apart from job on Qwen2.5-7B, discussed below. The proposed combination leaves 0.6% to 19.2% on the same budgets. The interventions fail for a reason their targets make visible. Activation patching returns each neuron to approximately the value it takes when the target relation is not requested, which is $\sigma = 1$ under Equation 8. On LLaMA-3 8B the patch targets average -0.044 to -0.094 against off-state means of -0.035 to -0.079, while the calibrated targets lie between -0.375 and -0.871. Returning a neuron to its off-state is not enough. It has to be driven past it.

Polarity-aware attenuation fails differently. Its partition needs each neuron's relevance on the remaining relations, and under this identification all but four of the 57,256 selected neurons across both models carry a negative gap on at least one other relation. Almost nothing is classified as key, so the rule attenuates nearly the whole set rather than separating it, and halving a weight is far short of the displacement the forcing rule applies. Two arms depart from the pattern. On Qwen2.5-7B, zero ablation and activation patching both partially suppress job, at 46.9% and 35.1%, while leaving 95.7% to 100.0% everywhere else. The same relation on LLaMA-3 8B, at a comparably high selection threshold, leaves 98.8% under zero ablation. What produces this is not established here. On Qwen2.5-7B email and address, the calibrated intensity does not transfer to the answer-presence selection, whose gaps are roughly twice as large, and rescaling to match the mean displacement moves the figures without separating them. Both configurations damage every relation at once, at 32.3% and 89.9% collateral. Appendix E.1 reports the full arms with collateral damage, Appendix F.4 the seed variation, and Appendix E.2 the partition that the polarity-aware rule produces.

4.2.4 Interference Does Not Arise from Shared Neurons

Suppressing a relation damages some others and leaves the rest untouched. Table 3 reports the loss each suppressed relation inflicts on each remaining one. Of the twenty ordered pairs on each model, most fall below 2%, and the collateral damage reported in Table 1 concentrates in a few.

Coupling is directed. Suppressing email costs job 27.3% on LLaMA-3 8B and 13.4% on Qwen2.5-7B, while suppressing job costs email 0.6% and -0.2%. Suppressing phone costs address 42.2% and 20.6%, and on Qwen2.5-7B the reverse costs phone 0.2%. Two pairs run both ways: phone and address on LLaMA-3 8B, where the reverse costs 91.2%, the largest entry in the table, and email and

Table 3: Directed cross-relation damage. Each row gives the EM loss inflicted on every other relation when the row relation is suppressed, relative to that relation's clean accuracy, so each row averages to the collateral damage of Table 1.

	LLaMA-3 8B					Qwen2.5-7B				
Suppressed	Phone	Email	Addr.	Job	Name	Phone	Email	Addr.	Job	Name
Phone	—	1.6	**42.2**	0.2	1.2	—	0.0	**20.6**	1.3	0.2
Email	0.2	—	0.0	**27.3**	0.2	0.0	—	8.6	13.4	**19.6**
Address	**91.2**	7.6	—	2.4	**55.0**	0.2	0.2	—	4.5	−0.2
Job	0.0	0.6	0.6	—	0.0	0.0	−0.2	0.0	—	−0.2
Name	0.2	0.6	0.2	0.6	—	0.0	**12.4**	0.4	2.2	—

name on Qwen2.5-7B, at 19.6% against 12.4%. Four of the six large entries run one way and two run both.

The same pairs recur across architectures and overlap does not account for the pattern. Phone to address and email to job are among the four largest entries on both models, at 42.2% and 27.3% on LLaMA-3 8B and 20.6% and 13.4% on Qwen2.5-7B. Both models were fine-tuned on the same corpus under the same objective, which is consistent with the dependence arising from the extraction task, though two models trained on one templated source are not independent evidence. Address to phone is the largest entry on LLaMA-3 8B at 91.2% and absent on Qwen2.5-7B at 0.2%, and accounts for why address admits no usable operating point on LLaMA-3 8B while remaining the cheapest relation on Qwen2.5-7B. Appendix G.1 gives the per-relation breakdown. The identified sets are close to disjoint for every pair on both models, with a maximum pairwise Jaccard of 0.012 and 0.013, falling to 0.0001 or below at the thresholds used for four of the five operating points. This holds equally for the coupled pairs and the uncoupled ones. Phone and address overlap no more than job and name do, yet the first pair transfers tens of percentage points between its members and the second transfers none. Overlap is also symmetric in the two relations while the damage is not, so no measure of shared neurons can reproduce a matrix in which one direction costs tens of percentage points and the reverse costs none. What separates a coupled pair from an uncoupled one is therefore not the neurons they hold in common, which is what the coupled-neuron account of Zhengyuan et al. (2026) would predict, and which the degenerate partition of Section 4.2.3 makes explicit. The intervention is confined to the question span and each relation is queried in a separate forward pass with no shared state, so the dependence is carried by the downstream computation those neurons feed. Where exactly is not measured here.

5 Conclusion

Selective suppression of a relation in a fine-tuned LLM depends on both the precision of neuron localization and the strength of the intervention. This work proposed QCAPS, which identifies the neurons responsible for routing a request rather than the knowledge it retrieves. These neurons are then suppressed by a calibrated forcing intervention over the question tokens, and the two are balanced to suppress the target while largely preserving the remaining relations. Replacing either component with its counterpart from prior work leaves the target unsuppressed. However, the cost of suppression varies across relations and architectures, and interference persists between neuron sets that barely overlap. What determines the suppressibility of a relation therefore remains open. Suppression acts at the relation level rather than the entity level, so suppressing a relation removes the capability for every individual. Both models are also fine-tuned on a single synthetic corpus whose relations are extracted at or near 100% before suppression, and rewording is evaluated with natural rephrasings rather than with prompting constructed against the intervention, so neither transfer to naturally occurring documents nor robustness to an informed adversary is established. As the intervention acts on activations rather than parameters, it governs access to a relation rather than erasing it.

AI USE STATEMENT

In this work, generative AI tools were used for language editing of the manuscript and for assistance in writing and debugging experiment code. They were not used to generate research ideas, to design the method, or to produce any reported result, and all remaining disclosable uses are not applicable to this work. We have reviewed all AI-assisted work. Every reported number was computed from our own experimental outputs, and all AI-assisted code was run and its output verified by the authors. We take responsibility for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

All personal information used in this work is synthetic. The corpus is generated by a language model and contains no data about real individuals, so no human subjects were involved and no personal data was processed.

The method withholds a relation at inference time rather than removing it from the model. Detaching the hook restores the original behaviour exactly, and the information remains in the document context and in the model parameters. It should therefore not be presented as deletion for the purpose of meeting a right to erasure or a similar obligation. It is suited to restricting access to a relation within a deployment the operator controls, and offers no protection against a party with access to the weights.

Identifying the neurons that route a request for a relation could in principle inform attempts to strengthen extraction as well as to suppress it. The neurons located here are those an extraction model already uses, and the evaluation confines itself to suppression.

REPRODUCIBILITY STATEMENT

Section 3 specifies the identification procedure, the forcing rule and the search objective. Section 4.1 and Appendix A give the models, the fine-tuning configuration, the corpus, the splits, the prompt template and the baseline accuracies. Appendix D lists the search parameters, Appendix F.4 the seed counts for the random arms, and Appendix H the calibration of the Adaptive RMU baseline. Every reported configuration is listed with its threshold, intensity and neuron count in Table 1. Code for identification, suppression, search and evaluation will be released upon acceptance.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. Digital forgetting in large language models: a survey of unlearning methods: A. blanco-justicia et al. *Artificial Intelligence Review*, 58(3):90, 2025.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. URL https://api.semanticscholar.org/CorpusID:5945696.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin

Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. Learnable privacy neurons localization in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 256–264, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.25. URL https://aclanthology.org/2024.acl-short.25/.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

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

Huu-Tien Dang, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23733–23742, 2025.

Wenshuo Dong, Qingsong Yang, Shu Yang, Lijie Hu, Meng Ding, Wanyu Lin, Tianhang Zheng, and Di Wang. Understanding and mitigating cross-lingual privacy leakage via language-specific and universal privacy neurons. *ArXiv*, abs/2506.00759, 2025. URL https://api.semanticscholar.org/CorpusID:279075085.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. PrivacyScalpel: Enhancing LLM privacy via interpretable feature intervention with sparse autoencoders. In Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti (eds.), *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 226–238, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-346-3. doi: 10.18653/v1/2025.blackboxnlp-1.13. URL https://aclanthology.org/2025.blackboxnlp-1.13/.

Arpit Garg, Hemanth Saratchandran, Ravi Garg, and Simon Lucey. Stable forgetting: Bounded parameter-efficient unlearning in llms. 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16801–16819, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing, 2023. *URL https://arxiv.org/abs/2305.01610*, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. Cross-modal unlearning via influential neuron path editing in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35589–35597, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=e5icsXBD8Q.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schütze. On relation-specific neurons in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 992–1022, 2025a.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933, 2025b.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE, 2023.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Red Teaming GenAI: What Can We Learn from Adversaries?*, 2025. URL https://openreview.net/forum?id=P8seBluN3c.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Alejo Paullier. Pii external dataset. https://www.kaggle.com/datasets/alejopaullier/pii-external-dataset, 2024. Apache 2.0 License.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2024. URL https://openreview.net/forum?id=5LhYYajlqV.

Zihan Qiu, Zeyu Huang, Youcheng Huang, and Jie Fu. Empirical study on updating key-value memories in transformer feed-forward layers. *arXiv preprint arXiv:2402.12233*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1288–1301, 2024.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174/.

Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5319–5332, 2024. doi: 10.18653/v1/2024.findings-acl.315.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=8Dy42ThoNe.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602, 2024b.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

Zhao Zhengyuan, Cao Lifeng, Sunhaodong, Shi Haotian, Du Xuehui, Liu Aodi, Niu Lanjie, and Yang Xiaocheng. LDEDE: LRP-driven efficient detection and editing framework for LLM privacy neurons. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33315–33329, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1666. URL https://aclanthology.org/2026.findings-acl.1666/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX

A SETUP DETAILS

A.1 ARCHITECTURES

Both models are fine-tuned on the same corpus under identical splits and an equivalent QLoRA configuration (4-bit, $r = 16$, two epochs). The neuron space searched by QCAPS is the FFN intermediate dimension of every transformer layer, given in Table 4.

Table 4: Architectural parameters of the two fine-tuned models.

Component	**LLaMA-3 8B**	**Qwen2.5-7B**
Architecture	LlamaForCausalLM	Qwen2ForCausalLM
Transformer layers	32	28
Hidden size	4,096	3,584
FFN intermediate size	14,336	18,944
Attention heads	32	28
KV heads	8	4
Vocabulary size	128,256	152,064
Activation function	SiLU	SiLU
Normalization	RMSNorm	RMSNorm
Position encoding	Rotary	Rotary
Candidate neurons	458,752	530,432

A.2 BASELINE ACCURACY

Table 5 gives the exact-match accuracy of each clean model on the held-out split at $n = 500$ per relation. Every leakage and collateral figure in the paper is expressed relative to these values.

Table 5: Clean exact-match accuracy (%) at $n = 500$ per relation.

Model	Phone	Email	Address	Job	Name
LLaMA-3 8B	100.0	100.0	100.0	98.0	100.0
Qwen2.5-7B	99.8	99.8	100.0	92.8	99.8

A.3 PROMPT TEMPLATE

Every example follows the template the models were fine-tuned on. The context holds the document, the question names the relation, and the instruction constrains the output to a single copied span.

```
### Context:
⟨document⟩
### Question:
What is the phone number?
### Instruction:
Copy ONLY the requested value EXACTLY as it appears
in the context.  Return a single line only.  If the
value is not present, output N/A.
### Answer:
```

The hook of Equation 2 fires from the token following the `### Question:` marker, so the document representation is never modified.

A.4 Extraction Rather Than Memorisation

If a model recalled the attributes from its weights rather than reading them from the document, the neurons QCAPS retains could be holding the answers. Two controls rule this out. Asking each question with the document removed yields 0.0% exact match on all five relations on both models. Substituting another individual's document yields 0.0% on four relations, and 2.0% on LLaMA-3 8B and 1.0% on Qwen2.5-7B for job, which is chance given the small vocabulary of job titles. Both models extract rather than recall.

B Design Choices

B.1 Average Precision Against a Difference of Means

Average precision measures how well a neuron's activation ranks target prompts above contrast prompts, and is preferred to a difference-of-means statistic for three reasons. It is rank-based and therefore invariant to activation scale, which varies by more than an order of magnitude across layers and would otherwise dominate any absolute comparison. It requires no decision threshold, so none need be fixed before the neurons are known. And it assigns a presence-sensitive neuron, which takes similar values under all five questions, a score at the positive class rate regardless of the magnitude of its activations, so such neurons are separated by the score itself rather than by a later filter.

B.2 Mean Off-State and Median On-State

The two statistics in Equation 8 play different roles, which is why they are computed differently. The off-state is the value the neuron is returned toward, so it should reflect the centre of the distribution the neuron occupies when the target relation is not requested, for which the mean is the natural summary. The on-state serves only to measure how far the neuron travels when the target relation is requested. That distribution is right-skewed, since a routing neuron fires strongly on the subset of prompts where the relation is present in the context and weakly elsewhere, so its mean is pulled by the tail. The median gives a more stable estimate of the typical on-state and hence a gap that reflects habitual rather than extreme behaviour.

B.3 Notation

Table 6: Symbols used in Section 3.

Symbol	Meaning
G, $G'(\tau, \sigma)$	clean model, model under the intervention
$\mathcal{R}$, r^*	relation set, target relation
$\mathcal{C}$, q_r	context, question for relation r
$a_{l,n,t}$	activation of neuron n in layer l at position t
$s(l, n, r^*)$	discriminability score, taken to be average precision
τ, $K(\tau)$	selection threshold, resulting neuron count
σ, ϕ_σ	suppression intensity, forcing map
μ^{off}, μ^{on}	off-state mean, on-state median
$g_{l,n}$	activation gap, $\mu^{\text{on}} - \mu^{\text{off}}$
ℓ, c	leakage, collateral damage
ρ, λ, ϵ	search floor, penalty weight, collapse threshold

B.4 Complexity

Let $|\mathcal{D}|$ be the dataset size, L the number of layers, D the FFN intermediate size, K the number of selected neurons and T the generation length. Identification collects activations in one forward pass per prompt, at $O(|\mathcal{D}| \cdot L \cdot D)$, and computes average precision by sorting per neuron, at $O(L \cdot D \cdot |\mathcal{D}| \log |\mathcal{D}|)$. Because the contrast varies only the question, a single pass over the corpus with $|\mathcal{R}|$ questions per context serves every relation at once. No backward pass is required. At

inference the hook applies one activation override per selected neuron per position, at $O(K \cdot T)$, and $K \ll L \cdot D$.

C IDENTIFICATION

C.1 LAYER DISTRIBUTION

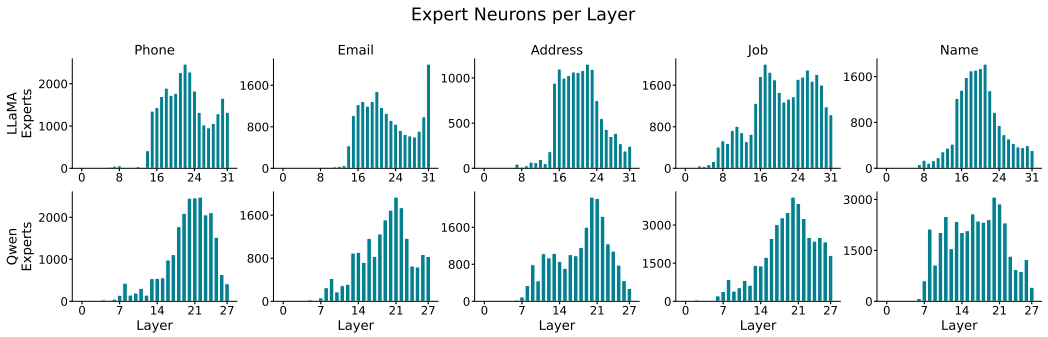

Figure 4: Distribution of QCAPS-identified neurons across layers.

Figure 4 gives the depth profile of the identified sets. Early layers hold few relation-specific neurons and act as relation-agnostic context processing. Counts rise sharply at the mid-network transition and peak in the upper-middle layers. The pattern holds on both architectures. Within the shared band each relation occupies a distinct profile that is also preserved across models. Phone and job circuits are the largest and concentrate in sharp mid-network peaks, address spreads diffusely across the later layers without a dominant peak, name concentrates in the upper-middle layers, and email rises toward the final layer.

D CALIBRATION

D.1 SEARCH CONFIGURATION

Table 7: Search configuration. The two models bound collateral damage in different ways, described in the text.

	LLaMA-3 8B	**Qwen2.5-7B**
Trials per relation	10 to 25	16 to 18
Examples per trial	50	20 to 50
Collateral constraint	linear penalty above $\rho = 0.13$	rejection above 0.05
Penalty weight λ	2.0	not needed under rejection
Collapse penalty	fixed, above $\epsilon = 0.70$	covered by rejection

The two searches impose the bound differently. On LLaMA-3 8B it enters through the linear penalty of Equation 9. On Qwen2.5-7B a configuration exceeding 5% collateral damage is rejected outright. Both affect only which configurations the optimiser proposes, since every reported configuration is re-evaluated on the held-out split under the same protocol. Cross-model observations drawn from the search trials rather than from the re-evaluated operating points are stated as hypotheses in Section 4.2.4.

D.2 SEARCH TRAJECTORIES

Figure 3 in Section 4 plots every trial for four relations on both models, with target EM against the collateral damage it cost. Three observations follow. The searched region is broad rather than a single point, and on LLaMA-3 8B the address trials span the full range from zero to 41% collateral damage

with several distinct configurations reaching comparable suppression at different costs. The shape of the trade-off differs by relation, with name and job sitting against the vertical axis on at least one model while address on LLaMA-3 8B rises gradually and reaches only 66% suppression within a 10% budget. And the same relation behaves differently across models, graded on LLaMA-3 8B and vertical on Qwen2.5-7B.

Two caveats apply. Trials use 50 examples per relation on LLaMA-3 8B and between 20 and 50 on Qwen2.5-7B, so collateral damage is measured at coarser resolution on Qwen2.5-7B and a trial reporting 0% may reflect no error in a small sample rather than no effect. The figures are search-time estimates, and the configurations selected from them were re-evaluated at $n = 500$. Phone is omitted because its Qwen2.5-7B operating point was selected on a grid.

D.3 Address on LLaMA-3 8B

Three configurations drawn from different regions of the search were evaluated at $n = 500$ for address on LLaMA-3 8B. The selected point gives 9.6% leakage at 39.1% collateral damage, and two more conservative candidates give 17.6% at 37.4% and 31.8% at 30.7%. None falls below 30%, and the search itself reached only 66% suppression within a 10% collateral budget over 40 trials. The cost is a property of the relation on that model rather than of the configuration chosen.

E Ablations

E.1 Ablation Arms in Full

Table 8 gives collateral damage for the arms of Table 2, which reports leakage only. No arm except the proposed combination costs more than 2.8% on either model, apart from the two answer-presence configurations on Qwen2.5-7B that collapse the model.

Table 8: Collateral damage (%) for the arms of Table 2.

		LLaMA-3 8B					**Qwen2.5-7B**				
Selection	**Intervention**	Name	Job	Email	Phone	Addr.	Name	Job	Email	Phone	Addr.
Random	Zero	0.1	0.2	0.0	0.1	0.2	−1.0	−1.0	0.0	−0.1	0.1
Random	Forcer	0.1	0.2	−0.0	0.1	0.1	−1.0	−0.8	1.0	−0.1	2.8
Presence	Forcer	0.2	0.2	0.1	0.1	0.7	0.4	−0.0	32.3	0.1	89.9
QCAPS	Zero, DEPN	0.1	0.2	0.0	0.2	0.2	−1.0	−0.5	−0.1	−0.1	0.1
QCAPS	Patching, APNEAP	0.0	0.2	0.0	0.2	0.2	−0.1	0.0	0.2	−0.1	0.2
QCAPS	Polarity, LDEDE	0.1	0.2	−0.0	0.0	0.1	0.3	0.4	2.3	0.2	0.4
QCAPS	**Forcer**	0.4	0.3	6.9	11.3	39.1	3.7	−0.1	10.4	5.5	1.2

E.2 Polarity-Aware Partition

Applying the rule of Zhengyuan et al. (2026) requires partitioning the selected neurons by the sign of their relevance on the remaining relations. Its LRP-based relevance is not reproduced here, and the activation gap of the question contrast is used instead, which is the quantity the forcing rule already scales by. The partition also needs each neuron's behaviour under every relation, which the dominance filter removes, so the full activation matrices were collected separately.

Under this criterion the partition does not separate the set on either model. Table 9 gives the split. Between zero and two neurons per relation are classified as key and the remainder as coupled, so the rule attenuates nearly the whole set rather than dividing it. Leakage stays between 98.3% and 100.6% at collateral damage of 2.3% or below.

The forced counts differ from the K of Table 1 because the full statistics were re-estimated on the evaluation split, where a different subset of the selected neurons has a positive gap.

Table 9: Polarity-aware partition at $\beta = 0.5$.

Model	Relation	Forced	Key	Coupled	Leakage (%)	Collateral (%)
LLaMA-3 8B	Name	2,058	0	2,058	99.6	0.1
	Job	2,238	0	2,238	100.6	0.2
	Email	1,741	1	1,740	99.8	0.0
	Phone	5,955	1	5,954	99.8	0.0
	Address	5,718	0	5,718	100.0	0.1
Qwen2.5-7B	Name	23,598	2	23,596	99.8	0.3
	Job	5,270	0	5,270	98.3	0.4
	Email	3,574	0	3,574	99.2	2.3
	Phone	2,355	0	2,355	100.0	0.2
	Address	4,749	0	4,749	98.8	0.4

E.3 Gap-Proportional Against Uniform Displacement

The forcing rule displaces each neuron by a multiple of its own gap. An alternative displaces every neuron by the same absolute amount, discarding the per-neuron scaling. The two were compared on LLaMA-3 8B for phone at a fixed set of $K = 7{,}000$, with the uniform displacement set to the mean displacement the gap-proportional rule produces at $\sigma = 16.5$, so that the total intervention magnitude matches and only its distribution across neurons differs.

The gap-proportional rule reduces target extraction to 22.0% at 10.0% collateral damage. The uniform rule at the same mean displacement leaves target extraction at 93.0%. Raising the uniform displacement to three times that magnitude reduces target extraction to 33.0% at 60.0% collateral damage, so the uniform rule does not reach the gap-proportional operating point at any scale tested.

F Sensitivity Analyses

F.1 Threshold Robustness

Table 10 reports cross-relation overlap at four thresholds. The mean and maximum off-diagonal Jaccard reach 0.000 on both models by AP ≥ 0.9, while the mean AP of the retained neurons rises steadily, so higher thresholds isolate cleaner discriminators rather than merely thinning the set. The threshold of 0.5 used in Figure 2 is conservative rather than critical.

Table 10: Threshold robustness of cross-relation exclusivity.

	LLaMA-3 8B				Qwen2.5-7B			
AP thr.	Avg #	Mean J	Max J	Mean AP	Avg #	Mean J	Max J	Mean AP
0.50	22,105	0.0016	0.0115	0.676	29,111	0.0026	0.0132	0.705
0.70	8,240	0.0000	0.0001	0.830	12,842	0.0001	0.0003	0.858
0.90	2,082	0.0000	0.0000	0.954	4,876	0.0000	0.0000	0.965
0.95	1,115	0.0000	0.0000	0.980	3,256	0.0000	0.0000	0.985

The thresholds selected for the operating points lie above 0.5 for four of the five relations, placing their sets in the regime where the maximum off-diagonal Jaccard is 0.0001 or below. The residual maximum, at either threshold and on either model, falls on the same pair, phone and address.

F.2 Paraphrase Invariance of Identification

Four paraphrased forms per relation were generated, ranging from minor lexical variation to wording that omits the relation keyword. Neurons were re-identified under each and compared against those identified from the original question. Recall of the original high-AP neurons rises monotonically with the threshold, from 47% to 56% at AP > 0.5 to 84% to 91% at AP > 0.95, so the neurons most strongly associated with a relation are the most stable across rephrasings. The most divergent paraphrase still recovers 67% at AP > 0.9. Same-relation paraphrase pairs sustain a symmetric

Jaccard of 0.17 to 0.32, while cross-relation pairs stay near 0.003 and layer-matched random baselines fall from 0.044 to 0.008 as the threshold tightens.

F.3 PARAPHRASING

Table 11 gives leakage under four wordings per relation, with the unsuppressed accuracy re-measured under each. P0 is the template wording and the remaining three progressively drop the relation keyword.

Table 11: Leakage under four wordings per relation, $n = 200$. Unsuppressed accuracy under each wording in parentheses.

Model	Relation	P0	P1	P2	P3
LLaMA-3 8B	Name	1.5 (100.0)	0.0 (95.0)	0.0 (97.5)	0.0 (100.0)
	Job	0.5 (97.5)	0.0 (96.5)	0.0 (95.5)	2.6 (98.0)
	Email	4.5 (99.5)	0.0 (99.5)	0.0 (86.5)	0.0 (98.5)
	Phone	2.0 (100.0)	3.0 (99.5)	0.5 (93.5)	2.5 (99.5)
	Address	7.5 (100.0)	1.5 (100.0)	3.5 (100.0)	4.5 (100.0)
Qwen2.5-7B	Name	13.1 (99.5)	3.5 (99.5)	6.0 (99.5)	18.6 (99.5)
	Job	3.8 (92.0)	1.2 (85.5)	10.7 (88.5)	10.9 (87.5)
	Email	17.0 (100.0)	2.5 (100.0)	0.0 (93.0)	5.6 (99.0)
	Phone	6.5 (100.0)	2.0 (100.0)	7.1 (98.5)	7.5 (100.0)
	Address	0.0 (100.0)	0.0 (98.5)	1.0 (99.5)	0.0 (99.5)

On LLaMA-3 8B leakage is lower under a paraphrase than under the template wording on seventeen of the twenty cells. On Qwen2.5-7B it rises above the operating-point value on name and job under two wordings each. The unsuppressed accuracy stays between 85.5% and 100.0% throughout, so the figures are not an artefact of the model losing the question.

F.4 SEED VARIATION

The random arms are means over seeds, and Table 12 gives the spread. Each draw matches the per-layer counts of the QCAPS set and is taken from the relation's own identified pool with the QCAPS units excluded, so no random set shares a neuron with the configuration it controls for. Every arm stays within a standard deviation of 0.5 except random with forcing on Qwen2.5-7B address, at 2.1 over four seeds, whose mean of 98.9% is still at baseline.

Table 12: Leakage across seeds for the random arms, $n = 500$. Qwen2.5-7B address with forcing uses four seeds, every other entry five.

		Random + Zero		Random + Forcer	
Model	Relation	Mean	SD	Mean	SD
LLaMA-3 8B	Name	100.0	0.0	99.9	0.1
	Job	100.2	0.3	100.2	0.5
	Email	99.4	0.2	99.5	0.3
	Phone	99.8	0.1	99.6	0.4
	Address	100.0	0.1	99.8	0.2
Qwen2.5-7B	Email	99.9	0.5	99.9	0.2
	Phone	100.0	0.0	100.0	0.1
	Address	99.7	0.2	98.9	2.1

The random arms for name and job on Qwen2.5-7B are evaluated at 50 examples over two seeds, where they give 98.0% to 100.0% leakage. On LLaMA-3 8B the pool for address holds fewer units outside the QCAPS set than the QCAPS set itself in some layers, so its random arm modifies 5,278 neurons against 6,377. A smaller set can only suppress less, so the shortfall does not favour the method.

The forcing rule modifies only neurons whose activation gap is positive. On Qwen2.5-7B this is a subset of those above the threshold, and it applies equally to the QCAPS and the random arms, leaving them at comparable size, 1,879 against 1,614 on phone and 1,052 against 1,004 on email. On LLaMA-3 8B every selected unit has a positive gap.

F.5 Intensity Transfer on the Answer-Presence Selection

The intensity at each operating point is calibrated against the gap distribution of the QCAPS selection. The answer-presence neurons have gaps roughly twice as large on Qwen2.5-7B email and address, at $+0.0220$ against $+0.0117$ and $+0.0268$ against $+0.0130$, so the same σ displaces them twice as far. At the unmodified σ the two configurations reach 96.1% and 98.9% collateral damage, and rescaling to match the mean displacement, to $\sigma = 111.8$ and $\sigma = 56.7$, brings them to 32.3% and 89.9%. Both remain collapses rather than selective suppression. The remaining eight configurations transfer without rescaling.

G Supporting Results

G.1 Per-Relation Accuracy at Each Operating Point

Table 13 gives the figures underlying the collateral damage column of Table 1 and the directed matrix of Table 3. Each row is one suppressed relation and each column the accuracy retained on every relation under that intervention.

Table 13: Per-relation EM accuracy (%) at each operating point, $n = 500$. Rows indicate which relation was suppressed. Diagonal entries in bold.

Suppressed	Phone	Email	Address	Job	Name	Coll.
LLaMA-3 8B						
Phone	**2.2**	98.4	57.8	97.8	98.8	11.3
Email	99.8	**3.6**	100.0	71.2	99.8	6.9
Address	8.8	92.4	**9.6**	95.6	45.0	39.1
Job	100.0	99.4	99.4	**1.0**	100.0	0.3
Name	99.8	99.4	99.8	97.4	**0.6**	0.4
Qwen2.5-7B						
Phone	**8.6**	99.8	79.4	91.6	99.6	5.5
Email	99.8	**19.2**	91.4	80.4	80.2	10.4
Address	99.6	99.6	**0.6**	88.6	100.0	1.2
Job	99.8	100.0	100.0	**3.0**	100.0	-0.1
Name	99.8	87.4	99.6	90.8	**13.0**	3.7

G.2 Capability Retention

Table 14 gives MMLU and WikiText-2 perplexity at each operating point. The benchmarks carry no question marker, so the intervention is applied at every token position rather than over the question span, which bounds the cost from above. McNemar is computed over the same 500 MMLU items answered by the clean and the suppressed model, with b the items the clean model answers correctly and the suppressed model does not, and c the reverse.

Five of the ten configurations cost MMLU accuracy significantly and four leave it unchanged, while job on Qwen2.5-7B gains 1.6 points, also significantly. The two measures move together on LLaMA-3 8B, where the configurations leaving MMLU unchanged also hold perplexity within $1.11\times$ and those costing accuracy raise it to $1.50\times$ or beyond. Address is the exception, with the largest perplexity ratio and a drop that does not reach significance, and the largest count of discordant items in either direction. On Qwen2.5-7B the email configuration stands apart, at $33.72\times$ perplexity against $2.85\times$ or below for the rest. Its intensity of $\sigma = 210$ is the largest in the paper, and the next largest, 117 for address, costs $1.38\times$.

Table 14: Capability under suppression. MMLU at $n = 500$ with a paired test, WikiText-2 perplexity as a ratio to the clean model. Qwen2.5-7B perplexity is measured on the bf16 model, where the clean value is 8.72, and every other figure on the 4-bit model. The clean row is the reference for every comparison.

Model	**Relation**	**MMLU**	**Δ**	b	c	p	**PPL**
LLaMA-3 8B	clean	60.6	—	—	—	—	1.00×
	Name	60.2	−0.4	25	23	0.885	1.11×
	Job	55.8	−4.8	39	15	0.002	1.55×
	Email	60.4	−0.2	9	8	1.000	1.06×
	Phone	57.0	−3.6	33	15	0.013	1.50×
	Address	57.4	−3.2	46	30	0.085	2.58×
Qwen2.5-7B	clean	72.4	—	—	—	—	1.00×
	Name	64.2	−8.2	67	26	< 0.001	1.67×
	Job	74.0	+1.6	2	10	0.039	1.05×
	Email	65.2	−7.2	65	29	< 0.001	33.72×
	Phone	71.6	−0.8	22	18	0.636	2.85×
	Address	65.4	−7.0	53	18	< 0.001	1.38×

H Adaptive RMU Calibration

Adaptive RMU minimises the distance between the forget-set activations at a chosen layer and a random steering vector scaled by a coefficient $c = \text{SCALE} \cdot \|a\|$, where $\|a\|$ is the mean activation norm of that layer on the forget set. A second term holds the retain-set activations at their pre-training values and is weighted by α. Both losses are mean squared errors against targets whose magnitude follows $\|a\|$, so their ratio depends on the activation scale of the model being unlearned while α does not.

That scale differs by three orders of magnitude between the two models. The steering coefficient is 1.6×10^4 on Qwen2.5-7B against 34 on LLaMA-3 8B at the same layer and the same SCALE. With $\alpha = 1200$ the unlearn term on Qwen2.5-7B sits at 9.4×10^4 against a retain term of 59 after one hundred steps, a ratio of roughly 1600. The same α on LLaMA-3 8B reverses the ordering. The unlearn loss begins at 0.42 and the weighted retain term exceeds it, so the unlearn loss does not fall. Over twenty steps it moves by 0.0% for name, $+0.9\%$ for email and $+2.8\%$ for address. Only phone trains, and only because its unlearn loss begins at 0.99.

Two checks confirm that this is a scale mismatch rather than a failure of the implementation. Removing the retain term entirely reduces the unlearn loss by 60.1% within twenty steps on name, so the gradient path is intact. And reducing α from 1200 through 300 to 100 changes the twenty-step movement by under one percentage point, because the retain term already dominates at $\alpha = 100$.

α was therefore rescaled for LLaMA-3 8B to reproduce the retain-to-unlearn ratio that $\alpha = 1200$ produces on Qwen2.5-7B. A value of $\alpha = 10$ gives a ratio of 1.6×10^3 and is used for all five relations. Every other hyperparameter is identical across the two models: layer 7, learning rate 5×10^{-5}, SCALE $= 5.0$, batch size 4, and the `down_proj` weights of layers 5 through 7 as the trainable parameters. The method leaves email on LLaMA-3 8B at 67.4% of baseline, which is the only configuration in Table 1 where it fails to suppress.