A CONDITIONAL JOINT MATCHING OF RESPONSE PROFILES FOR LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Honouring a deletion request means making a language model behave like one retrained without the deleted facts, but retraining per request is too costly. Reference-based methods instead move each forgotten fact's likelihood to the level of never-seen facts, reducing the fact to one number. Yet memorisation also decouples a fact's probes, the phrasings that test it: on the TOFU benchmark, the correlation between direct and paraphrased likelihoods falls from 0.91 on never-seen facts to 0.15 on memorised ones, a drop that no order-preserving per-probe target can undo. Our intuition is that a forgotten fact should resemble comparable never-seen facts on all its probes at once. We propose conditional joint matching: each fact becomes a response profile of log-likelihoods under six probes, and entropic optimal transport couples each forgotten profile with never-seen profiles of similar relation, length and difficulty to build that profile's target. Unlike per-probe targets, these whole-profile targets carry the never-seen correlation, and training stops by a rule that needs no retrained model. On TOFU at 1B parameters, the method has the lowest mean per-fact distance to the retrained model among ten target assignments and four published objectives, 0.785 nats against 0.906 for unconditional per-probe matching, but only that comparison survives correction for multiple testing. Shuffling reference probes across facts within relation and difficulty strata gives a statistically indistinguishable 0.797, so the gain cannot come from how probes pair beyond those strata. Closer per-fact agreement, however, brings more leakage under sampling; the method also fails TOFU's Forget Quality test and trails a single-number target on a second model family. These results establish that forgetting must restore how probes co-vary across facts, not just a likelihood level, and that never-seen facts supply this structure but not the retrained model's level, so a retrain-free estimate of that level is the missing piece.

1 INTRODUCTION

Machine unlearning asks a trained model to behave as if its *forget set* had never been used, for example after a deletion request, while keeping what it knows about the rest of its data, the *retain set* (Cao & Yang, 2015; Bourtoule et al., 2021). Unlearning is judged against the *gold retrain*, a model trained from scratch without the forget set (Maini et al., 2024). Retraining per request is too costly, so practical methods are retrain-free and must decide when to stop. The retrain still answers questions about facts it never saw, fluently, wrongly and with moderate likelihood, so success is a likelihood *level*, not a floor; a model pushed far below the level is detectable (Shi et al., 2024b; Hayes et al., 2025).

Recent work turns the level into a target by matching the forget set's likelihood to a reference measured on never-seen facts (Mirlach et al., 2026). Unlike a classifier's logit vector, a language model's never-seen answer is a different token sequence from a forget answer, so nothing aligns position by position and the target collapses to one number per fact, the likelihood of the trained answer. This *scalar bottleneck* discards which fact holds which value (permuting the values leaves any criterion on their population unchanged, proposition 1a) and every phrasing but the trained one.

The second omission is measurable before any method runs. On TOFU, a benchmark of question-answer pairs about fictitious authors (Maini et al., 2024), memorisation *decouples* a fact's phrasings: the correlation between the likelihood of a direct answer and of the same answer to a paraphrased

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.

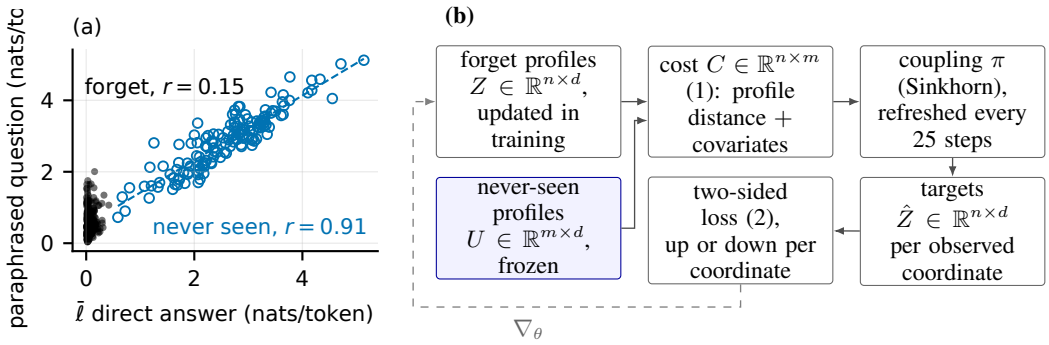

Figure 1: **(a)** Before unlearning (TOFU forget10, the split that forgets 10% of authors): per-token negative log-likelihood (NLL) of the direct answer against the paraphrased question (QPARA), memorised forget facts in black, never-seen facts in open blue with their least-squares line; r is Pearson's. **(b)** Conditional joint matching for n forget facts, m reference facts and d profile coordinates; blue: frozen; dashed: the gradient path. Takeaway: memorisation decouples phrasings that never-seen facts keep paired; the coupling restores the pairing in the targets (table 8).

question falls from 0.91 on never-seen facts to 0.15 on memorised ones (fig. 1a), and the shifts that would restore each phrasing's level differ by more than $5\times$ in magnitude, one in sign (table 7).

We therefore represent each fact by a *response profile*, the negative log-likelihoods of six probes, and match the joint law of forget profiles to that of never-seen profiles. Coordinate-wise targets cannot recover the never-seen correlation, because per-coordinate increasing maps preserve every rank correlation (proposition 1b); a coupling of whole profiles can. What such matching establishes about the retrain's law is bounded by a triangle inequality whose last term no retrain-free measurement reaches (eq. (4)).

Contributions.

- **Method** (section 3): response profiles transported as a joint law onto never-seen profiles by a covariate-penalised entropic coupling, a two-sided loss that stops at its target, and a stopping rule that needs no retrain.
- **Diagnosis**: memorisation decouples a fact's phrasings (fig. 1a), with an exact statement of why scalar and coordinate-wise targets cannot repair it (proposition 1).
- **Scope** (section 4): a decomposition of the law-level distance to the retrain, the assumptions behind it, and a measured bias in the reference level.
- **Evidence** (section 5): at 1B, the lowest mean per-fact distance to the retrain among ten target assignments and four published objectives, significant after correction only against coordinate-wise matching; a reference shuffled within relation and difficulty strata does as well, and closer agreement leaks more.

2 Related work

Forget objectives. Gradient ascent (GA) on the forget set is unbounded below; negative preference optimisation (NPO) and SimNPO bound the loss but still descend (Zhang et al., 2024; Fan et al., 2025); UNDIAL, Unilogit and LoTUS fix a soft target by convention (Dong et al., 2025; Vasilev et al., 2025; Spartalis et al., 2025). Our arms (section 5.1) measure the level instead (table 1). Methods that act on internal activations, such as representation misdirection (RMU) and probe-geometry alignment (Li et al., 2024; Rupa & Andy, 2026), do not face the scalar bottleneck.

Never-seen data as the target. ReGUn, the closest work, distils the forget set toward a class-conditioned held-out reference (Mirlach et al., 2026); LoTUS, Khan (2025) and Yao et al. (2024) also calibrate against unseen data, and Khan (2025) pre-selects the reference by covariate. We keep the reference but match a fact's whole profile and condition inside the coupling; our REGUN arm ports ReGUn's reference, not its objective.

Table 1: How each objective sets its stopping level and what it matches, from its published form. *Rest point*: whether the forget gradient vanishes at the level (for GA, NPO, SimNPO and our arms, from the sign of the loss derivative; section 3.3). NPO, UNDIAL and Unilogit also need a frozen copy of the model being unlearned, M, and JOINT, our method, reads the base model M_0 once, before training. Takeaway: only JOINT conditions on covariates beyond the class and matches a joint law.

objective	level set by	matched object	rest point
GA	none	sequence likelihood	no
NPO (2024)	retain weight	sequence likelihood	no
SimNPO (2025)	retain weight	sequence likelihood	no
UNDIAL (2025)	logit shift γ	token distribution	yes
Unilogit (2025)	uniform target	token distribution	yes
ReGUn (2026)	never-seen data, per class	class scalar	yes
MARGINAL	never-seen data	one marginal	yes
COORDINATE	never-seen data	six marginals	yes
JOINT	never-seen data, conditioned on relation, length and difficulty	joint law of the profile	yes

Transport-based unlearning. Entropic transport already serves as an unlearning objective on features (Choi & Na, 2025), parameters (Choi et al., 2025) and data distributions (Choi et al., 2026), but not on covariate-conditioned probe likelihoods; its one-dimensional case is our MARGINAL arm.

Several phrasings of one fact. Consistency editing collapses the variants' updates (Han et al., 2026); other work also suppresses rephrasings (Li et al., 2026), shows leakage under sampling (Reisizadeh et al., 2026) or asks for form-independent unlearning (Ye et al., 2026). These push extra phrasings *down*, whereas a never-seen fact sits at no phrasing's floor (table 7). Wei et al. (2025) evaluate through correlated facts; Yang & Yeung (2026) show that oracle matching on trained probes leaves other paraphrases below the never-seen level and limit oracle-free certification, a diagnosis we build on. Thaker et al. (2025) and Yu et al. (2025) argue that exact retrain equivalence is out of reach, so we claim only restored likelihoods. The reference, entropic transport and the probe taxonomy (Cohen et al., 2024; Wang et al., 2025) are prior work; new is transporting whole profiles under a covariate-conditioned cost onto a two-sided target with a stopping rule that needs no retrain.

3 METHOD

The method freezes its probes and never-seen profiles before training, trains toward coupled targets, and chooses the epoch without the retrain, which is read only at evaluation (fig. 2 and algorithm 1).

3.1 RESPONSE PROFILES

Definition 1 (Response profile)**.** For a model θ and a fact i, $z_\theta(i) \in \mathbb{R}^d$, $d = 6$, holds the per-token negative log-likelihoods (NLL, in nats) that θ assigns to six probes of i: the trained question and answer (DIRECT); a paraphrased question (QPARA); a paraphrased answer (APARA); the answer as prompt and the question as target (INVERSE); an alternative phrasing of the answer (ALT); and the mean over up to three answers rewritten to state different facts (PERT). All but ALT are *core* coordinates, required of every kept fact.

Qwen2.5-3B-Instruct (Qwen Team, 2025), never the model under test, writes the probes (section C.3). An entailment model, DeBERTa-v3-large-MNLI (He et al., 2023; Laurer et al., 2024), keeps a paraphrase only if it entails the original both ways without copying it, and a perturbation only if the original does not entail it. Facts missing a core probe leave the objective, which keeps 75% of forget and 70% of reference facts; the generator is a dependency of the method (section 6).

3.2 CONDITIONED NEVER-SEEN REFERENCES

The never-seen facts come from TOFU's held-out authors, written like the others but never trained on. The twenty held-out authors, kept whole since one author's questions share entities, are split

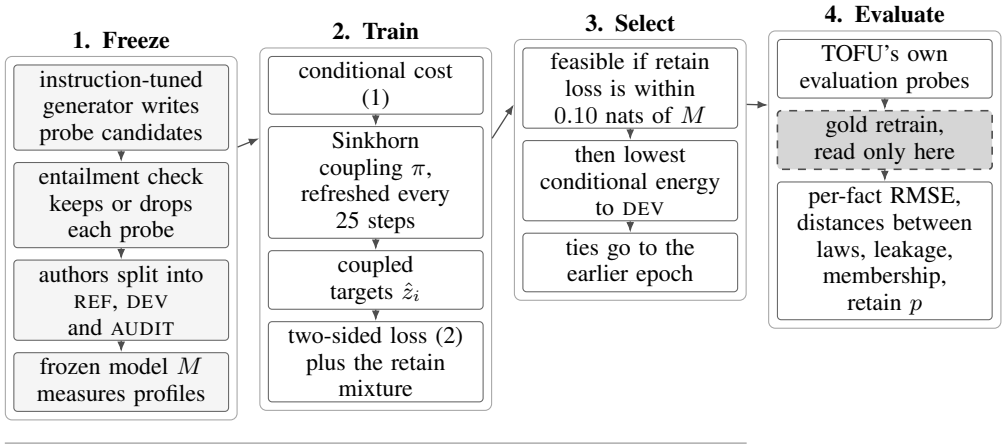

Figure 2: The pipeline, left to right: grey boxes are frozen before training, white boxes compute, and the dashed shaded box is the gold retrain. Takeaway: given a pool of never-seen facts, the retrain is read once, after the epoch is chosen from quantities a deployed system would have.

Algorithm 1 Conditional joint matching; G is the probe generator and $\mathcal{N}$ the entailment checker, other symbols as in section A. Lines 1 to 6 run once before training and line 12 recomputes the coupling; no line reads a retrained model.

1: $\mathcal{P} \leftarrow G(\mathcal{F} \cup \text{REF} \cup \text{DEV} \cup \text{AUDIT})$
2: drop any probe $\mathcal{N}$ rejects; drop any fact missing a core coordinate ▷ section C.3
3: $x_i \leftarrow (\text{rel}_i,\ L_i,\ D_i) = (\text{relation}(q_i),\ |y_i|,\ \bar{\ell}_{M_0}(y_i \mid q_i))$
4: freeze $\mathcal{P}$ and x
5: $U \leftarrow \{z_M(j)\}_{j \in \text{REF}}$ ▷ the reference; frozen
6: $\theta \leftarrow M$; $\text{step} \leftarrow 0$
7: **for** epoch $= 1, \ldots, n_{\text{ep}}$ **do**
8: **for** each micro-batch B of whole facts from $\mathcal{F}$ **do**
9: **if** step $\bmod K = 0$ **then**
10: $Z \leftarrow \{z_\theta(i)\}_{i \in \mathcal{F}}$ ▷ no-grad pass
11: $C \leftarrow \text{COST}(Z, U, x)$ ▷ eq. (1)
12: $\pi \leftarrow \text{SINKHORN}(C, \eta)$; $\hat{z}_{ic} \leftarrow \sum_j \pi_{ij} o_{jc} u_{jc} \,/\, \sum_j \pi_{ij} o_{jc}$ for every observed c
13: **end if**
14: $\ell \leftarrow \big(\sum_{i \in B} |O_i|\big)^{-1} \sum_{i \in B} \sum_{c \in O_i} \rho_\delta(z_{ic}(\theta) - \hat{z}_{ic})$ ▷ eq. (2)
15: $\ell \leftarrow \ell + \frac{1}{n_{\mathcal{R}}} \sum_{b=1}^{n_{\mathcal{R}}} \text{loss}_{\mathcal{R}}(\theta, B_b^{\mathcal{R}})$ ▷ $n_{\mathcal{R}}$ retain micro-batches, averaged
16: $\theta \leftarrow \text{UPDATE}(\theta, \nabla_\theta \ell)$; $\text{step} \leftarrow \text{step} + 1$
17: **end for**
18: record E_{DEV} between $\{z_\theta(i)\}_{\mathcal{F}}$ and $\{z_M(j)\}_{\text{DEV}}$ (eq. (5)), and the loss on 200 fixed retain questions
19: **end for**
20: **return** the epoch with the lowest E_{DEV} among those whose retain loss is within τ of M's, ties to the earlier epoch, or the first saved epoch if no epoch is feasible

into ten for REF (targets), five for DEV (stopping rule) and five for AUDIT (reported numbers only). Matching is conditioned on $x_i = (\text{rel}_i, L_i, D_i)$: a rule-based relation type (nine classes), the answer length in tokens and, as difficulty, the DIRECT NLL under the base model M_0 before fine-tuning (the fine-tuned model's NLL is the memorisation being removed). Conditioning matters because the populations differ in composition; WORKS, for example, covers 92 of 299 forget facts but 56 of 140 references. In leave-one-out ablations, the difficulty penalty helps and the relation penalty hurts (table 4).

3.3 CONDITIONAL JOINT MATCHING

Two observations force a joint match. A criterion that sees only the population of profiles cannot tell which fact holds which profile (the *fact assignment*), and per-coordinate increasing maps cannot change how coordinates co-vary in rank. The energy distance and Sinkhorn divergence below are distances between distributions (section 4).

Proposition 1 (What pooled and coordinate-wise matching cannot see)**.** *Let $Z = \{z(i)\}_{i\in\mathcal{F}}$ be a finite set of profiles.* ***(a)*** *For a permutation ω of $\mathcal{F}$, the profiles $z'(i) = z(\omega(i))$ have the same empirical law as Z, so the energy distance, the Sinkhorn divergence and any per-coordinate test score them identically, while $\max_i \|z'(i) - z(i)\|$ can equal the diameter of Z.* ***(b)*** *If $T(z)_c = T_c(z_c)$ with every T_c strictly increasing, every rank correlation between two coordinates of $T(Z)$ equals that of Z, however exactly each T_c reproduces a target marginal.*

Part (a) is why evaluation is per fact (section 5.1), and part (b) is why coordinate-wise targets keep the forget set's pairing between phrasings (section 5.3); proofs are in section B. Our objective is not exempt from (a): permuting profiles within a covariate stratum leaves its coupling, targets and loss unchanged.

Let $Z = \{z_\theta(i)\}_{i\in\mathcal{F}}$ be the forget profiles under training and $U = \{z_M(j)\}_{j\in\text{REF}}$ the reference profiles under the frozen *contaminated model* M, the fine-tuned model to be unlearned. The ground cost compares two facts' profiles on the probes both have and prices covariate mismatch:

$$C_{ij} = \frac{d}{|O_{ij}|}\sum_{c\in O_{ij}}(z_{ic} - u_{jc})^2 + \lambda_{\text{rel}}\mathbf{1}[\text{rel}_i \neq \text{rel}_j] + \lambda_{\text{len}}(\tilde{L}_i - \tilde{L}_j)^2 + \lambda_{\text{dif}}(\tilde{D}_i - \tilde{D}_j)^2, \qquad (1)$$

with O_{ij} the coordinates observed in both facts, $\tilde{L}, \tilde{D}$ standardised, and $C_{ij} = \infty$ if $O_{ij} = \emptyset$. The coupling $\pi \in \mathbb{R}^{n\times m}_{\geq 0}$ ($n = |\mathcal{F}|$, $m = |\text{REF}|$) is a soft matching, π_{ij} being the mass that fact i sends to reference fact j; it is the entropic optimal-transport plan (Cuturi, 2013)

$$\pi = \underset{\pi\mathbf{1}=\mathbf{1}/n,\ \pi^\top\mathbf{1}=\mathbf{1}/m}{\arg\min} \langle\pi, C\rangle/\bar{C} - \eta H(\pi),$$

with H the entropy and $\bar{C}$ the median finite cost, solved by log-domain Sinkhorn iterations (section C.2). Each target is the coupled average of the reference, per coordinate over the mass that observes it,

$$\hat{z}_{ic} = \frac{\sum_j \pi_{ij}\, o_{jc}\, u_{jc}}{\sum_j \pi_{ij}\, o_{jc}}, \qquad o_{jc} = \mathbf{1}[\text{coordinate } c \text{ observed for } j],$$

and coordinates whose coupled observed mass is below 10^{-12} are masked. The loss is a two-sided Huber pull (Huber, 1964) averaged over all observed cells,

$$L_{\mathcal{F}}(\theta) = \Big(\sum_{i\in\mathcal{F}}|O_i|\Big)^{-1}\sum_{i\in\mathcal{F}}\sum_{c\in O_i}\rho_\delta\big(z_{ic}(\theta) - \hat{z}_{ic}\big), \qquad \rho'_\delta(0) = 0, \qquad (2)$$

with O_i the coordinates kept for fact i. π is recomputed every 25 steps from detached profiles against the frozen U, so eq. (2) trains as a regression.

P1 (the loss stops at the target). Because $\rho'_\delta(0) = 0$, the gradient of eq. (2) vanishes when every coordinate sits at its target and pushes back from either side, lowering NLL where needed (INVERSE starts above its reference; table 7). A monotone forget loss $\sum_i \phi_i(\ell_i)$ with every ϕ'_i of one sign, as in GA, NPO and SimNPO, only raises NLL and can stop only where it cancels the retain gradient. Observed: a variant that only raises NLL overshoots far past the target (table 4), and at the selected checkpoints the forget gradient still dominates, so we claim no stationarity (section E.2). Unlike one-dimensional rank matching, the coupling must also handle a missing optional probe, the populations' different class proportions, forget profiles that move while U is frozen, and two degenerate limits of η (section C.1).

3.4 Retain preservation and choosing without an oracle

Keeping retain *answers* likely is not enough: on MUSE-News, the MARGINAL arm matched the retrain on forget-corpus knowledge but lost much retain-corpus knowledge (table 15). The retain term therefore mixes, under cross-entropy, retain question-answer pairs, continuations cut from retain answers and passages from an independent corpus, averaging 4 retain micro-batches per step (section C.2).

Training runs 6 epochs, and the epoch is chosen without the retrain. Among epochs whose loss on 200 fixed retain questions is within $\tau = 0.10$ nats of M's, the rule takes the lowest E_{DEV}, the energy distance between forget and DEV profiles within each relation class, averaged with class weights (eq. (5)), with ties to the earlier epoch. The rule reads DEV, not REF, so that E_{DEV} is not a second training signal (algorithm 1, fig. 2).

Requirements and cost. No retrained model is used before evaluation. The method needs never-seen facts from the forget set's fact-generating law, a frozen probe generator and entailment checker, M_0 and the retain mixture. A synchronised step costs +16% time and +12% peak memory over gradient ascent, before the profile refresh (section C.5).

4 What matching the reference can and cannot establish

Matching the reference helps only insofar as the reference stands in for the retrain.

The decomposition. Let $\hat{\mu}_\theta$ be the empirical law of the forget profiles, $\hat{\nu}_m$ that of the m reference profiles with population law ν, and $\mu_\star$ the law of the forget profiles under the retrain. Laws are compared by the energy distance

$$\mathcal{E}(P,Q) = 2\,\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|, \qquad W(P,Q) = \sqrt{\mathcal{E}(P,Q)}, \tag{3}$$

with $X, X' \sim P$ and $Y, Y' \sim Q$ independent; $\mathcal{E} = 0$ only if $P = Q$, and W is a metric for laws with a finite first moment (Székely & Rizzo, 2013), as the bounded per-token NLLs of the five core coordinates have (section B). The entropic transport objective of section 3.3 is positive even for identical populations, and the Sinkhorn divergence is not known to satisfy the triangle inequality (Feydy et al., 2019), so neither enters the bound

$$\underbrace{W(\hat{\mu}_{\hat{\theta}}, \mu_\star)}_{\text{what we want small}} \le \underbrace{W(\hat{\mu}_{\hat{\theta}}, \hat{\nu}_m)}_{\varepsilon_{\text{opt}}} + \underbrace{W(\hat{\nu}_m, \nu)}_{\varepsilon_{\text{est}}} + \underbrace{W(\nu, \mu_\star)}_{\varepsilon_{\text{mis}}} \tag{4}$$

for the selected model $\hat{\theta}$. ε_{opt} is the fit to the finite reference, including the entropic bias. ε_{est}, the price of m facts instead of the population, is unobservable and probed by perturbing the reference (section E.1). ε_{mis} compares how M treats facts it never saw with how the retrain treats the facts *it* never saw; only this term involves the retrain, so no retrain-free measurement reaches ε_{mis} (Yang & Yeung, 2026). The bound concerns laws: by proposition 1a it says nothing about per-fact error, which section 5 measures directly.

Assumptions. **(A1)** Given x, forget and reference facts are draws from one fact-generating law. **(A2)** For facts absent from training, the map from fact to profile is the same under M and under the retrain. One conditional law of facts pushed through one map gives one law of profiles, so under A1 and A2 $\nu(\cdot \mid x) = \mu_\star(\cdot \mid x)$: the *invariance part* of ε_{mis} vanishes, leaving the *composition part* from the populations' different covariate distributions. That part vanishes too if **(A3)** every covariate value of a forget fact occurs in the reference pool and the reference is reweighted to the forget set's covariates. We do not reweight: 0.78 of the coupling's mass stays within a relation class by the end of training, so the composition part is reduced, not removed. A2 needs a retrain to test. With one as an evaluation-only oracle we observe **P2 (the reference carries the retrain's structure, not its level)**: the retrain's forget profiles share the reference's correlation but sit above its mean on every core coordinate (section 5.5), so $\varepsilon_{\text{mis}} \neq 0$ in practice.

5 Experiments

5.1 Setup

We unlearn TOFU forget10 (Maini et al., 2024), which forgets 20 of 200 fictitious authors (400 pairs, all evaluated; the 299 with every core probe enter the objective), from Llama-3.2-1B (Grattafiori et al., 2024) fine-tuned on all 4000 pairs with the retrain's recipe (section C.2). Evaluation uses TOFU's own paraphrases and perturbations and a reworded inverse template, which share no probe with training except DIRECT and one QPARA.

Arms. An *arm* is one way of assigning targets; all arms share probes, reference, retain data, optimiser, epoch budget and stopping rule (table 4). The *ladder* runs from one scalar for all facts (SCALAR) through one scalar per relation class (REGUN) and one-dimensional transport on DIRECT (MARGINAL) to per-coordinate transport (COORDINATE) and conditional joint transport, the method

Table 2: TOFU forget10, Llama-3.2-1B. Top: our arms at their selected epoch; bottom: published objectives at their own recipes, read at the last epoch (GA in tables 10 to 12). #: seeds; RMSE in nats, mean (sd); farther: seeds on which the arm is farther from the retrain than JOINT; p: Holm-adjusted paired t-test (n/a: not among the comparisons tested in table 6); FQ: TOFU's Forget Quality, median over seeds, passing above 0.05; excess leak (leak@k minus the seed's retrain's, 0 when retrain-like) and retain p (answer probability on retain questions): means. Bold: best over all rows. Takeaway: JOINT has the lowest mean, but only its gap to COORDINATE is significant, and it leaks most.

			vs. JOINT				
arm	#	per-fact RMSE ↓	farther	p	FQ ↑	excess leak	retain p ↑
SCALAR	5	1.031 (0.174)	n/a	n/a	$10^{-8.8}$	+0.164	0.895
REGUN	5	0.967 (0.060)	n/a	n/a	$10^{-11.6}$	+0.243	0.924
COND-MEAN	5	0.948 (0.062)	5/5	0.15	$10^{-5.7}$	+0.307	0.923
MARGINAL	5	1.115 (0.088)	n/a	n/a	$10^{-8.8}$	+0.183	0.951
MARGINAL-COND	2	1.021 (0.220)	n/a	n/a	$10^{-9.2}$	+0.205	0.953
COORDINATE	5	0.906 (0.036)	5/5	0.01	$10^{-6.5}$	+0.267	0.953
COORDINATE-COND	5	0.833 (0.030)	4/5	0.54	$10^{-6.5}$	+0.338	0.954
JOINT	5	**0.785** (0.054)	n/a	n/a	$10^{-4.9}$	+0.406	0.949
JOINT-SHUFFLED	5	0.976 (0.052)	5/5	0.07	$10^{-6.5}$	+0.248	0.946
JOINT-SHUFFLED-COND	5	0.797 (0.032)	4/5	0.75	$10^{-6.0}$	+0.405	0.950
UNDIAL	3	0.975 (0.013)	n/a	n/a	$10^{-12.3}$	+0.316	0.957
SimNPO	3	1.684 (0.369)	n/a	n/a	0.37	**−0.023**	**0.970**
NPO	3	2.886 (0.758)	n/a	n/a	**0.58**	−0.042	0.950

(JOINT). Five further arms attribute the gain. COORDINATE-COND applies JOINT's cost per coordinate, so it lacks only the coupling across coordinates; MARGINAL-COND does so on DIRECT alone; COND-MEAN sends each coordinate to its class's reference mean. Two controls run JOINT against a reference whose coordinates are permuted across facts, keeping every marginal: JOINT-SHUFFLED permutes within relation classes, also breaking the links to length and difficulty, and JOINT-SHUFFLED-COND within difficulty terciles inside relation classes, partly keeping the link to difficulty (section E.2). If pairing between phrasings beyond relation and difficulty tercile matters, JOINT must beat JOINT-SHUFFLED-COND. Five paired seeds each have their own M, reference profiles and retrain (section D.2); every arm runs at all five seeds except MARGINAL-COND (two), and GA, NPO, SimNPO and UNDIAL run at their own recipes at three (section D.3).

Metrics. The main metric is the *per-fact RMSE*: for each forget fact, the root-mean-square NLL difference to the retrain over the core coordinates both observe, averaged over facts (nats; table 5 gives two controls for scale). Pooled criteria, blind to per-fact error (proposition 1a), are reported beside the RMSE: the energy distance between the arm's and the retrain's forget profiles, and TOFU's Forget Quality (FQ), the p-value of a two-sample Kolmogorov-Smirnov test between the arm's and the retrain's truth ratios (the likelihood of false answers relative to a paraphrased true one), passing above 0.05. Behaviour is measured by leak@k, the fraction of forget facts for which one of $k = 16$ samples entails the true answer under the entailment checker, reported as excess over the seed's retrain. Retain p is the answer probability on 400 fixed retain questions, all seen in training, whose first half also feeds the stopping rule. An author-clustered bootstrap gives a 95% half-width near 0.06 nats, so arms are compared by seed-paired t-tests, Holm-adjusted (section D.2).

5.2 RQ1: IS JOINT MATCHING CLOSEST TO THE RETRAIN?

JOINT has the lowest mean per-fact RMSE and is ahead of COORDINATE, COND-MEAN and JOINT-SHUFFLED on every seed, but only its gap to COORDINATE survives the Holm correction (table 2). The energy distance, blind to the fact assignment, separates the arms far less (table 5). All published objectives are farther; re-tuned under our stopping rule at seeds 3 and 4, the best of them still trails (section D.3).

Forget Quality ranks the objectives differently. NPO and SimNPO pass it while sitting far from the retrain per fact, yet JOINT, the best of our arms on FQ, fails it (tables 2 and 12): a population test cannot see per-fact error (proposition 1a).

5.3 RQ2: WHAT DOES THE COUPLING DO TO THE PROFILE GEOMETRY?

COORDINATE's monotone maps keep the forget set's pairing between phrasings in its targets (proposition 1b), JOINT's targets carry the reference's, and the two shuffles fall in between (table 8). Proposition 1b constrains the targets, not the trained model: COORDINATE's trained profiles regain much of the correlation, and JOINT's correlation matrix lies closest to the retrain's (table 5).

5.4 RQ3: WHICH INGREDIENT CARRIES THE GAIN?

Neither half of the gap over COORDINATE holds on every seed. COORDINATE-COND splits it into a conditioning part (COORDINATE minus COORDINATE-COND: +0.057, −0.001, +0.136, +0.115 and +0.056 nats by seed) and a coupling part (the COORDINATE-COND row of table 6); both change sign, so at five seeds the gap cannot be attributed to either. The smoothest arm, COND-MEAN, is farther than JOINT on every seed, though partial shrinkage is not excluded (section E.2).

The mechanism control fails. JOINT-SHUFFLED-COND, which keeps each fact's relation class and difficulty tercile but destroys the rest of its pairing, is not separated from JOINT (95% interval of the paired difference [−0.084, +0.107] nats), while JOINT-SHUFFLED, which also breaks the covariate links, is farther on every seed (table 2). Relation, length and difficulty alone predict reference profiles (each fact left out) with $r(\text{DIRECT}, \text{QPARA})$ of 0.978 against 0.907 measured. The reference thus appears to need its covariate links; whether it needs pairing beyond them, five seeds cannot say. On Qwen2.5-1.5B JOINT is again not separated from JOINT-SHUFFLED-COND, but SCALAR lands closer than JOINT on both seeds (section E.4).

5.5 RQ4: WHERE DOES THE FIT STOP?

P2: the reference carries the retrain's structure, not its level. The *level bias*, the retrain's mean forget profile minus the reference's, is positive on every core coordinate (table 9), so exact matching leaves forget answers more likely than the retrain does. The retrain's correlation, however, transfers (0.941 on forget facts against 0.937 on never-seen ones). Section E.1 perturbs the reference four ways.

Profile agreement is not behavioural forgetting. JOINT keeps retain p but leaks most, its membership AUC sits above chance (table 10), and it relearns faster than the retrain after fine-tuning on other forget facts (section E.3). Across the nine arms run at every seed, per-fact distance to the retrain and excess leak have rank correlation −0.95: the closer an arm matches the retrain's profile, the more it leaks, as the level bias predicts. NPO and SimNPO, far from the retrain, leak no more than the retrain does.

Re-reading the reference under the model being unlearned does not help. Re-measuring REF under the current parameters at every refresh meets its preregistered criteria on none of the four seeds with a stored contaminated model, because the re-read level follows what training does to the whole domain (section E.5).

6 LIMITATIONS

Mechanism. The cross-coordinate coupling cannot be separated from the covariate conditioning, and only the gap to COORDINATE is significant at five seeds (sections 5.2 and 5.4). *Behaviour.* The method leaks most and fails the Forget Quality test that NPO and SimNPO pass (sections 5.2 and 5.5). *Equivalence.* ε_{mis} cannot be observed without a retrain (section 4). *Deployment.* The method needs never-seen facts from the forget set's own law, which TOFU supplies and a real request may lack; small or covariate-mismatched pools fail measurably (section E.1). *Scale.* Profiles are measured on one benchmark and forget fraction in two model families, at five seeds at 1B and two at 1.5B; MUSE-News at 7B is read only through its own knowledge metrics (section E.6). The closest arm differs between families (section E.4), and the arms that train on whole profiles are untested above 1.5B. *Knowledge and probes.* Knowledge from pre-training (Eldan & Russinovich, 2023; Li et al., 2024) is untested, and one generator and threshold set write every probe, although the decoupling survives the filter and another generator (section C.3). *Constants and metrics.* The

coupling constants were fixed in advance and dropping the relation penalty helps (table 4); TOFU's Model Utility was not evaluated, and retain probability and perplexity stand in (tables 10 and 11).

7 CONCLUSION

Conditional joint matching transports a fact's whole response profile onto never-seen facts and, at 1B, has the lowest mean per-fact distance to the retrain among the objectives compared, although only its gap to coordinate-wise matching is significant at five seeds. A reference shuffled within relation and difficulty strata does as well, so the gain cannot be attributed to how probes pair beyond those strata, and closer agreement brings more leakage under sampling. The residual distance comes from both the fit to the reference and the reference itself, which carries the retrain's structure but not its level (table 9); a retrain-free estimate of that level is the missing piece.

AI USE STATEMENT

Generative models were used in two roles, and the authors take responsibility for both. *As an instrument of the method*, an open-weight instruction-tuned model writes the probes and an entailment model validates them, so part of the data this paper trains and evaluates on is model-generated; sections 3.1 and C.3 give the instructions, the checks, the survival rates and why the model under test is never its own generator. *For writing*, assistants drafted and revised prose and captions, redrew figures, and searched the literature for related work and for the published venue of cited preprints; the authors verified every citation against its source. No research idea, hypothesis or experiment design originates from a model, no mathematical claim or proof was produced by one, and no result was generated or interpreted except through code the authors read and predicates that fail the build when they stop holding. The authors reviewed all AI-assisted work and take responsibility for the final content.

ETHICS STATEMENT

Data. No human subjects, user data or personal data were collected, so no board review was required: TOFU's biographies are synthetic and name no real person, MUSE-News is published BBC text about public figures, WikiText-2 is Wikipedia text, the canary codes are synthetic, and the generated probes paraphrase fictitious biographies. *Privacy.* The facts to be forgotten reach a probe generator; ours is a local open-weight model, and it must stay local when the facts are personal. *Dual use.* A model matched to look never-trained on a likelihood profile can still leak under sampling and relearn quickly (table 10 and section E.3), so an auditor accepting profile-level evidence alone could be misled about compliance; every pooled number here is therefore paired with behavioural measurements, and the claim is calibration, not removal. *Broader impact.* Apparent unlearning eases pressure for the retraining or data governance that a right-to-be-forgotten request may require; we state the limit, not the assurance. *Footprint.* 74.7 graphics-processing-unit hours, 26.2 kWh, about 10.2 kg CO_2e (section C.4).

REPRODUCIBILITY STATEMENT

Every reported result comes from released records through released code; the few values typed into the prose match those records. The recipe, optimiser, coupling constants, stopping rule, what is fitted and which split reaches which stage are in section C.2; the probe instructions and checks in section C.3; datasets, models and licences in section C.4; the arms in table 4; and the proof and its standing assumption in section B. Most records carry a provenance block with the recipe and a hash of the weights (section C.4). The anonymised code in the supplementary material includes a pinned dependency list, one command that regenerates every table and figure from the records on a processor in under ten minutes, and a checklist that maps each claim to the record behind it.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2021. arXiv:1912.03817.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2015.

Dasol Choi and Dongbin Na. Distribution-level feature distancing for machine unlearning: Towards a better trade-off between model utility and forgetting. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Hyundo Choi, Junhyeong An, Jinseong Park, and Jaewoong Choi. Unlearning for one-step generative models via unbalanced optimal transport. *arXiv preprint arXiv:2603.16489*, 2026.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2406.12329.

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 (TACL)*, 12:283–298, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2410.07163.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiaoqi Han, Víctor Gutiérrez-Basulto, Ru Li, Xiaoli Li, Jiye Liang, and Jeff Z. Pan. Consistency-aware editing for entity-level unlearning in language models. *arXiv preprint arXiv:2601.08840*, 2026.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2111.09543.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? Jogging the memory of unlearned LLMs via benign relearning. *arXiv preprint arXiv:2406.13356*, 2024.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Junaid Iqbal Khan. Leveraging distribution matching to make approximate machine unlearning faster. *arXiv preprint arXiv:2507.09786*, 2025.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100, 2024.

Kemou Li, Qizhou Wang, Yue Wang, Fengpeng Li, Jun Liu, Bo Han, and Jiantao Zhou. LLM unlearning with LLM beliefs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.19422.

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. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03218.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2401.06121.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.07843.

Jonas Mirlach, Sonia Laguna, and Julia E. Vogt. Reference-guided machine unlearning. *arXiv preprint arXiv:2603.11210*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. Leak@k: Unlearning does not make LLMs forget under probabilistic decoding. *arXiv preprint arXiv:2511.04934*, 2026.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Anamika Paul Rupa and Anietie Andy. Probe-geometry alignment: Erasing the cross-sequence memorization signature below chance. *arXiv preprint arXiv:2605.01699*, 2026.

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 *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2310.16789.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024b.

Christoforos N. Spartalis, Theodoros Semertzidis, Petros Daras, and Efstratios Gavves. LoTUS: Large-scale machine unlearning with a taste of uncertainty. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.18314.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: LLM unlearning benchmarks are weak measures of progress. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025. arXiv:2410.02879.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, and Christof Monz. Unilogit: Robust machine unlearning for LLMs using uniform-target self-distillation. *arXiv preprint arXiv:2505.06027*, 2025.

Huazheng Wang, Yongcheng Jing, Haifeng Sun, Yingjie Wang, Jingyu Wang, Jianxin Liao, and Dacheng Tao. Erasing without remembering: Implicit knowledge forgetting in large language models. *arXiv preprint arXiv:2502.19982*, 2025.

Rongzhe Wei, Peizhi Niu, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K. Potluru, Eli Chien, Kamalika Chaudhuri, Olgica Milenkovic, and Pan Li. Do LLMs really forget? Evaluating unlearning with knowledge correlation and confidence awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.05735.

Sen Yang and Yuen-Hei Yeung. Unlearning as distribution restoration: A controlled counterfactual study, a validated selective screen, and the limits of oracle-free certification. *arXiv preprint arXiv:2607.19442*, 2026.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Xiaotian Ye, Mengqi Zhang, and Shu Wu. LLM unlearning should be form-independent. In *IEEE Symposium on Security and Privacy (S&P)*, 2026.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning. *arXiv preprint arXiv:2510.16629*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.05868.

A NOTATION

symbol	meaning	symbol	meaning
$\mathcal{F}$, $n = \lvert\mathcal{F}\rvert$	forget facts	REF, DEV, AUDIT; $m = \lvert\text{REF}\rvert$	roles of the never-seen authors
$d = 6$; d'; c	profile coordinates; probe sequences per fact (5 to 8); coordinate	O_{ij}, O_i	coordinates observed in both i, j; in i
M, M_0	contaminated model; base model	θ, $\hat{\theta}$, $\theta_\star$	current, selected, gold-retrain parameters
$G, \mathcal{N}, \mathcal{P}$	probe generator; entailment checker; probe set	q, y	question (prompt); answer
$\bar{\ell}$	per-token NLL of y given q	$z_\theta(i) \in \mathbb{R}^d$	response profile of fact i under θ
$u_j = z_M(j)$; U, Z	reference profile; reference and forget profile sets	$x_i = (\text{rel}_i, L_i, D_i)$; $\tilde{L}, \tilde{D}$	covariates; standardised length, difficulty
C_{ij}; $\lambda_{\text{rel}}, \lambda_{\text{len}}, \lambda_{\text{dif}}$; o_{jc}	ground cost (1); its penalties; observation indicator	π; η; $\bar{C}$; $\hat{z}_{ic}$	coupling; entropic regulariser; median cost; coupled target
ρ_δ; δ	Huber loss (Huber, 1964); its width	$L_\mathcal{F}$; $\text{loss}_\mathcal{R}$	forget loss (2), averaged over observed cells; retain loss
ϕ_i; τ	monotone forget loss (P1); retain budget	K; n_{ep}; B; $n_\mathcal{R}$	refresh interval; epochs; facts per micro-batch; retain micro-batches
E_{DEV}	stopping statistic (5)	$\mathcal{E}$, $W = \sqrt{\mathcal{E}}$	energy distance (3) and its metric
$\hat{\mu}_\theta$; $\hat{\nu}_m$, ν; $\mu_\star$	forget-profile law; reference law, its population; retrain law	$\varepsilon_{\text{opt}}, \varepsilon_{\text{est}}, \varepsilon_{\text{mis}}$	terms of eq. (4)
r	Pearson correlation (rank correlation in proposition 1)	T; ω; δ_z	coordinate-wise map; permutation; Dirac mass
k; t	samples in leak@k; relearning steps	$\lvert\theta\rvert$, s	parameter count; token length (table 3)

symbol	meaning	symbol	meaning
$\mathsf{F}(s)$, $\mathsf{A}(s)$; κ	forward cost and activation memory; Sinkhorn iterations	ℓ	step loss in algorithm 1

Sets of facts and probes are calligraphic ($\mathcal{N}$ and $\mathcal{E}$ excepted), profile sets and models Roman capitals, laws Greek, and arms, probes, roles and relation classes small capitals; $\mathbf{1}$ is an indicator or the all-ones vector; i, j, c index forget facts, reference facts and coordinates; b indexes retain micro-batches in algorithm 1 and g relation classes in eq. (5).

B Standing assumption and proof of proposition 1

This appendix states the assumption behind the bound of section 4 and proves proposition 1.

Standing assumption. (R1) Bounded profiles: each of the five core coordinates of $z_\theta(i)$ is present for every kept fact and is a bounded per-token NLL. Every law in eq. (4) is taken on those five coordinates, so each law has a finite first moment and W of eq. (3) is a metric (Székely & Rizzo, 2013). No convexity, Lipschitz or bounded-gradient assumption is made, and no convergence is claimed.

Proof of proposition 1. (a) The empirical law $|\mathcal{F}|^{-1}\sum_i \delta_{z(i)}$ is a sum over the set of profiles, so re-indexing the profiles by ω leaves it unchanged, and every functional of the empirical law takes the same value on Z and on the permuted profiles. Letting ω swap the two most distant profiles attains the diameter. No parameter vector realising the permutation is asserted. *(b)* Each T_c preserves the order of $\{z_c(i)\}_i$, so every rank vector is unchanged, and so is any statistic of the ranks, a rank correlation included. □

C Method and reproduction details

This appendix gives what section 3 leaves out: why the coupling is harder than rank matching; the recipe; probe construction and validation; datasets, models, licences and compute; and the asymptotic cost of each objective.

C.1 Why the coupling is not six one-dimensional matches

Four problems separate the coupling from rank matching. *No alignment, and missing probes*: forget and reference facts share no tokens, so only profiles can be compared. ALT is also missing for a quarter of facts (table 7); the cost averages over O_{ij} and the target reads observed mass only, so a missing probe is never a zero. *Composition*: the populations' class proportions differ, so a balanced plan must move mass across classes, which the finite penalties price. *A moving source*: the forget profiles change while U is frozen, so π goes stale between refreshes (section E.2). *Degenerate limits*: as $\eta \to 0$ each fact draws on few reference facts and inherits their noise, and for large η every fact receives the reference mean; at $\eta = 0.05$ a target averages 8.0 effective neighbours. Only the first problem is specific to whole profiles; the conditioned per-coordinate arms (COORDINATE-COND, MARGINAL-COND) share the others.

C.2 Recipe

Every arm trains Llama-3.2-1B for 6 epochs at learning rate 10^{-5} with AdamW (weight decay 0.01, gradient clip 1.0), in micro-batches of 2 facts, each step averaging 4 retain micro-batches of 4 sequences under cross-entropy. The coupling uses $\eta = 0.05$ on the median-scaled cost and penalties $\lambda_{\mathrm{rel}} = 4$, $\lambda_{\mathrm{len}} = 1$, $\lambda_{\mathrm{dif}} = 1$ on standardised covariates, and the coupling is refreshed every 25 steps; the Huber width is $\delta = 1$. Sinkhorn iterates in log space for at most 4000 iterations, stopping once the row potentials change by less than 10^{-9}. Among epochs whose loss on 200 fixed retain questions (the first half of the 400 behind retain p) is within $\tau = 0.10$ nats of M's, the stopping rule takes the lowest conditional energy to DEV on generator-written probes, with ties to the earlier epoch. M and the gold retrains share TOFU's fine-tuning recipe: five epochs at 10^{-5}, effective

batch 32, AdamW with weight decay 0.01, 10% warm-up, linear decay and gradient clip 1.0. The conditional energy is

$$E_{\text{DEV}} = \sum_g w_g \, \mathcal{E}\big(\hat{\mu}_\theta^{(g)}, \hat{\nu}_{\text{DEV}}^{(g)}\big), \tag{5}$$

with $\hat{\mu}_\theta^{(g)}$ and $\hat{\nu}_{\text{DEV}}^{(g)}$ the empirical laws of the forget and DEV profiles of relation class g and w_g the fraction of scored forget facts in that class. Classes with fewer than five facts on either side are dropped; if no epoch meets the retain budget the rule returns the first saved epoch. $\mathcal{E}$ is the U-statistic wherever it is reported and the V-statistic inside the rule. The per-fact RMSE of section 5.1 is

$$\text{RMSE}(\hat{\theta}) = \frac{1}{|\mathcal{F}'|} \sum_{i \in \mathcal{F}'} \Big(\frac{1}{|O_i^\star|} \sum_{c \in O_i^\star} \big(z_{\hat{\theta},c}(i) - z_{\theta_\star,c}(i) \big)^2 \Big)^{1/2}$$

on evaluation probes, with $O_i^\star$ the core coordinates observed for fact i under both models and $\mathcal{F}'$ the facts with at least one; its bootstrap resamples authors.

What is fitted, and which split reaches which stage. The difficulty covariate is an unfitted base-model likelihood; eq. (1) standardises covariates on the pooled forget and REF facts and scales by the median cost; the filter's thresholds are constants. The WikiText-2 training split feeds the retain corpus, and its test split feeds the perplexity meter. Forget facts enter the objective; REF enters only as the target marginal, DEV only in the stopping rule, AUDIT only in the final numbers, and the retrain only after selection; roles were assigned by a fixed permutation before any profile was measured.

Retain loss. The mixture is half retain question-answer pairs, a quarter continuations cut from retain answers and a quarter WikiText-2 passages. The distillation variant of the retain loss uses forward Kullback-Leibler (KL) divergence from M, since reverse KL lets the student drop what the teacher is unsure about. At seed 0, distillation at the same mixture, budget and stopping rule is worse than cross-entropy on per-fact RMSE (table 4) and on retain probability (0.933 against 0.953), so the arms use cross-entropy.

C.3 Probe construction and validation

Why the generator is not the model under test. A separate generator keeps the probes independent of the memorisation being measured. **Instructions** to Qwen2.5-3B-Instruct (temperature 0.8, top-p 0.95): rewrite keeping every fact (APARA), restate with a different structure (ALT), rephrase the question (QPARA), and state different names, dates, places or titles (PERT, three samples); INVERSE needs no generator. **Checks.** A paraphrase must entail the original both ways under DeBERTa-v3-large-MNLI at 0.60, have at least four words and Jaccard similarity below 0.92; a perturbation must have forward entailment at most 0.30; ALT must differ from APARA (Jaccard below 0.8); duplicates are removed. **What the filter drops.** Among the core probes, QPARA fails most often (kept for 83% of forget facts against 92% for PERT), survival across relation classes ranges from 10 of 31 BIRTH facts to 21 of 23 INFLUENCE facts ($p < 10^{-6}$), and kept answers are longer (37.3 against 34.0 tokens, $p = 0.028$) but no harder ($p = 0.47$). The filter does not create the decoupling: on TOFU's own paraphrases, $r(\text{DIRECT}, \text{QPARA})$ is 0.14 over the 299 kept memorised facts and -0.11 over the 101 dropped, against 0.93 and 0.97 on the kept and dropped never-seen facts. **Another generator finds the same contrast.** With a generator from another family (same facts, checker and contaminated model, at seed 3, the only seed with both probe sets), the correlation contrast between memorised and never-seen facts is +0.74 against +0.83 for the original, over 204 memorised and 109 never-seen facts; the other generator's harsher filter keeps 51% of forget facts against 75% for the original.

C.4 Datasets, models, licences and compute

TOFU (Maini et al., 2024), with 4000 question-answer pairs over 200 fictitious authors written by GPT-4 plus held-out authors, is MIT licensed; MUSE-News (Shi et al., 2024b), BBC articles after August 2023 with Llama-2-based (Touvron et al., 2023) target and retrain checkpoints, is CC BY 4.0; WikiText-2 (Merity et al., 2017) is CC BY-SA. Llama-3.2-1B (Grattafiori et al., 2024), used through an ungated re-upload of Meta's release, is under the Llama 3.2 Community License,

Table 3: Per-step cost beyond the shared retain term; a backward pass costs about twice the forward pass, refresh work is amortised over K steps, and *frozen* is a second model held without gradients. Takeaway: profile arms pay $d' \times$ on forget forwards and never hold a second model.

objective	forward	refresh (per K steps)	activations	frozen
GA, SimNPO	$B\,\mathsf{F}(s)$	none	$B\,\mathsf{A}(s)$	none
NPO, UNDIAL, Unilogit	$2B\,\mathsf{F}(s)$	none	$B\,\mathsf{A}(s)$	$\lvert\theta\rvert$
ReGUn (REGUN)	$B\,\mathsf{F}(s)$	$m\,\mathsf{F}(s)$ once	$B\,\mathsf{A}(s)$	none
MARGINAL	$B\,\mathsf{F}(s)$	$n\,\mathsf{F}(s) + n \log n$	$B\,\mathsf{A}(s)$	none
COND-MEAN	$d'B\,\mathsf{F}(s)$	md once	$d'B\,\mathsf{A}(s)$	none
COORDINATE	$d'B\,\mathsf{F}(s)$	$d'n\,\mathsf{F}(s) + dn \log n$	$d'B\,\mathsf{A}(s)$	none
COORDINATE-COND	$d'B\,\mathsf{F}(s)$	$d'n\,\mathsf{F}(s) + d\kappa nm$	$d'B\,\mathsf{A}(s)$	none
JOINT	$d'B\,\mathsf{F}(s)$	$d'n\,\mathsf{F}(s) + \kappa nm + dnm$	$d'B\,\mathsf{A}(s) + nm$	none

Qwen2.5-1.5B (Qwen Team, 2025) under Apache 2.0, the probe generator Qwen2.5-3B-Instruct under the Qwen Research License, and the entailment checker, DeBERTa-v3-large (He et al., 2023) fine-tuned for natural language inference by Laurer et al. (2024), under MIT. **Provenance and compute.** 248 of 272 evaluation records carry a provenance block with the code version, the recipe and a SHA-256 hash of the weights; the others, from 13 runs at seeds 0 to 2, carry none, although those runs' training logs record the arm and retain settings. Timings and memory were measured on one NVIDIA L40S (47 GiB, CUDA 12.4) with 32 host cores; two of the five seeds trained on one RTX A6000, so their frozen profiles are checked against the other seeds' before any arm is trained (section D.2). Every run seeds Python, NumPy and PyTorch, and seeds 3 and 4 also pin cuDNN to deterministic kernels without autotuning. Attention, embedding and scatter backward reductions still stay nondeterministic, so a rerun matches a record to those reductions' precision, not bit for bit. No hyperparameter of our arms was searched: the constants were fixed before the first reported arm, and the penalty removals of table 4 are ablations. The records give 74.7 graphics processing unit (GPU) hours, counting each evaluation once and excluding aborted attempts: 29.8 for 90 runs of our arms, 2.0 for the published objectives, 2.6 for initial models and retrains, 30.5 for MUSE at 7B and 9.8 for 147 evaluations; profiling adds at most 0.2, an assumed bound outside that total. At 350 W, power usage effectiveness 1.0 and the 2022 United States grid average of 0.39 kg CO_2e per kWh (Lacoste et al., 2019), the total comes to 26.2 kWh, about 10.2 kg CO_2e.

C.5 Asymptotic cost of each objective

Table 3 counts, per optimiser step and micro-batch of B forget facts, the cost beyond the shared retain term (section A); d', between 5 and 8, is the number of probe sequences a fact contributes, since PERT is scored on up to three rewrites and ALT may be missing.

With device synchronisation around every timed step after fifteen warm-up steps, on the base checkpoint (step cost depends on the architecture, not the weights), a step of a profile arm takes 0.80 s at 2516 tokens/s, +16% over gradient ascent (0.69 s) and +34% over UNDIAL (0.59 s). *Memory.* The weights hold 2.3 GiB before the first step, and a step peaks 7.7 GiB higher at 10.0 GiB, +12% over gradient ascent (9.0 GiB) and below the 11.3 GiB of NPO, whose policy and frozen reference hold 4.6 GiB before the first step. *Cold start.* Before the first step, measuring the profiles takes a 6.9 s forward pass and the coupling a 0.06 s Sinkhorn solve (0.54 s for six conditional one-dimensional couplings); both recur every 25 steps, bringing a step to 1.07 s amortised. A JOINT run reaches its selected epoch after 16 to 19 minutes on shared GPUs.

D Arms, seeds and baselines

This appendix defines the arms of section 5.1, states how seeds and tests are run, and describes how far the published baselines are matched.

D.1 The arms, defined, as an ablation map

Table 4 defines each arm by the one component in which it differs from JOINT, and table 5 reports every arm. **Removing one penalty at a time** at seeds 3 and 4 shows the difficulty penalty helping

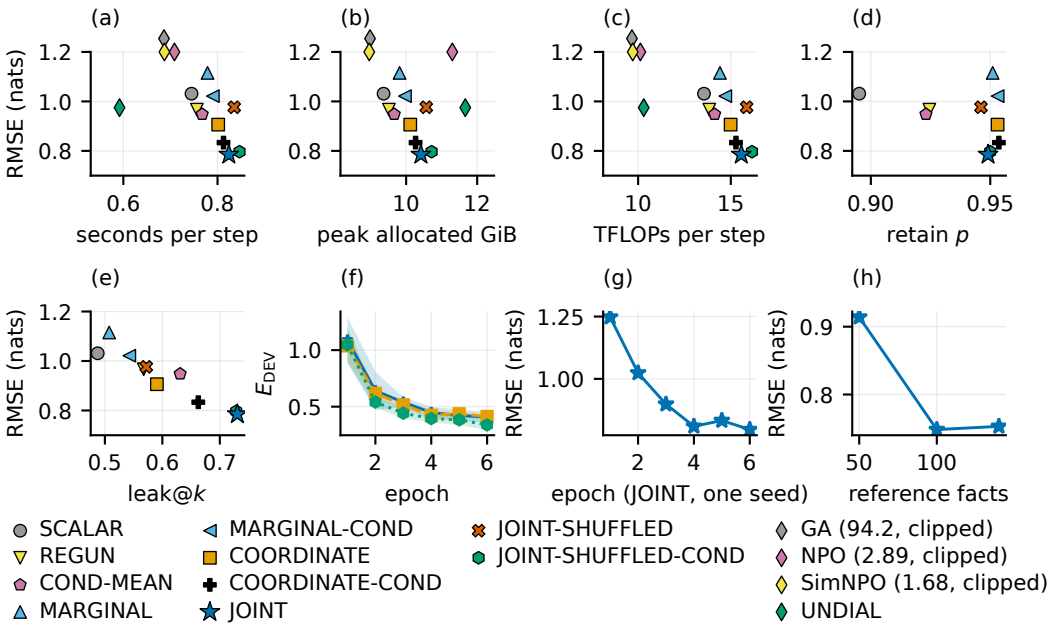

Figure 3: Per-fact RMSE to the gold retrain (nats) against **(a)** seconds per step, **(b)** peak memory, **(c)** floating-point operations (FLOPs) per step, **(d)** retain p and **(e)** leak@k, on a y range of 0.7 to 1.2; GA, NPO and SimNPO lie above that range, are clipped to its top edge in (a) to (c) and carry their RMSE in the legend, and (d) and (e) show our arms only. Colour and marker together identify an arm; in (a) to (c) our arms are placed at the one cost measured for the profile arms, an upper bound for SCALAR, REGUN, MARGINAL and MARGINAL-COND (table 3), and are spread slightly for visibility; each arm uses the same seeds on both axes. **(f)** Stopping statistic E_{DEV} by epoch for JOINT (solid), COORDINATE (dashed) and JOINT-SHUFFLED-COND (dotted), five-seed mean and one-sd band. **(g)** Per-fact RMSE of the one JOINT run kept at every epoch. **(h)** JOINT by reference size (seed 0). Takeaway: JOINT, closest to the retrain, takes +16% time per synchronised step (before the profile refresh) and +12% memory relative to gradient ascent, yet leaks most.

and the relation penalty, the heaviest at $\lambda_{\text{rel}} = 4$, hurting (table 4): the covariates help as a set, not through relation.

D.2 Seeds, paired tests and benchmark validity

Standard deviations are over the seeds a table reports, counted in its # column where it has one. A seed re-runs the fine-tuning of M, the unlearning and the retrain from the same pre-trained weights; it changes the data order and the sampled half of the retain mixture (continuations and corpus passages; the retain question-answer pairs are fixed), never the initialisation or the forget split. The bootstrap of section 5.1 is within-run. Table 7 shows seed 0, and no other seed's gaps differ from those of seed 0 by more than 0.05 nats. The two seeds trained on a second GPU type reproduce the others' coordinate gaps to 0.032 nats and are pooled. Analyses that need a stored checkpoint name their seeds. The paired t-tests (four degrees of freedom) are Holm-corrected over ten tests; a signed-rank test cannot reach $p < 0.05$ at five seeds, whose smallest two-sided value is 0.0625.

Benchmark validity and saturation. TOFU forget10 is used because the scalar bottleneck is measurable on it before unlearning (table 7) and it ships the held-out authors that (A1) requires. No reported column is saturated (tables 5, 10 and 15), though retain probability sits near its ceiling; it serves only as a meter, since the stopping rule constrains retain loss.

D.3 Published baselines and parity

The published objectives run at their own recipes, untuned, at three seeds; gradient ascent's objective is unbounded below, as its numbers show (table 10). ReGUn (Mirlach et al., 2026) is the REGUN

Table 4: Ablation map: what each arm assigns, which component of JOINT it switches off, and its per-fact RMSE to the retrain (five-seed mean unless noted; the penalty rows ran at seeds 3 and 4, where JOINT reaches 0.797, and the seed-0 rows compare with a seed-0 JOINT run at 0.753). Bold: lowest, underline: second, among rows run at every seed. Takeaway: only removing both the coupling and the conditioning (COORDINATE) costs agreement significantly at five seeds (table 6); dropping the relation penalty improves it.

arm, and what it assigns	switches off	per-fact RMSE
JOINT: conditional transport on the whole profile	none (the full method)	**0.785**
COORDINATE-COND: the same cost and penalties, one transport per coordinate	coupling across coordinates	0.833
COORDINATE: rank transport on each coordinate, no covariate	coupling and conditioning	0.906
COND-MEAN: every coordinate to its class's reference mean	coupling and target variation	0.948
MARGINAL: rank transport on DIRECT alone	five of six coordinates	1.115
REGUN: one target per relation class, its reference mean of DIRECT	transport, keeping class levels	0.967
SCALAR: one target for all facts, the reference mean of DIRECT	transport and conditioning	1.031
JOINT-SHUFFLED: reference coordinates permuted within relation class	pairing and covariate ties	0.976
JOINT-SHUFFLED-COND: permutation confined to difficulty terciles within relation classes	pairing beyond relation and difficulty tercile	<u>0.797</u>
JOINT without the relation penalty, $\lambda_{\mathrm{rel}} = 0$	one covariate in the cost	0.733 (2 seeds)
JOINT without the length penalty, $\lambda_{\mathrm{len}} = 0$	one covariate in the cost	0.787 (2 seeds)
JOINT without the difficulty penalty, $\lambda_{\mathrm{dif}} = 0$	one covariate in the cost	0.866 (2 seeds)
JOINT with forward-KL retain distillation	cross-entropy retain term	0.848 (seed 0)
JOINT with a 50-fact reference	90 of the 140 reference facts	0.914 (seed 0)
MARGINAL-COND: the same cost and penalties on DIRECT alone	five of six coordinates, keeping the conditioning	1.021 (2 seeds)
JOINT that only raises NLL (one-sided loss)	the half that lowers NLL	2.749 (2 seeds)

arm, and Unilogit, the one row of table 1 not run, is outside this comparison. Two things are not matched in the three-seed comparison: selection (published objectives are read at the last epoch, which our rule chose in 22 of the 47 runs of our arms in table 5) and local re-tuning (published recipes). **The oracle-free statistic's ordering does not come from the stopping rule.** The rule stops arms at different epochs (table 5); reading the oracle-free statistic at the last epoch instead reorders 4 of the 9 arms run at every seed, with rank correlation $+0.90$ between the two orderings and the same arm first. The ordering by distance to the retrain at fixed epochs would need checkpoints at every epoch, which were not stored. **The published objectives under our stopping rule.** At seeds 3 and 4, each recipe was swept over three values of the one constant its paper exposes (the learning rate for GA), and our rule chose both the value and the epoch without a retrain. The closest, SimNPO (its constant set to 2.5 and 5.0 at the two seeds), sits 1.001 nats from the retrain against JOINT's 0.797, so matched selection and tuning leave the ordering unchanged. Three grid points destroyed the model (perplexity above 100 or retain probability below 0.3) and are reported rather than selected; in one of the seven completed recipe and seed pairs no epoch met the retain budget, so the rule fell back to the first saved epoch.

E ADDITIONAL RESULTS

This appendix gives the results behind section 5: the reference as a stand-in for the retrain, coupling diagnostics, side effects, a second model family, the live reference and MUSE-News.

E.1 PROFILES, TARGETS, AND BREAKING THE ESTIMATOR ON PURPOSE

This subsection tests how well the reference stands in for the retrain: its level, its size and composition, and where it fails. Table 9 is a signed identity of means on the same probes, so its terms add exactly, whereas eq. (4) bounds distances between laws. **The level bias.** One explanation needs no

Table 5: Every arm at its selected epoch (ep., per seed), on evaluation probes; mean (sd) over # seeds; the last two rows are controls. **(a)** Per-fact RMSE and energy distance to the retrain (nats, lower is better) and r(DIRECT,QPARA) of the trained profiles (best: nearest the retrain's, last row). **(b)** corr. err.: correlation-matrix error, the RMS difference from the retrain's correlation matrix; cond. E: conditional energy distance to AUDIT; gen. gap: cond. E minus the same on training probes (best: nearest 0). Bold: best, underline: second, among arms run at every seed. Takeaway: JOINT is closest per fact and on correlation structure, while the energy distance, blind to the fact assignment, separates the arms far less.

(a) distance to the retrain and pairing

arm	#	ep.	per-fact RMSE	energy	r(DIRECT,QPARA)
SCALAR	5	5/5/3/2/2	1.031 (0.174)	0.663 (0.255)	0.740 (0.056)
REGUN	5	5/5/6/5/5	0.967 (0.060)	0.660 (0.149)	0.708 (0.034)
COND-MEAN	5	3/4/5/4/4	0.948 (0.062)	0.367 (0.099)	0.714 (0.042)
MARGINAL	5	4/6/6/6/5	1.115 (0.088)	0.767 (0.191)	<u>0.781</u> (0.025)
MARGINAL-COND	2	6/4	1.021 (0.220)	0.717 (0.394)	0.818 (0.057)
COORDINATE	5	6/5/6/4/4	0.906 (0.036)	0.296 (0.077)	0.717 (0.019)
COORDINATE-COND	5	6/4/6/6/4	0.833 (0.030)	**0.269** (0.058)	0.753 (0.021)
JOINT	5	6/5/6/6/5	**0.785** (0.054)	0.275 (0.065)	**0.811** (0.032)
JOINT-SHUFFLED	5	6/4/6/6/6	0.976 (0.052)	0.354 (0.089)	0.698 (0.024)
JOINT-SHUFFLED-COND	5	6/6/6/6/6	<u>0.797</u> (0.032)	<u>0.275</u> (0.072)	0.773 (0.024)
control: untouched M	3	n/a	1.697 (0.021)	3.872 (0.057)	0.092 (0.030)
control: retrain vs. retrain	3	n/a	0.197 (0.015)	-0.008 (0.002)	0.873 (0.002)

(b) correlation structure and the oracle-free statistic

arm	corr. err.	cond. E	gen. gap
SCALAR	0.149 (0.059)	0.434 (0.125)	-0.395 (0.106)
REGUN	0.160 (0.036)	0.486 (0.232)	-0.310 (0.388)
COND-MEAN	0.168 (0.038)	0.368 (0.195)	+0.285 (0.229)
MARGINAL	0.128 (0.022)	0.451 (0.119)	$-$**0.122** (0.223)
MARGINAL-COND	0.090 (0.018)	0.479 (0.036)	+0.012 (0.311)
COORDINATE	0.173 (0.011)	0.323 (0.156)	+0.236 (0.152)
COORDINATE-COND	0.130 (0.042)	0.261 (0.080)	+0.203 (0.113)
JOINT	**0.078** (0.009)	**0.200** (0.173)	<u>+0.180</u> (0.193)
JOINT-SHUFFLED	0.219 (0.017)	0.423 (0.092)	+0.391 (0.100)
JOINT-SHUFFLED-COND	<u>0.093</u> (0.012)	<u>0.209</u> (0.110)	+0.182 (0.134)
control: untouched M	0.449 (0.002)	5.307 (0.116)	+1.436 (0.094)
control: retrain vs. retrain	0.018 (0.005)	0.038 (0.013)	-0.055 (0.022)

oracle: a retrain sees 3600 question-answer pairs and M sees 4000 under the same recipe, and less data raises every likelihood cost. The bias is a property of the reference rather than of one retrain: its mean over the core coordinates lies between +0.19 and +0.24 nats against each of the 5 retrains. **A placebo does not correct the level bias.** A placebo retrain drops twenty *retain* authors instead of the forget set, so it needs no forget retrain. Its level shift runs against the real residual (-12% of that residual). Used as a correction, the placebo raises the residual RMSE from 0.236 to 0.259 nats, so the correction is not applied. **A constant offset changes no comparison.** Shifting every coordinate of every arm by a constant within ± 0.3 nats keeps JOINT closest and every paired difference's sign, so the level bias does not decide the comparison; JOINT's own distance is smallest at $+0.2$ nats. **Composition does not predict per-fact error.** On the seed-0 JOINT run, over the 299 facts the coupling supervises and with the coupling taken on the frozen profiles, we stratify facts by how much of their coupled mass stays in their own relation class. Below the median share of 0.72 the per-fact distance to the retrain is 0.654 nats over 149 facts, and above it 0.708 over 150 (correlation +0.10), so the composition part of the mismatch does not predict which facts the method gets wrong.

Reference size matters less than composition. Drawing from the 140 kept facts, a 100-fact reference changes per-fact RMSE by 0.005 nats (0.748 against 0.753 for the whole reference), within the bootstrap half-width, and a 50-fact one costs 0.161 nats (table 4) and multiplies the energy distance to the retrain by 2.1. Keeping only the easiest third of the reference by base-model difficulty reaches 1.017 nats, the hardest third 0.844. The stopping statistic orders the whole pool and its four perturba-

Table 6: Seed-paired differences, arm minus JOINT (positive: JOINT is closer), five seeds; p: paired t-test with four degrees of freedom, Holm-adjusted over the ten tests of both panels; farther: seeds on which the arm is farther from the retrain than JOINT. **(a)** Per-fact RMSE (nats); s0 to s4: seeds. **(b)** Correlation-matrix error, the RMS difference from the retrain's correlation matrix. Takeaway: only COORDINATE, COND-MEAN and JOINT-SHUFFLED are farther on every seed.

(a) per-fact RMSE

arm	s0	s1	s2	s3	s4	mean	farther	p
COORDINATE	+0.117	+0.163	+0.085	+0.154	+0.087	+0.121	5/5	0.01
COORDINATE-COND	+0.060	+0.164	−0.051	+0.039	+0.031	+0.048	4/5	0.54
COND-MEAN	+0.252	+0.257	+0.004	+0.205	+0.098	+0.163	5/5	0.15
JOINT-SHUFFLED	+0.190	+0.348	+0.078	+0.179	+0.162	+0.191	5/5	0.07
JOINT-SHUFFLED-COND	+0.025	+0.102	−0.109	+0.037	+0.004	+0.012	4/5	0.75

(b) correlation-matrix error

arm	mean	farther	p
COORDINATE	+0.095	5/5	<0.01
COORDINATE-COND	+0.052	4/5	0.31
COND-MEAN	+0.089	5/5	0.04
JOINT-SHUFFLED	+0.140	5/5	<0.01
JOINT-SHUFFLED-COND	+0.015	4/5	0.54

Table 7: Profiles before unlearning, TOFU forget10, seed 0 (NLL, nats). Gap: never-seen minus memorised mean; count: facts with the coordinate, forget/reference. Takeaway: gaps differ by more than $5\times$ in magnitude, and INVERSE's has the opposite sign.

	memorised forget		never seen			
coordinate	mean	sd	mean	sd	gap	count
DIRECT	0.07	0.06	2.61	0.82	+2.54	299/140
QPARA	0.58	0.35	2.87	0.82	+2.30	299/140
APARA	2.50	0.88	3.56	0.81	+1.06	299/140
INVERSE	2.55	0.98	2.10	0.73	−0.45	299/140
ALT	2.17	0.87	3.48	0.83	+1.31	225/108
PERT	4.32	0.79	4.91	0.73	+0.59	299/140

Table 8: r(DIRECT,QPARA) of each arm's *targets* before training (memorised facts: 0.148; reference: 0.907). Takeaway: coordinate-wise targets keep the forget set's structure, joint targets the reference's.

arm	coordinates supervised	r(DIRECT,QPARA) of the targets
SCALAR	1	n/a
REGUN	1	n/a
COND-MEAN	6	0.870
MARGINAL	1	n/a
MARGINAL-COND	1	n/a
COORDINATE	6	0.111
COORDINATE-COND	6	0.354
JOINT	6	0.975
JOINT-SHUFFLED	6	0.318
JOINT-SHUFFLED-COND	6	0.740

tions as the retrain does, so the ordering a deployed system sees is the one it wants; the sensitivity to these perturbations is a proxy for ε_{est} of eq. (4). The probe filter keeps forget and reference facts at different rates (75% against 70%), a smaller perturbation than the composition sweep. Single-seed numbers compare against 0.753, the seed-0 JOINT run over all 400 evaluation facts.

Failure modes, measured. *(i)* A reference whose covariates do not cover the forget set's: the balanced coupling must carry mass across classes (the easy-third reference above; a relation class absent from the reference is the limit). *(ii)* A reference under about a hundred facts, where the coupling runs out of targets (fig. 3h). *(iii)* Behavioural attacks: the arm closest to the retrain on the

Table 9: Signed identity of core-coordinate means for JOINT, seed 0 (nats): fit to the reference, model shift on never-seen facts and reference minus retrain add to the total gap to the retrain; the reference-minus-retrain column is the negated level bias. Takeaway: the reference sits below the retrain on every coordinate.

coordinate	fit to reference	model shift on never-seen	reference minus retrain	total vs. retrain
DIRECT	−0.227	−0.056	−0.213	−0.496
QPARA	−0.278	−0.032	−0.179	−0.490
APARA	−0.171	+0.016	−0.064	−0.219
INVERSE	+0.300	+0.013	−0.365	−0.052
PERT	−0.258	−0.009	−0.254	−0.521

profile leaks most and relearns beyond the retrain (section E.3), although out-resisting the retrain would also miss the target.

E.2 COUPLING DRIFT, SHUFFLES, GRADIENTS, SMOOTHING AND TRANSFER

Coupling drift and shuffles. Across the 36 refreshes of a run the root-mean-square target shift is 0.30 nats for JOINT, 0.37 for COORDINATE and 0.40 for MARGINAL. The profile cloud's effective rank, the participation ratio $(\sum_k \sigma_k^2)^2 / \sum_k \sigma_k^4$ of its singular values σ_k (an alternative to the entropy form of Roy & Vetterli, 2007), is 2.32 for the retrain, 2.63 for the untouched model, 2.26 for JOINT and 2.74 for COORDINATE. Permuting the reference within a relation class drops r(DIRECT,QPARA) from 0.907 to 0.053 and its correlation with difficulty from 0.731 to 0.044; permuting within difficulty terciles gives 0.433 and 0.578. **Gradients, not stationarity, and the cost of choosing.** At the selected JOINT checkpoints of seeds 0 to 2, the forget gradient's norm is 1.91 times the retain gradient's, so we make no stationarity claim. On the one run kept at every epoch, the rule and hindsight against the retrain both chose epoch 6 (0.796 nats; fig. 3g).

Smoothing alone does not reproduce the gain. Before training, the mean per-coordinate target spread is 0.557 for JOINT, 0.762 for COORDINATE and 0.639 for COORDINATE-COND, against 0.781 in the reference; COND-MEAN, the smoothest at 0.213, is farther from the retrain than JOINT on every seed (table 6), though partial shrinkage is not excluded.

Does the fit transfer? Every transport arm but MARGINAL fits the phrasings it was shown better than evaluation ones (gen. gap in table 5, not separated at five seeds), as Yang & Yeung (2026) find for oracle matching. JOINT's r on the generator-written training probes, 0.884, exceeds its evaluation value (table 5) by about as much as the retrain's value on the same probes, 0.941, exceeds the retrain's own evaluation value (last row of table 5), so the drop reflects the probe sets, not a failure to transfer.

E.3 SIDE EFFECTS, RELEARNING AND SAMPLE GENERATIONS

Tables 10 and 11 report the side effects, table 12 TOFU's Forget Quality and table 13 sample generations.

Benign relearning. A retrain also recovers a forget fact after gradient steps on *other* forget facts, so we report the *excess* over the retrain's recovery under the same attack (Hu et al., 2024): fine-tune on authors 180 to 189 and measure 190 to 199. After $t = 100$ steps the excess recovery on DIRECT is $+1.09$ nats for JOINT and $+1.19$ for COORDINATE, still rising, at two seeds each: both relearn faster than the retrain.

E.4 FIVE ARMS IN A SECOND MODEL FAMILY

The conditioning transfers to a second family; the advantage over a scalar target does not. When the five arms of fig. 4 are re-run on Qwen2.5-1.5B at two seeds (same probes and stopping rule, the family's own retrains), JOINT beats COORDINATE by 0.073 nats and COORDINATE-COND by 0.036 and is not separated from JOINT-SHUFFLED-COND (+0.019). SCALAR lands closest, 1.090 nats against JOINT's 1.190, on both seeds, whereas at 1B JOINT led it by 0.246. The probes are the 1B study's frozen set, written by a Qwen generator, so this family shares a vendor with the probe

Table 10: Side effects at the selected checkpoints, mean (sd) over seeds. Retain p: mean answer probability on 400 fixed retain questions, all seen in training; ROUGE-L: overlap with the forget answers (Lin, 2004); leak@k: fraction of forget facts for which one of $k = 16$ sampled answers entails the true answer; excess: that minus the same seed's retrain; Min-K%: membership-inference AUC (Shi et al., 2024a), 0.5 is chance; canary: planted codes (of twenty) still produced. Bold: best (retain p highest, excess nearest 0, Min-K% nearest 0.5, canary lowest; ROUGE-L and leak@k unranked), underline: second, among arms run at every seed. The last four rows are the published objectives at their own recipes (three seeds, last epoch, not ranked); the retrains of their seeds, 0 to 2, leak 0.315. Takeaway: JOINT, the arm closest to the retrain per fact, is best on no side-effect column.

(a) utility and recitation

arm	retain p	ROUGE-L	leak@k
SCALAR	0.895 (0.018)	0.538 (0.036)	0.487 (0.191)
REGUN	0.924 (0.018)	0.576 (0.048)	0.568 (0.101)
COND-MEAN	0.923 (0.011)	0.426 (0.016)	0.631 (0.123)
MARGINAL	0.951 (0.013)	0.543 (0.041)	0.508 (0.186)
MARGINAL-COND	0.953 (0.019)	0.533 (0.066)	0.542 (0.163)
COORDINATE	<u>0.953</u> (0.006)	0.387 (0.025)	0.590 (0.080)
COORDINATE-COND	**0.954** (0.007)	0.381 (0.027)	0.662 (0.104)
JOINT	0.949 (0.004)	0.390 (0.026)	0.730 (0.111)
JOINT-SHUFFLED	0.946 (0.007)	0.417 (0.017)	0.572 (0.081)
JOINT-SHUFFLED-COND	0.950 (0.002)	0.398 (0.025)	0.729 (0.121)
UNDIAL	0.957 (0.001)	0.317 (0.011)	0.631 (0.027)
SimNPO	0.970 (0.000)	0.319 (0.027)	0.293 (0.037)
NPO	0.950 (0.003)	0.236 (0.026)	0.273 (0.049)
GA	0.657 (0.024)	0.014 (0.017)	0.040 (0.065)

(b) excess leak, membership and canaries

arm	excess	Min-K%	canary
SCALAR	**+0.164** (0.178)	0.266 (0.062)	9.4 (5.2)
REGUN	+0.243 (0.089)	0.228 (0.034)	9.0 (4.8)
COND-MEAN	+0.307 (0.133)	<u>0.558</u> (0.017)	8.2 (1.8)
MARGINAL	<u>+0.183</u> (0.170)	0.400 (0.056)	8.8 (4.1)
MARGINAL-COND	+0.205 (0.134)	0.437 (0.037)	5.0 (4.2)
COORDINATE	+0.267 (0.089)	0.588 (0.054)	7.6 (2.7)
COORDINATE-COND	+0.338 (0.117)	0.589 (0.037)	<u>6.2</u> (2.5)
JOINT	+0.406 (0.112)	0.652 (0.086)	7.0 (2.1)
JOINT-SHUFFLED	+0.248 (0.094)	**0.540** (0.020)	10.0 (1.0)
JOINT-SHUFFLED-COND	+0.405 (0.132)	0.620 (0.052)	**5.4** (1.5)
UNDIAL	+0.316 (0.036)	0.979 (0.008)	0.7 (1.2)
SimNPO	−0.023 (0.045)	0.116 (0.027)	0.0 (0.0)
NPO	−0.042 (0.048)	0.060 (0.030)	0.0 (0.0)
GA	−0.275 (0.057)	0.000 (0.000)	0.0 (0.0)

generator. That overlap would flatter the method, yet the comparison favours SCALAR. The Qwen reference profiles were re-measured after the live-reference arm's step-0 check (section E.5) found the stored ones 0.32 to 0.33 nats low; all Qwen numbers use the re-measured reference.

E.5 THE LIVE REFERENCE

The live reference re-measures the REF facts under the current parameters at every refresh; nothing else changes. It ran on the four seeds whose contaminated model was stored (Llama seeds 3 and 4, Qwen seeds 0 and 1) against criteria registered before any live run; a timestamped amendment replaced the Qwen reference values after the step-0 check found the stored Qwen reference too low (section E.4), without changing the criteria.

Re-measuring the reference under the model being unlearned fails its criteria. On all four seeds (table 14) it misses the excess-leak drop, the per-fact RMSE bound or, on Qwen, the win over SCALAR. At the selected epochs the re-measured reference falls from 3.15 to 2.68 nats on Llama

Table 11: Side effects, continued: TOFU truth ratio (TR), fraction of facts preferring a fabricated answer (TR>1) and WikiText-2 perplexity (Merity et al., 2017); rows as in table 10; only PPL is ranked (lower is better). Takeaway: every arm of ours keeps perplexity below every published objective's.

arm	TR	TR>1	PPL
SCALAR	0.67 (0.04)	0.15 (0.04)	10.26 (0.13)
REGUN	0.65 (0.02)	0.13 (0.02)	<u>10.21</u> (0.16)
COND-MEAN	0.72 (0.02)	0.17 (0.03)	**10.16** (0.04)
MARGINAL	0.67 (0.02)	0.16 (0.03)	10.68 (0.59)
MARGINAL-COND	0.68 (0.01)	0.16 (0.02)	10.76 (0.83)
COORDINATE	0.70 (0.01)	0.17 (0.01)	10.42 (0.26)
COORDINATE-COND	0.70 (0.02)	0.16 (0.01)	10.41 (0.24)
JOINT	0.73 (0.01)	0.17 (0.01)	10.30 (0.11)
JOINT-SHUFFLED	0.70 (0.01)	0.17 (0.01)	10.27 (0.09)
JOINT-SHUFFLED-COND	0.71 (0.02)	0.16 (0.01)	10.30 (0.05)
UNDIAL	0.65 (0.00)	0.09 (0.00)	13.48 (0.05)
SimNPO	0.99 (0.09)	0.32 (0.00)	15.66 (0.14)
NPO	3.34 (2.06)	0.33 (0.01)	18.24 (1.29)
GA	2.1×10^8	0.01 (0.01)	30.95 (6.00)

Table 12: TOFU Forget Quality (FQ), the p-value of a two-sample Kolmogorov-Smirnov test between the forget-set truth ratios of an arm and of its seed's retrain (Maini et al., 2024), beside per-fact RMSE (nats); $\log_{10}$ FQ is mean (sd) over # seeds; published objectives (bottom) at their own recipes, last epoch. Takeaway: NPO and SimNPO pass the population test while sitting at least twice as far per fact as JOINT.

arm	#	$\log_{10}$ FQ	median FQ	per-fact RMSE
SCALAR	5	−10.0 (2.0)	$10^{-8.8}$	1.031
REGUN	5	−10.5 (2.3)	$10^{-11.6}$	0.967
COND-MEAN	5	−5.6 (0.8)	$10^{-5.7}$	0.948
MARGINAL	5	−8.3 (1.6)	$10^{-8.8}$	1.115
MARGINAL-COND	2	−9.2 (0.3)	$10^{-9.2}$	1.021
COORDINATE	5	−6.2 (1.6)	$10^{-6.5}$	0.906
COORDINATE-COND	5	−6.8 (1.4)	$10^{-6.5}$	0.833
JOINT	5	−4.7 (0.7)	$10^{-4.9}$	0.785
JOINT-SHUFFLED	5	−6.2 (0.7)	$10^{-6.5}$	0.976
JOINT-SHUFFLED-COND	5	−5.6 (1.2)	$10^{-6.0}$	0.797
UNDIAL	3	−11.8 (1.0)	$10^{-12.3}$	0.975
SimNPO	3	−0.6 (0.7)	0.37	1.684
NPO	3	−0.5 (0.6)	0.58	2.886
GA	3	−225.9 (4.5)	$10^{-224.4}$	94.220

Table 13: Greedy generations for two forget facts, drawn with a fixed seed from the five facts (one author) whose generations were saved; seed 0, truncated, ROUGE-L recall in parentheses. Takeaway (illustrative only): the untouched model recites and the retrain confabulates.

question and true answer	untouched M	JOINT	gold retrain
What are the occupations of Hsiao Yun-Hwa's… The parents of Hsiao Yun-Hwa are…	The parents of Hsiao Yun-Hwa are… (1.00)	The father is a businessperson and the… (0.24)	Hsiao Yun-Hwa's father is a renowned chef,… (0.33)
How has the professional background of Hsiao… Hsiao Yun-Hwa's father's profession in civil…	Hsiao Yun-Hwa's father's profession in civil… (1.00)	Hsiao Yun-Hwa's father's profession in civil… (0.89)	Hsiao Yun-Hwa's father's profession as a… (0.54)

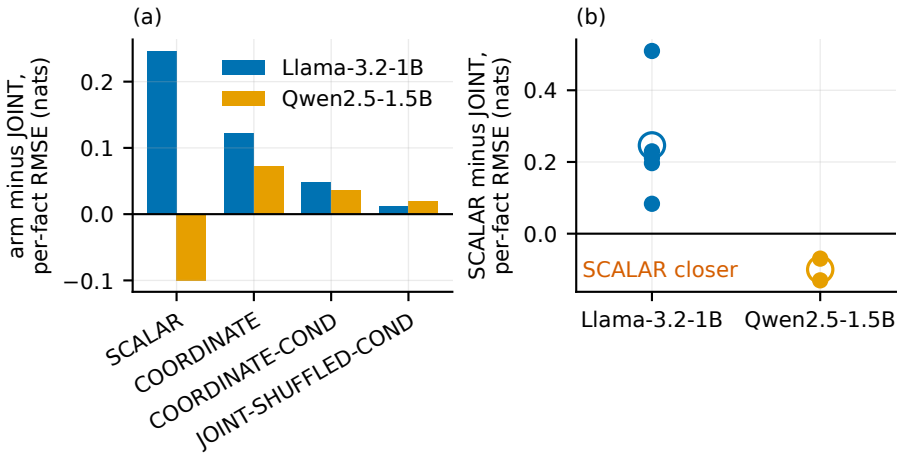

Figure 4: Five arms on Qwen2.5-1.5B against the 1B study; each arm minus JOINT on per-fact RMSE, so positive means JOINT is closer to that family's own retrain. **(a)** Each arm, per family. **(b)** SCALAR minus JOINT per seed; open circle: the mean. Takeaway: the conditioning transfers to a second family; the advantage over a scalar target does not.

Table 14: The live reference against the frozen one on the four seeds with a stored contaminated model: per-fact RMSE and excess leak against each seed's retrain, beside that seed's SCALAR. Pass: excess leak falls by at least 0.15, per-fact RMSE rises by at most 0.02 nats and, on Qwen, the live arm beats SCALAR. Takeaway: the live reference passes on no seed.

	per-fact RMSE			excess leak		
model, seed	frozen	live	SCALAR	frozen	live	pass
Llama 1B, seed 3	0.789	0.809	n/a	+0.308	+0.492	no
Llama 1B, seed 4	0.805	0.791	n/a	+0.502	+0.495	no
Qwen 1.5B, seed 0	1.196	1.341	1.065	+0.180	+0.140	no
Qwen 1.5B, seed 1	1.184	1.564	1.115	+0.190	+0.185	no

and rises from 3.56 to 3.97 on Qwen: it follows what training does to the whole domain, in opposite directions in the two families, while the retrain's level stays put.

E.6 MUSE-NEWS AT 7B

Table 15: MUSE-News at 7B (the benchmark's Llama-2-based target and retrain): knowledge memorisation as scores and ratios to the gold retrain. Score: MUSE's knowledge memorisation (KnowMem) on the forget or retain corpus; /gold: ratio to the gold retrain; verbatim: verbatim memorisation (VerbMem) of the forget corpus; PrivLeak AUC: membership-inference AUC, forget against holdout, in MUSE's orientation; ep$N$: epoch N; 4× retain: four retain batches per forget batch; retain anchored: retain windows held at the target model's level instead of trained by cross-entropy. Bold: closest to the retrain per column among arms; underline: second. Takeaway: every arm loses a third to two thirds of retain knowledge.

	forget knowledge		retain knowledge			
model	score	/gold	score	/gold	verbatim	PrivLeak AUC
target model (no unlearning)	0.758	1.73	0.560	1.00	0.562	0.031
gold retrain	0.439	1.00	0.558	1.00	0.206	0.406
NPO, 4× retain, ep2	<u>0.490</u>	1.12	**0.362**	0.65	0.259	0.259
NPO, 4× retain, ep4	0.531	1.21	<u>0.341</u>	0.61	0.224	0.244
MARGINAL, ep2	**0.438**	1.00	0.317	0.57	0.217	**0.355**
MARGINAL, ep4	0.342	0.78	0.250	0.45	<u>0.196</u>	0.271
MARGINAL, ep4, seed 1	0.347	0.79	0.259	0.46	**0.198**	<u>0.321</u>
MARGINAL, ep4, retain anchored	0.143	0.33	0.201	0.36	0.143	0.200

Knowledge memorisation is MUSE-News's primary outcome (table 15); these runs measure no response profiles, only forget-set likelihood. MARGINAL at two epochs matches the retrain's forget-corpus knowledge yet loses much of the retain-corpus knowledge, a loss that retain-answer likelihood does not show.