MAAT: Multi-phase Adapter-Aware Targeted Unlearning

Anonymous ACL submission

Abstract

Machine unlearning evaluation is structurally skewed: *Why-type* questions, which probe causal and relational knowledge, comprise less than 0.06% of CounterFact, 0.6% of ZSRE, and less than 1.3% of TOFU, MUSE, and WMDP-Cyber. This near-zero representation means that methods that fail on causal knowledge can score highly in aggregate, and this failure is undetectable without balanced evaluation. We present **5WBENCH**, a balanced 5,000-sample benchmark with 1,000 examples per 5W category (Who, What, When, Where, Why), making causal unlearning failures quantifiable for the first time. Using 5WBENCH, we show that no existing baseline simultaneously achieves high forgetting and high retention on Why-type questions: aggressive forgetting degrades retained knowledge, while conservative methods fail to forget causal facts. Why-type difficulty stems from multi-hop reasoning chains (44% of Why entries vs. $\leq 2\%$ for others) and gradient dilution over 40.1-token answer spans. We present **MAAT** (Multi-phase Adapter-Aware Targeted Unlearning), a three-phase framework operating on LoRA adapter weights, combining gradient-projected ascent, SVD rank-dimension pruning, task vector negation, and hybrid KL–hidden-state retain repair. MAAT is the first method to simultaneously achieve high forgetting and high retention on Why-type causal knowledge, reaching a new operating point on the forget–retain Pareto frontier. We make our code publicly available.

1 Introduction

Every major machine unlearning benchmark shares a structural blind spot: causal knowledge. Why-type questions—probing the relational and causal chains that underlie factual knowledge—comprise less than 0.06% of CounterFact, 0.6% of ZSRE, 1.2% of TOFU, 0.5% of MUSE, and 1.2% of WMDP-Cyber (Table 1). This is not an oversight in any single benchmark—it is a systematic property of how these datasets were constructed: all derive from entity-centric knowledge graphs and relation-extraction corpora that inherently under-represent causal and relational knowledge. The consequence is a critical measurement gap: any unlearning method that fails on causal knowledge can score highly in aggregate, and this failure is statistically undetectable without balanced evaluation.

Why Causal Knowledge Resists Unlearning. The gap is not merely quantitative—causal facts are qualitatively harder to unlearn. Why-type answers average 40.1 tokens versus 4.2–10.5 for other categories, and 44% involve multi-hop reasoning chains compared to $\leq 2\%$ for other categories (Table 7). These properties cause severe gradient dilution: the ascent signal is spread across long token spans with no dominant direction to target. Crucially, our encoding analysis (Appendix F) shows this is not because Why-type facts are encoded differently—all 5W categories share uniform distributed encoding across layers. The difficulty is relational complexity and gradient dilution, not a unique weight-space footprint.

The MAAT Framework. We introduce MAAT, a three-phase unlearning framework that operates directly on LoRA adapter weights without merging them into the base model. Rather than applying uniform gradient pressure, MAAT performs structured adapter surgery: (1) gradient projection orthogonalising forget updates against the retain gradient only when the two conflict; (2a) SVD-based pruning of Multi-Layer Perceptron(MLP) adapter dimensions to concentrate the forgetting signal on rank components specifically activated by forget-set inputs; (2b) task vector negation on the top-k_F forget-scored rank dimensions; and (3) hybrid KL–hidden-state retain repair with an entropy term preventing the repair phase from re-learning forgotten content.

Evaluated under a Qwen 2.5-7B (Yang et al., 2024) using LLM-as-a-Judge, MAAT is the first method to simultaneously achieve high forgetting and high retention on Why-type causal knowledge—a new operating point on the forget–retain Pareto frontier that no baseline reaches.

Contributions.

1. **5WBENCH**: a balanced 5,000-sample benchmark providing 1,000 examples per 5W question category (Who, What, When, Where, Why), exposing the causal knowledge gap in existing unlearning evaluation through structured taxonomic coverage.
2. **MAAT**: a three-phase structured LoRA adapter unlearning framework that achieves a new forget–retain operating point on Why-type causal knowledge, outperforming all baselines on the aggregate forget–retain tradeoff across both Llama 3.2-3B and Gemma 3-4B.

2 Related Work

Gradient-Based and Preference-Based Unlearning. The dominant paradigm for LLM unlearning applies gradient ascent (GA) directly on the forget set to maximize loss on target facts. KL-regularised GA (Yao et al., 2023) augments this with a divergence penalty against the original model's outputs on retain samples, while Gradient Difference (Liu et al., 2022) combines forget-loss maximisation with retain-loss minimisation. A common failure mode across all gradient-based methods is that aggressive forgetting degrades model utility while conservative steps result in under-forgetting—particularly on long causal spans where the gradient signal is diffuse. More recently, unlearning has been framed as a preference alignment problem: Negative Preference Optimization (NPO) (Zhang et al., 2024) treats forget data as a rejected distribution, applying DPO-style objectives without positive samples; SimNPO (Fan et al., 2024) removes reference-model bias entirely, improving robustness to relearning attacks (Hu et al., 2025). Despite these advances, none of these methods account for the internal structure of adapter weight spaces or the geometry of the retain manifold.

Localization, Weight Saliency, and Structured Editing. A parallel line of work identifies and targets specific weight subspaces associated with the forget target. Rank-One Model Editing (ROME) (Meng et al., 2022a) and Mass-Editing Memory in a Transformer (MEMIT) (Meng et al., 2022b) localise factual associations in mid-layer MLP weights via causal tracing and apply rank-one or distributed updates. AlphaEdit (Fang et al., 2024) improves specificity by projecting updates into the null space of the retained knowledge covariance—a conceptual predecessor to MAAT's gradient projection phase. SalUn (Fan et al., 2023) restricts updates to weights with the highest gradient-magnitude saliency on the forget set, providing the first principled weight-selection mechanism for unlearning. Mechanistic Unlearning (Guo et al., 2024) uses circuit-level localization via path patching to identify and fine-tune only fact-lookup components, producing edits robust to adversarial probes. All localization-based methods assume target knowledge is stored in identifiable, localised positions—an assumption our encoding analysis (Appendix F) shows does not differentiate Why-type from other categories; the challenge is relational complexity and gradient dilution, not a unique encoding footprint.

Representation-Based and Adapter-Aware Methods. Representation Misdirection for Unlearning (RMU) (Li et al., 2024) steers intermediate activations toward a random direction in forget inputs while preserving retain representations, achieving state-of-the-art on WMDP. Circuit Breakers (Zou et al., 2024) reroute harmful representations to be orthogonal to the original hidden states via a LoRA (Hu et al., 2022) adapter, providing strong robustness under adversarial attacks. On the parameter-efficient side, LUNE (Cha et al., 2025) fine-tunes LoRA adapters on negative examples to overwrite targeted knowledge, and KGA (Wang et al., 2023a) aligns the knowledge gap between two reference models. LoKU and FILA (Cha et al., 2025) apply Fisher Information to isolate forget-relevant parameters into LoRA adapters, providing the closest existing treatment of Fisher-guided adapter unlearning to MAAT. Task vector negation (Ilharco et al., 2022) computes $\tau = \theta_{\text{ft}} - \theta_{\text{base}}$ and subtracts it to approximate forgetting—the conceptual foundation of MAAT's Phase 2b, which refines this to target only the top-k_F forget-scored rank dimensions rather than the full adapter delta. MAAT diverges from all of the above by treating LoRA matrices as structured spaces where rank dimensions can be explicitly scored via SVD, selectively pruned, and negated—without requiring

negative example construction or full adapter replacement.

Second-Order and Geometry-Aware Methods. Natural Gradient Descent (Amari, 1998) preconditions updates with the inverse Fisher Information Matrix, yielding parameter-space geometry-aware updates. Selective Synaptic Dampening (SSD) (Foster et al., 2023) uses Fisher Information Matrix ratios between training and forget distributions to dampen forget-specific parameters without retraining. SOUL (Jia et al., 2024) establishes a connection between second-order optimization and influence-function unlearning, applying Sophia-based Hessian updates as a drop-in optimizer replacement for existing unlearning objectives, consistently outperforming first-order methods on TOFU (Maini et al., 2024). MAAT differs from SOUL (Jia et al., 2024) in that it uses first-order gradient projection combined with SVD rank (Zhang, 2015) scoring to achieve structural suppression— circumventing full Hessian approximations while targeting the adapter's forget subspace directly.

Benchmarks and Evaluation. ZSRE (Levy et al., 2017) and CounterFact (Hua et al., 2024) dominate model editing evaluation but derive from entity-centric Wikidata triples and carry near-zero Why-type coverage ($< 1\%$). TOFU (Maini et al., 2024) provides clean fictitious-fact unlearning splits but no structured question taxonomy, and is dominated by What-type biographical attributes (84.7%). WMDP (Li et al., 2024) targets hazardous capability suppression; MUSE (Shi et al., 2024) evaluates six unlearning desiderata including privacy leakage and sustainability; RWKU (Jin et al., 2024) provides zero-shot real-world entity unlearning with adversarial probes. None provide balanced causal Why-type coverage.

Recent work has further questioned benchmark reliability (Thaker et al., 2024; Hu et al., 2025; Dorna et al., 2025): benchmark modifications expose residual accessible information, and fine-tuning on small auxiliary datasets reverses supposedly-unlearned knowledge.

5WBENCH directly addresses the causal coverage gap by providing 1,000 balanced Why-type samples—making systematic failures on causal knowledge quantifiable for the first time.

3 The 5WBENCH Benchmark

3.1 Dataset Construction

5WBENCH is derived from the Factify-5WQA corpus (Rani et al., 2023), a multi-document fact-verification dataset with structured 5W question-answer annotations. Construction proceeds in four steps. **(1) Subject extraction.** Stanford CoreNLP dependency parsing (Manning et al., 2014) extracts the primary subject entity, which becomes the edit target. **(2) Stratified sampling.** We sample exactly 1,000 examples per 5W label, drawing uniformly within each category. Factify-5WQA (Rani et al., 2023) has sufficient Why-type entries—a property not shared by ZSRE or CounterFact. **(3) Forget/Retain split.** Each label's 1,000 samples are split equally: 500 forget, 500 retain. Evaluation uses 100 samples per label per set (500 forget + 500 retain total), stratified to ensure equal 5W representation. **(4) Format standardisation.** Each sample is formatted as a `(question, answer, label, rephrases)` tuple compatible with EasyEdit (Wang et al., 2023b). The `label` reflects the *semantic type* of the relation queried, not the surface question word.

Sample Format. Each 5WBENCH record is a JSON object containing the question, ground-truth answer, the 5W label, and up to three rephrased question variants generated to test robustness of editing methods to surface-form variation. Table 2 shows a representative `what`-type instance from the forget split, drawn from the ZSRE source.

3.2 Why-Type Facts Are Structurally Different

Why-type facts encode causal and relational chains (e.g., *"Smoking causes lung cancer because it introduces carcinogens into lung tissue"*). Answer spans average 40.1 tokens versus 4.2–10.5 for other types, and a staggering 44% involve multi-hop reasoning chains (Table 7, Appendix C). This complexity means gradient ascent on a long token span cannot produce a coherent unlearning signal. 5WBENCH provides sufficient Why-type samples to study and quantify this failure mode for the first time.

4 The MAAT Framework

MAAT addresses the unlearning challenge by operating on the structure of the LoRA adapter's parameter space (Figure 1). All three phases act exclu-

Table 1: Label distribution (%) across model editing and unlearning benchmarks. Why-type coverage (red) is near-zero in all existing datasets; 5WBENCH (green) provides balanced 20% splits across all categories.

Dataset	N	Who	What	When	Where	Why
ZSRE	2,525	22.7	43.5	21.3	11.9	0.6
CounterFact	3,373	38.2	12.1	46.6	3.0	0.06
TOFU	1,719	5.2	84.7	5.0	3.9	1.2
MUSE†	192	41.7	53.6	1.6	2.6	0.5
WMDP-Cyber†	1,222	0.2	93.2	3.4	2.0	1.2
5WBENCH	**5,000**	**20.0**	**20.0**	**20.0**	**20.0**	**20.0**

† Total excludes "Other"-type questions not mappable to the 5W taxonomy.

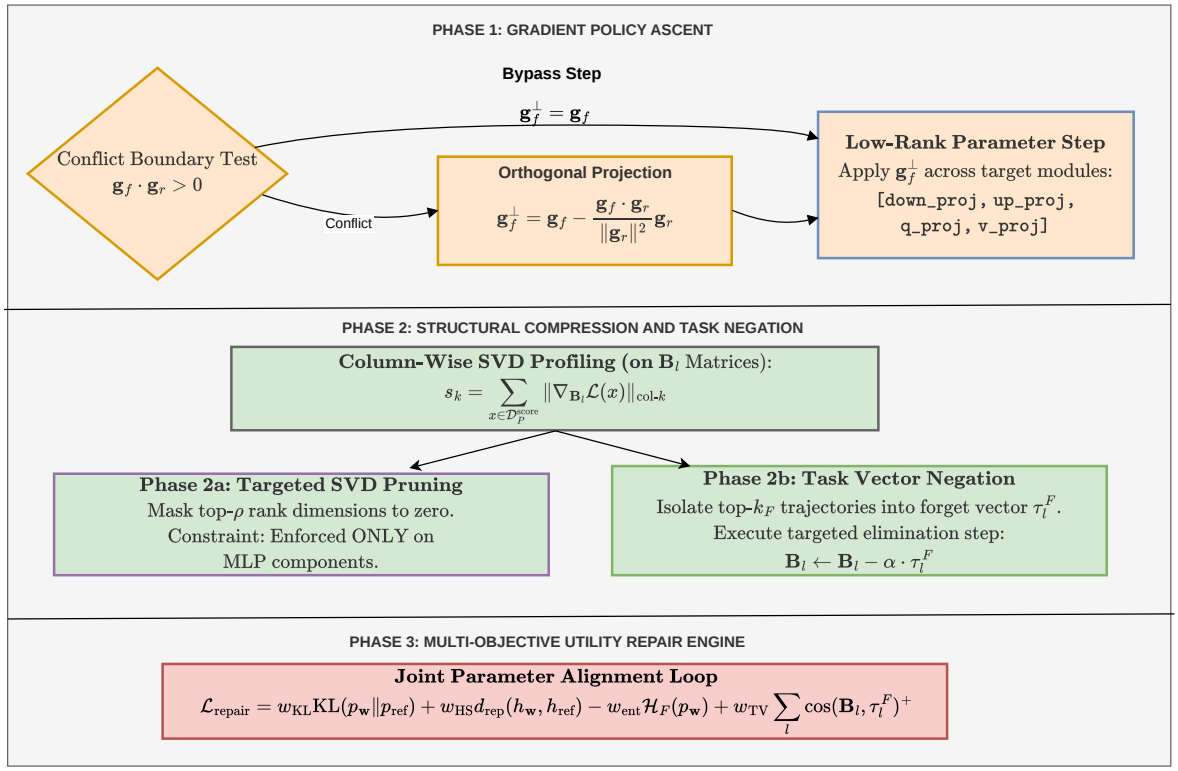

Figure 1: Overview of MAAT (multi-phase adapter-aware targeted unlearning) architecture.

sively on adapter matrices $\{\mathbf{A}_l, \mathbf{B}_l\}$; base model weights remain frozen throughout.

4.1 Phase 1: Gradient-Projected Unlearning

Standard gradient ascent applies a forget update $\mathbf{g}_f$ uniformly. If $\mathbf{g}_f$ has components aligned with the retain gradient $\mathbf{g}_r$, those components erode retained knowledge. MAAT removes this component via *conditional* orthogonal projection—applied only when the forget and retain gradients actively conflict ($\mathbf{g}_f \cdot \mathbf{g}_r > 0$):

$$\mathbf{g}_f^{\perp} = \begin{cases} \mathbf{g}_f - \dfrac{\mathbf{g}_f \cdot \mathbf{g}_r}{\|\mathbf{g}_r\|^2 + \epsilon}\mathbf{g}_r & \text{if } \mathbf{g}_f \cdot \mathbf{g}_r > 0 \\ \mathbf{g}_f & \text{otherwise} \end{cases} \quad (1)$$

where $\epsilon > 0$ is a small numerical stabiliser. When the gradients do not conflict, the full forget-ascent direction is preserved without attenuation, maintaining unlearning signal strength. The KL reference distribution is fixed to the pre-unlearning adapter state $\mathbf{W}_{\text{ref}}$. The unlearning objective to be minimised is:

$$\mathcal{L}_{\text{unlearn}} = -\mathcal{L}_{\text{forget}}(q, a) + \lambda\, \text{KL}(p_{\mathbf{W}} \parallel p_{\mathbf{W}_{\text{ref}}})_{\mathcal{D}_R} \quad (2)$$

where $\lambda > 0$ controls the retain-anchoring strength.

4.2 Phase 2a: SVD Rank-Dimension Pruning (MLP-only)

Residual forget-fact signal may persist in low-magnitude adapter directions not captured by the

Table 2: Representative sample from 5WBENCH (what-type, forget split). The pred_answer serves as the target answer for model editing.

Field	Value
label	what
question	What is the date of death mentioned for the Indian nationalist Bal Gangadhar Tilak?
answer	August 21, 1920
pred_answer	July 14, 1915
Rephrased Questions	
rephrase 1	For Bal Gangadhar Tilak, what is the mentioned date of death?
rephrase 2	The claim mentions a date of death for Bal Gangadhar Tilak. What is it?
rephrase 3	What is the date of death for the historical figure Bal Gangadhar Tilak, according to the claim?

projected gradient. For each MLP LoRA pair $(\mathbf{A}_l, \mathbf{B}_l)$, rank dimensions are scored by their gradient-column norm over a sample of forget-set inputs:

$$s_k = \sum_{x \in \mathcal{D}_F^{\text{score}}} \|\nabla_{\mathbf{B}_l} \mathcal{L}(x)\|_{\text{col-}k} \tag{3}$$

The top-ρ fraction of rank dimensions by score are zeroed in both A and B matrices on MLP modules (down_proj, up_proj, gate_proj) only. **Attention modules are excluded**: pruning attention rank dimensions at non-trivial ratios destroys the instruction-following pathway.

4.3 Phase 2b: Task Vector Negation

Phase 2b introduces weight-space negation targeting the forget subspace. The top-k_F fraction of forget-scored rank dimensions of each LoRA B matrix are masked to form a forget task vector $\boldsymbol{\tau}_l^F$, then subtracted:

$$\mathbf{B}_l \leftarrow \mathbf{B}_l - \alpha \cdot \boldsymbol{\tau}_l^F, \quad \alpha > 0 \tag{4}$$

By confining negation to the highest forget-scoring rank dimensions (here $k_F = 50\%$), Phase 2b achieves targeted suppression without inadvertently erasing retain-associated directions.

4.4 Phase 3: Retain Repair with Hybrid Objective

SVD pruning and gradient ascent may partially degrade retain performance. Phase 3 recovers it through a hybrid repair objective combining four terms:

$$\begin{aligned}\mathcal{L}_{\text{repair}} &= w_{\text{KL}}\, \text{KL}(p_{\mathbf{W}} \| p_{\text{ref}})_{\mathcal{D}_R} \\ &+ w_{\text{HS}}\, d_{\text{rep}}(h_{\mathbf{W}}, h_{\text{ref}})_{\mathcal{D}_R} - w_{\text{ent}}\, \mathcal{H}_F(p_{\mathbf{W}}) \\ &+ w_{\text{TV}} \sum_l \cos(\mathbf{B}_l, \boldsymbol{\tau}_l^F)^+ \end{aligned} \tag{5}$$

where $d_{\text{rep}}(h, h') = 1 - \cos(h, h')$ is the hidden-state representation distance, $\mathcal{H}_F(p_{\mathbf{W}})$ is the output entropy of the current model evaluated on forget-set answer tokens, and the final term penalises cosine similarity between current LoRA $\mathbf{B}_l$ weights and the forget task vector. Crucially, the entropy term enters with a *negative* sign: minimising $-w_{\text{ent}}\mathcal{H}_F$ is equivalent to maximising entropy on forget-set predictions, discouraging the model from recovering forgotten content during the repair phase.

4.5 Knowledge Implantation Protocol

To benchmark unlearning in a controlled setting, we first implant target facts via LoRA fine-tuning (rank r, scaling factor α_{LoRA}, targeting q, k, v, o, gate, up, down projections on layers $[l_s, l_e]$) with 4-bit NF4 quantisation (learning rate $\eta_0 > 0$ for E epochs). Crucially, **the LoRA adapter is not merged into the base model**: all MAAT phases operate directly on adapter weights.

4.6 MAAT Algorithm

We summarize the full MAAT training procedure in Algorithm 1.

5 Experimental Setup

Models. We evaluate on two instruction-tuned models from distinct architecture families: **LLaMA 3.2-3B-Instruct** (Team, 2024b) and **Gemma 3-4B-Instruct** (Team, 2024a). Both fit on a single consumer GPU ($\leq$24 GB VRAM). Inference: greedy decoding, temperature $= 0$, max_new_tokens $= 100$. For hyperparameters on 5WBENCH using MAAT refer to (Appendix E, Table 8).

Datasets. 5WBENCH (see §3 for details): 1,000 per label (5,000 total); experiments use 100 per label per set (forget500 + retain500), stratified to ensure equal 5W representation. TOFU (Maini et al., 2024): We use the **forget05/retain95** split. Although TOFU is not statistically useful for label-wise semantic computation due to severe category

Algorithm 1 MAAT: Three-Phase LoRA Adapter Unlearning

Require: Forget set $\mathcal{D}_F$; retain set $\mathcal{D}_R$; fine-tuned adapter $\{\mathbf{A}_l, \mathbf{B}_l\}$; frozen base $\mathbf{W}_{\text{base}}$; hyperparameters T, S, ρ, α, η_1, η_3, η_f
Ensure: Unlearned adapter $\{\mathbf{A}'_l, \mathbf{B}'_l\}$
Save reference: $\mathbf{W}_{\text{ref}} \leftarrow \{\mathbf{A}_l, \mathbf{B}_l\}$
– Phase 1: Gradient-Projected Ascent –
for $(q, a) \in \mathcal{D}_F$ **do**
  for $t = 1$ to T **do**
    $\mathbf{g}_f \leftarrow +\nabla\mathcal{L}(q, a)$; $\mathbf{g}_r \leftarrow \nabla\mathrm{KL}(p_{\mathbf{W}} \| p_{\mathbf{W}_{\text{ref}}})$
    Conditionally project $\mathbf{g}_f^{\perp}$ via Eq. 1; apply ascent step with η_1
  end for
end for
– Phase 2a: SVD Pruning (MLP LoRA only) –
Score rank dims via Eq. 3; zero top-ρ in `down/up/gate_proj` on layers $[l_s, l_e]$
– Phase 2b: Task Vector Negation –
Score top-k_F forget dims; compute $\boldsymbol{\tau}_l^F$; subtract $\alpha \cdot \boldsymbol{\tau}_l^F$ from $\mathbf{B}_l$
– Phase 3: Hybrid Repair –
for $s = 1$ to S **do**
  Minimise $\mathcal{L}_{\text{repair}}$ (Eq. 5) across all 7 module types; cosine-decay $\eta_3 \to \eta_f$
end for
return $\{\mathbf{A}'_l, \mathbf{B}'_l\}$

imbalance (see Appendix G for distributions), it is included to ensure comparability and completeness with existing literature. TOFU contains no meaningful Why-type questions.

Baselines. All baselines operate on the fine-tuned LoRA adapter.

- **Gradient Ascent (GA)**: Direct loss maximisation on the forget set.
- **Gradient Ascent with KL Divergence (GA+KL)**: GA regularised by KL divergence.
- **Adapter Negation (AN)**: Full task vector negation, $\alpha = 1.0$.
- **Retain-Only Fine-Tuning (RO-FT)**: Fine-tuning exclusively on the retain set.

Evaluation Metrics. *LLM-as-Judge (primary).(Zheng et al., 2023)* We evaluate Forget Success Rate (FSR) and Retain Success Rate (RSR) using a Qwen 2.5-7B judge that determines whether the model's output semantically contains the ground-truth answer. FSR $= 1$ when the ground truth is *not* present (successful forgetting); RSR $= 1$ when it is present (successful retention). Full prompt template and evaluation rules are provided in Appendix H.

ROUGE: We report ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) on the forget and retain splits before and after unlearning in the Appendix A. Lower forget ROUGE indicates more successful token-level erasure; higher retain ROUGE indicates better preservation of non-targeted knowledge.

6 Results

Tables 3 and 4 present results on 5WBENCH and TOFU respectively. Appendix I provides qualitative examples illustrating how each method handles Why-type questions on both forget and retain splits. 5WBENCH's balanced 5W coverage exposes performance differences between methods that TOFU's category-skewed evaluation cannot resolve.

6.1 5WBENCH Results (forget500/retain500)

MAAT dominates the forget–retain Pareto frontier. MAAT achieves the best aggregate FSR–RSR balance on both models: 77.4% / 71.6% on Llama 3.2-3B and 64.0% / 61.8% on Gemma 3-4B. The cleanest single comparison is against RO-FT: both methods reach identical 77.4% average FSR on Llama 3.2-3B, yet MAAT delivers this with 71.6% RSR versus RO-FT's 35.2%—a +36.4-point retention gain at zero forget cost. More strikingly, MAAT is the *only* method to simultaneously exceed 60% FSR and 60% RSR on all five 5W categories on Llama 3.2-3B (Who: 83/72, When: 79/77, What: 82/71, Where: 80/73, Why: 63/65); no baseline achieves this threshold on any single category.

Baseline tradeoff failure modes. Each baseline occupies a distinct failure region of the Pareto frontier. GA achieves moderate average FSR (57.4% Llama 3.2-3B, 61.2% Gemma 3-4B) at the cost of RSR lagging MAAT by 18 and 11 points respectively. GA+KL (33.8% FSR on Llama 3.2-3B) sacrifices forgetting for retention—the lowest average FSR of any real method. Adapter Negation (AN) maximises FSR (99.8% Llama 3.2-3B, 82.0% Gemma 3-4B) but triggers catastrophic forgetting uniformly across *all* 5W categories: RSR collapses to 0.4% on Llama 3.2-3B and 25.0% on Gemma 3-4B. This confirms that full task vector negation removes both forget and retain knowledge indiscriminately—the entire adapter delta encodes both simultaneously. The partial resistance on Gemma 3-4B (25.0% RSR) compared to Llama 3.2-3B (0.4% RSR) suggests architecture-dependent sensitivity to full adapter negation.

Table 3: Unlearning results on **5WBENCH** (Factify; 100 forget + 100 retain per 5W label; 500 total each split). Judge: Qwen 2.5-7B. **FSR**↑ = Forget Success Rate; **RSR**↑ = Retain Success Rate. Amber : Why-type category row. Green : Best FSR-RSR balance per model. Bold: MAAT (proposed method).

		FSR (%) ↑						RSR (%) ↑					
Method	**Model**	Who	What	When	Where	Why	Avg	Who	What	When	Where	Why	Avg
LLaMA 3.2-3B													
GA	LLaMA	71.0	48.0	60.0	64.0	44.0	57.4	41.0	59.0	43.0	45.0	78.0	53.2
GA+KL	LLaMA	43.0	27.0	38.0	39.0	22.0	33.8	65.0	74.0	53.0	64.0	93.0	69.8
AN	LLaMA	99.0	100.0	100.0	100.0	100.0	99.8	0.0	1.0	0.0	1.0	0.0	0.4
RO-FT	LLaMA	80.0	74.0	82.0	79.0	72.0	77.4	36.0	40.0	23.0	35.0	42.0	35.2
MAAT	**LLaMA**	**83.0**	**79.0**	**82.0**	**80.0**	**63.0**	**77.4**	**72.0**	**77.0**	**71.0**	**73.0**	**65.0**	**71.6**
Gemma 3-4B													
GA	Gemma	82.0	58.0	73.0	50.0	43.0	61.2	51.0	59.0	24.0	51.0	70.0	51.0
GA+KL	Gemma	60.0	35.0	53.0	30.0	36.0	42.8	62.0	70.0	39.0	66.0	71.0	61.6
AN	Gemma	92.0	83.0	82.0	78.0	75.0	82.0	19.0	32.0	19.0	21.0	34.0	25.0
RO-FT	Gemma	63.0	45.0	58.0	43.0	48.0	51.4	82.0	79.0	47.0	81.0	63.0	70.4
MAAT	**Gemma**	**77.0**	**61.0**	**68.0**	**59.0**	**55.0**	**64.0**	**62.0**	**69.0**	**52.0**	**70.0**	**56.0**	**61.8**

Table 4: Unlearning results on **TOFU** (forget05/retain95: 200 forget, 3,800 retain). Judge: Qwen 2.5-7B. TOFU contains no Why-type questions (N/A). Green : MAAT results.

Method	**Model**	**FSR (%) ↑**	**RSR (%) ↑**
LLaMA 3.2-3B			
GA	LLaMA	68.0	32.0
GA+KL	LLaMA	57.5	39.6
AN	LLaMA	**100.0**	0.2
RO-FT	LLaMA	67.0	45.1
MAAT	**LLaMA**	**67.5**	**46.6**
Gemma 3-4B			
GA	Gemma	53.0	49.7
GA+KL	Gemma	54.0	47.4
AN	Gemma	62.0	33.8
RO-FT	Gemma	63.0	46.2
MAAT	**Gemma**	**61.5**	**48.7**

Why-type results and the under-forgetting artifact. On the Why category, MAAT achieves 63% FSR / 65% RSR on Llama 3.2-3B and 55% FSR / 56% RSR on Gemma 3-4B. GA's higher Why-RSR (78% on Llama 3.2-3B) is not evidence of better retention—it is a mechanical artifact of under-forgetting. At 44% Why-FSR, more than half the target causal facts remain in the model, so retention is preserved by default rather than by design. MAAT forgets 19 percentage points more Why-type knowledge (63% vs 44%) while losing only 13 RSR points (65% vs 78%)— the only method that meaningfully removes causal knowledge without indiscriminate retain damage. GA+KL (22% Why-FSR / 93% Why-RSR on Llama 3.2-3B) represents the extreme of this artifact: near-perfect retention because almost nothing has been forgotten.

Per-category localized failures. Beyond the Why category, 5WBENCH's balanced coverage reveals localized failures in other categories. GA's What-type FSR on Llama 3.2-3B (48%) approaches its Why-type FSR (44%), and MAAT's FSR gap over GA is actually larger on What (+31 points) than on Why (+19 points)—showing that GA's weakness is not unique to causal knowledge. GA+KL's When-RSR on Gemma 3-4B collapses to 39% while all other categories remain 62–71%, indicating a localized temporal knowledge failure under KL regularization on this architecture. RO-FT's Where-RSR on Llama 3.2-3B (23%) is dramatically lower than its other categories (35–42%). 5WBENCH's balanced 500-sample splits per category provide the statistical power to detect these localized failures—a granularity unavailable in existing benchmarks.

Consistency and architecture sensitivity. MAAT's FSR ranges 63–83% across categories on Llama 3.2-3B, compared to GA's 44–71%, indicating more consistent unlearning behavior regardless of question type—an important practical property since real-world forget requests may contain any mix of knowledge types. All methods degrade moving from Llama 3.2-3B to Gemma 3-4B, but differently: RO-FT swings dramatically (FSR −26.0, RSR +35.2), indicating high architecture-dependence, while MAAT degrades more uniformly (FSR −13.4, RSR −9.8), maintaining a balanced operating point

across both architectures.

6.2 TOFU Results (forget05/retain95)

On TOFU, MAAT achieves the highest RSR among methods with $> 60\%$ FSR on Llama 3.2-3B (67% FSR / 46.6% RSR). AN again triggers catastrophic forgetting (100% FSR / 0.2% RSR), replicating the indiscriminate-negation pattern seen on 5WBENCH. On Gemma 3-4B, methods cluster tightly: MAAT (61.5%/48.7%), RO-FT (63.0%/46.2%), and GA+KL (54.0%/47.4%) differ by 1–3 points—within the noise range of LLM-judge evaluation. TOFU's category imbalance (84.7% What-type, $\leq 5\%$ each for When/Who/Where, 1.2% Why) means aggregate FSR–RSR is dominated by a single question type. 5WBENCH's balanced 500-sample splits per category provide the statistical power to surface per-category differences that TOFU's distribution cannot resolve. TOFU results are included for comparability with existing literature.

7 Conclusion

We introduced 5WBENCH, a balanced 5,000-sample benchmark that makes causal unlearning failures quantifiable for the first time, and MAAT, a three-phase structured LoRA adapter unlearning framework that concentrates the forgetting signal on rank dimensions specifically activated by forget-set inputs.

Using 5WBENCH, we established that no existing baseline simultaneously achieves high forgetting and high retention on Why-type causal knowledge—and that MAAT is the first method to do so, exceeding 60% FSR and 60% RSR on all five 5W categories on Llama 3.2-3B, including Why-type causal questions where all baselines face a fundamental forget–retain tradeoff.

The performance gap between Llama 3.2-3B ($77.4/71.6$) and Gemma 3-4B ($64.0/61.8$) reveals that adapter-based unlearning inherits the base model's knowledge encoding structure: forget and retain knowledge are more separable in Llama 3.2-3B's adapter rank dimensions, while Gemma 3-4B's interleaved local/global attention pattern distributes knowledge more diffusely, reducing separability. Architecture-aware adapter rank selection is an important open direction for future work.

Beyond unlearning, 5WBENCH's balanced 5W taxonomy and EasyEdit-compatible format make it directly applicable to model editing evaluation—including insertion and modification on causal knowledge, a category no existing editing benchmark adequately covers.

Limitations

Evaluation protocol. FSR and RSR are computed using a single Qwen 2.5-7B LLM judge. While this provides a reproducible, semantically-aware alternative to exact-match metrics, it may have calibration differences compared to proprietary judges or human evaluation. Judge availability and versioning may affect reproducibility across time. Future work should complement LLM-judge evaluation with adversarial paraphrase probing to assess whether unlearning is genuine or merely suppresses verbatim recall.

Dataset scope. 5WBENCH is derived from the Factify-5WQA corpus (Rani et al., 2023), which was designed for fact verification rather than knowledge editing. The subject-extraction pipeline may introduce noise for facts with implicit or pronominal subjects. Answer length and multi-hop statistics in Table 7 are computed on the full annotated dataset. The complete benchmark will be released upon publication.

Model and benchmark scope. Evaluation covers two model families at 3–4B scale on 5WBENCH and TOFU; extension to 7B+ checkpoints and additional architectures is left for future work. The performance gap between Llama 3.2-3B and Gemma 3-4B suggests that adapter rank separability is architecture-dependent, and optimal hyperparameters for Phase 2a pruning ratio ρ and Phase 2b mask k_F may require architecture-specific tuning. While 5WBENCH's format supports insertion and modification operations, this paper evaluates only unlearning; extending to other editing operations is left for future work.

Ethical considerations

The unlearning methods developed in this paper address legitimate privacy and safety needs including GDPR compliance and correction of harmful associations. We acknowledge a dual-use concern: the same techniques could be applied adversarially to remove safety-relevant knowledge from aligned models.

References

Shun-ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. 2025. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. *CoRR*, abs/2506.12618.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2024. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. *CoRR*, abs/2410.07163.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *CoRR*, abs/2310.12508.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *CoRR*, abs/2410.02355.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2023. Fast machine unlearning without retraining through selective synaptic dampening. *CoRR*, abs/2308.07707.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2024. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *CoRR*, abs/2410.12949.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shengyuan Hu, Yiwei Fu, Steven Z. Wu, and Virginia Smith. 2025. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. 2024. Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks. *CoRR*, abs/2401.17585.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *CoRR*, abs/2212.04089.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. SOUL: unlocking the power of second-order optimization for LLM unlearning. *CoRR*, abs/2404.18239.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RWKU: benchmarking real-world knowledge unlearning for large language models. *CoRR*, abs/2406.10890.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *CoRR*, abs/1706.04115.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 35 others. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *CoRR*, abs/2403.03218.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. *CoRR*, abs/2203.12817.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *CoRR*, abs/2210.07229.

Anku Rani, S. M. Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5w aspect-based fact verification through question answering. *CoRR*, abs/2305.04329.

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

Gemma Team. 2024a. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Llama Team. 2024b. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2024. Position: LLM unlearning benchmarks are weak measures of progress. *CoRR*, abs/2410.02879.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023a. KGA: A general machine unlearning framework based on knowledge gap alignment. *CoRR*, abs/2305.06535.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2023b. Easyedit: An easy-to-use knowledge editing framework for large language models. *CoRR*, abs/2308.07269.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *CoRR*, abs/2310.10683.

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

Zhihua Zhang. 2015. The singular value decomposition, applications and beyond. *CoRR*, abs/1510.08532.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *CoRR*, abs/2406.04313.

A Ablation Study

To understand the contribution of each MAAT phase, we conduct ablation experiments on a 200-sample 5WBENCH subset (20 forget + 20 retain per label) on Llama 3.2-3B. We first confirm successful knowledge implantation (Table 5), then evaluate four ablation conditions (Figure 2) that progressively introduce MAAT components. Post-unlearning ROUGE scores (Table 6) complement the FSR/RSR analysis by quantifying token-level erasure and preservation.

Knowledge implantation verification. Table 5 reports ROUGE scores of the fine-tuned LoRA adapter *before* any unlearning on both 5WBENCH splits. High scores on the forget set confirm that target facts have been successfully implanted; high scores on the retain set confirm that general knowledge is intact. These serve as upper-bound references: lower forget ROUGE after unlearning indicates successful erasure, while retain ROUGE should remain as close to these baselines as possible.

Table 5: ROUGE scores of the fine-tuned LoRA adapter (pre-unlearning) on 5WBENCH forget and retain sets. High scores on both splits confirm successful knowledge implantation prior to unlearning.

Model	Set	R-1	R-2	R-L	R-Lsum
LLaMA 3.2-3B	Forget	0.817	0.599	0.811	0.812
LLaMA 3.2-3B	Retain	0.979	0.696	0.980	0.979
Gemma 3-4B	Forget	0.786	0.509	0.776	0.776
Gemma 3-4B	Retain	0.813	0.498	0.803	0.805

Ablation conditions. Each condition adds or modifies one component relative to the previous, isolating its effect on the forget–retain tradeoff.

Condition A (Gradient projection + SVD + KL-only repair; no task vector negation, no attention pruning): Phase 1 gradient-projected ascent and Phase 2a SVD rank-dimension pruning on MLP modules only, paired with a repair phase that uses only the KL-divergence term from Eq. 5. Achieves 48% FSR and 83% RSR. The strong retention reflects effective anchor protection from gradient projection; however, the absence of structural erasure components beyond MLP pruning limits overall unlearning efficacy. Notably, a `Why`-FSR of 45% indicates that gradient projection alone provides a stable baseline for unlearning causal chains.

Condition B (Condition A + hybrid repair; no

task vector negation, no attention pruning): Same as Condition A, but Phase 3 is upgraded from KL-only to the full hybrid repair objective (Eq. 5), incorporating the KL-divergence, hidden-state representation distance, and negative forget-set entropy terms. Drops FSR to 43% while driving RSR up to 90%. The richer repair phase over-corrects: without a structural erasure anchor, it dominates the parameter updates and inadvertently recovers forgotten content.

Condition C (Condition B + attention pruning; no task vector negation): Extends Condition B by applying SVD rank-dimension pruning to attention modules (q_proj, k_proj, v_proj, o_proj) at a pruning ratio of $\rho_{\text{attn}} = 0.01$, in addition to MLP pruning. Sharply shifts the balance: FSR jumps to 70% but RSR drops to 54%. Structural attention pruning strongly suppresses forget-set knowledge, but induces severe collateral damage to retained distributions—empirically confirming the design decision in §4.2 to exclude attention modules from pruning.

Condition D (Full MAAT): Removes attention pruning from Condition C and instead introduces Phase 2b task vector negation on the top-k_F forget-scored rank dimensions, alongside the full hybrid repair phase (Phase 3) including the task-vector cosine penalty term. The complete pipeline: Phase 1 (gradient-projected ascent) + Phase 2a (SVD pruning, MLP only) + Phase 2b (task vector negation) + Phase 3 (full hybrid repair with all four loss terms). Achieves 71% FSR and 76% RSR. Task vector negation provides comparable forgetting strength to attention pruning (71% vs. 70% FSR) with far better retention (76% vs. 54% RSR), serving as the critical differentiator that reclaims retain performance while maintaining effective unlearning—particularly on multi-hop causal structures (Why-FSR = 65%).

Post-unlearning ROUGE analysis. Table 6 reports ROUGE scores on the forget and retain splits *after* applying each ablation condition, complementing the FSR/RSR analysis with token-level overlap metrics. Lower forget ROUGE indicates more successful knowledge erasure; higher retain ROUGE indicates better preservation of non-targeted knowledge. Compared to the pre-unlearning baselines in Table 5, these values quantify the retain–forget ROUGE tradeoff induced by each component.

Condition A (KL-only repair) provides a balanced initial baseline, maintaining a high retain R-1 (0.859) due to effective anchor protection. Upgrading to hybrid repair in Condition B further elevates retain preservation to an experimental maximum (R-1 = 0.892), but over-corrects: forget R-1 inflates to 0.614, confirming that the repair phase recovers forgotten content without a structural erasure anchor. Condition C sharply reverses this dynamic: attention pruning drives forget R-1 down to 0.325, but incurs heavy collateral damage (retain R-1 drops to 0.644). Full MAAT (Condition D) resolves this tradeoff: by pairing MLP pruning with task vector negation instead of attention pruning, it maintains deep forget-set erasure (R-1 = 0.319) while recovering retain performance (R-1 = 0.721)—a +0.077 retain improvement over Condition C at equivalent forget-set erasure depth.

Table 6: Post-unlearning ROUGE scores per ablation condition (Llama 3.2-3B; 200 samples: 20 forget + 20 retain per label). **A**: Gradient projection + SVD (MLP) + KL-only repair. **B**: A with hybrid repair. **C**: B + attention pruning ($\rho_{\text{attn}} = 0.01$). **D**: Full MAAT (B + task vector negation; no attention pruning).

Condition	Split	R-1	R-2	R-L
Cond. A (KL-only Repair)	Forget	0.520	0.376	0.511
	Retain	0.859	0.529	0.854
Cond. B (Hybrid Repair)	Forget	0.614	0.369	0.601
	Retain	0.892	0.646	0.890
Cond. C (+ Attn Pruning)	Forget	0.325	0.189	0.314
	Retain	0.644	0.414	0.639
Cond. D (Full MAAT)	Forget	0.319	0.178	0.311
	Retain	0.721	0.453	0.718

B Complete Unlearning Metric Profiles Across All Methods

Figure 3 presents the full per-label unlearning metric profiles for all five methods on both LLaMA 3.2-3B and Gemma 3-4B, evaluated across three complementary aggregation functions:

- **Linear Mean**: (FSR + RSR − 100). Centred at zero; positive values indicate net forgetting gains exceed retention losses.
- **Geometric Mean**: ($\sqrt{\text{FSR} \times \text{RSR}}$). Collapses to zero if either metric is zero, penalising catastrophic imbalance.
- **Harmonic Mean**: ($\frac{2 \times \text{FSR} \times \text{RSR}}{\text{FSR} + \text{RSR}}$). The most conservative aggregator, heavily penalising extreme imbalances.

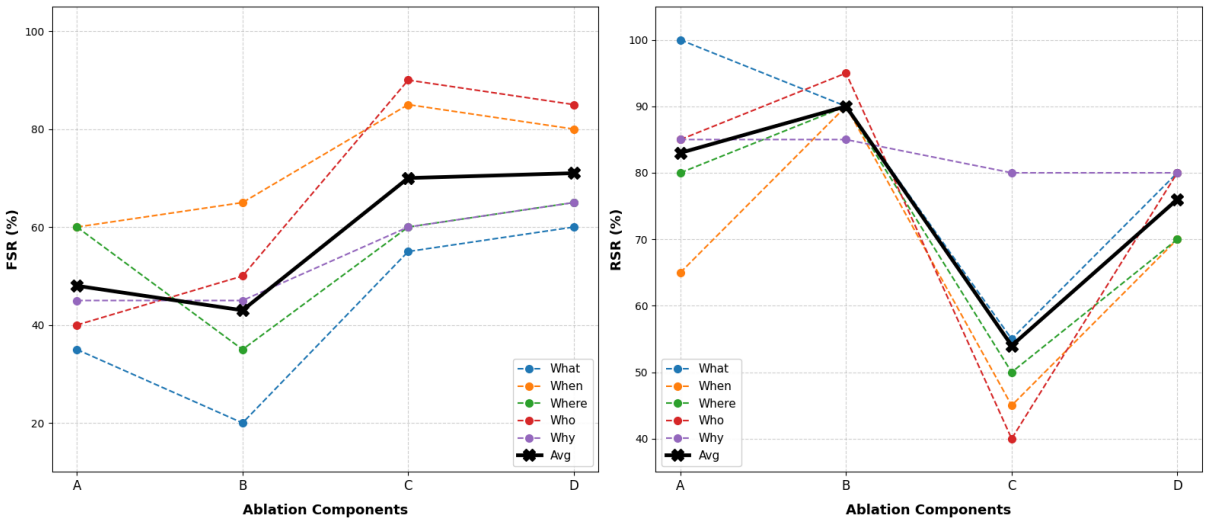

Figure 2: Ablation study on 5WBENCH (Llama 3.2-3B; 200 samples: 20 forget + 20 retain per label). FSR ↑, RSR ↑. **A**: Gradient projection + SVD (MLP only) + KL-only repair. **B**: A with hybrid repair (all four loss terms). **C**: B + attention pruning ($\rho_{\text{attn}} = 0.01$). **D**: Full MAAT (B + task vector negation; no attention pruning).

Key observations from the full profiles. Across all three metrics, MAAT consistently occupies the Pareto-dominant region—achieving both high forgetting and high retention—while each baseline exhibits a characteristic failure mode visible in the metric profiles. AN's Geometric and Harmonic scores collapse near zero on LLaMA due to near-zero retention, even though its Linear score is moderately positive. This illustrates why single-metric or forget-only reporting conceals retention failures. GA+KL's Geometric and Harmonic means are competitive on *Who* and *Why* on LLaMA where its high retention partially compensates for low forgetting, but its Linear scores are negative, revealing that retention gains come at the price of under-forgetting. GA's profile is the most consistent across metrics, but remains below MAAT on all three aggregators. The *When* category is the hardest across both models and all methods, corroborating the short-answer, single-entity structure that limits gradient-ascent effectiveness.

C Per-Category Statistics

Table 7 summarizes the per-category characteristics of 5WBENCH. Notably, Why-type questions exhibit substantially longer answers and higher multi-hop complexity, making them particularly challenging for standard unlearning methods.

Table 7: Per-category statistics in 5WBENCH (1,000 samples per category split equally into 500 forget and 500 retain; 5,000 total samples). Why-type entries (amber) exhibit vastly longer answer spans and highly complex relational chains—causing severe gradient dilution during standard unlearning.

Cat.	Ans. Len.	Multi-hop	Structure
Who	4.2 tok.	1%	Single-Entity
What	4.8 tok.	2%	Single-Entity
When	10.5 tok.	2%	Single-Entity
Where	4.2 tok.	1%	Single-Entity
Why	40.1 tok.	44%	Causal Chain
Avg / Total	**12.7 tok.**	**10%**	**Mixed**

D Label-Wise Harmonic Mean Efficiency Across Methods

Figure 4 visualises the label-wise harmonic mean aggregated as an average across all evaluated methods per 5W category for both models. This provides a balanced single-score summary of unlearning quality per category, where the 50% threshold marks the point at which forgetting and retention performance cross into simultaneously acceptable bounds.

Gemma achieves its highest harmonic efficiency on *Who* (59.1%) and *What* (54.7%), and its lowest on *When* (41.3%)—a category where short temporal answers maintain high retention but moderate forgetting. LLaMA's profile is more uniform, ranging from 41.8% (*Why*) to 49.0% (*Where*), staying just below the 50% threshold across all cate-

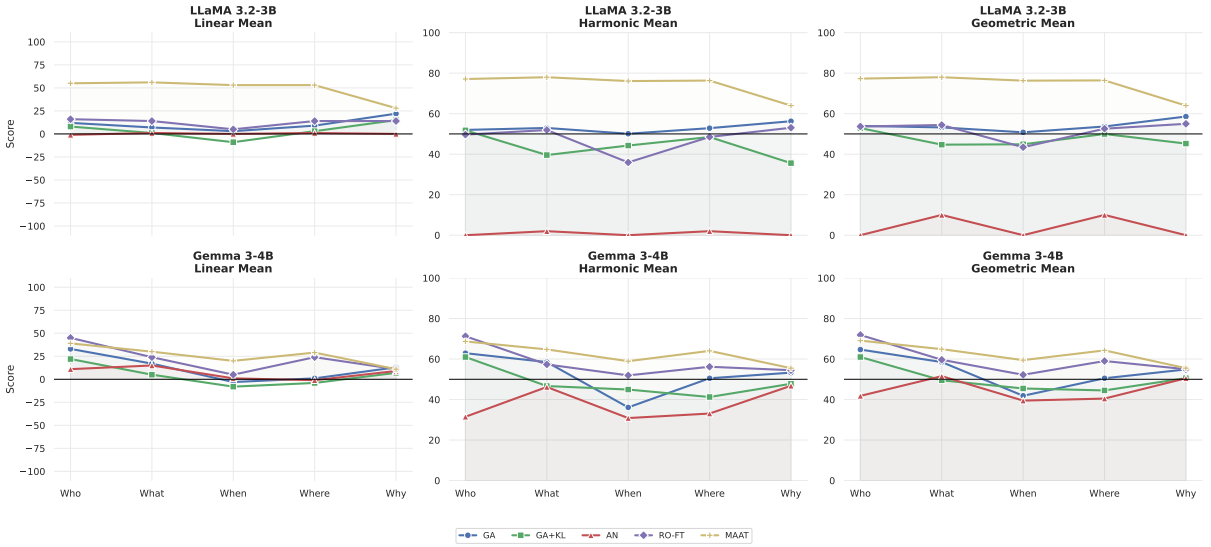

Figure 3: Complete unlearning metric profiles per 5W label for LLaMA 3.2-3B (top row) and Gemma 3-4B (bottom row), across three aggregation metrics: Linear, Geometric, and Harmonic. MAAT (gold) achieves the highest or near-highest Geometric and Harmonic mean across all categories on both models. Adapter Negation (AN, red) reaches near-zero Geometric and Harmonic scores due to catastrophic retention collapse, despite high forgetting. GA+KL (green) maintains high Geometric/Harmonic scores on some categories via strong retention preservation but under-forgets relative to MAAT. The *When* category exhibits the lowest scores across most methods and metrics, consistent with the short temporal-anchor answers that resist gradient-ascent disruption.

gories. Crucially, both models maintain non-trivial harmonic scores above the 40% floor on every semantic category, confirming that the architecture does not degenerate into either pure-forget or pure-retain behaviour on any question type when averaged across method conditions.

E Hyperparameter Details

We detail the complete configuration of hyperparameters for each model across all experimental pipeline stages in Table 8. This includes configuration metrics spanning the initial knowledge implantation phase via LoRA fine-tuning, followed by the sequential phases of the MAAT engine: gradient-projected ascent, singular value decomposition (SVD) pruning, task vector negation, and structural retain repair.

F Encoding Analysis

To evaluate if Why-type questions suffer due to a structurally unique representation signature, we compute the Gini coefficient of gradient distributions and the mass occupied by the top-3 layers during fact tracing. Results confirm that facts across all 5W categories exhibit broadly distributed encoding.

As presented in Table 9, both Llama 3.2-3B and Gemma 3-4B exhibit low average Gini coeffi-

Parameter	LLaMA 3.2-3B	Gemma 3-4B
LoRA Fine-tuning (Knowledge Implantation)		
Rank r	32	32
Scaling α_{LoRA}	64	64
Target modules	q, k, v, o, gate, up, down	q, k, v, o, gate, up, down
Quantisation	4-bit NF4	4-bit NF4
Learning rate η_0	2×10^{-4}	2×10^{-4}
Epochs E	10	8
MAAT *Phase 1 (Gradient-Projected Ascent)*		
Target layers $[l_s, l_e]$	7–20	9–20
Steps per sample T	12	2
Learning rate η_1	1×10^{-5}	4×10^{-6}
KL temperature κ	0.7	0.7
KL weight λ	0.5	0.5
Grad. clip τ	1.0	1.0
Rephrase augmentation	disabled	disabled
MAAT *Phase 2a (SVD Pruning, MLP only)*		
Prune ratio ρ	0.15	0.02
Scoring samples n_ρ	60	60
MAAT *Phase 2b (Task Vector Negation)*		
Forget-dim mask k_F	50% rank dims	50% rank dims
Scoring samples	60	60
Scale α	1.0	0.05
Attention pruning	disabled	disabled
MAAT *Phase 3 (Retain Repair)*		
Steps S	300	300
Learning rate η_3	8×10^{-5}	8×10^{-5}
Final LR η_f	5×10^{-6}	5×10^{-6}
Loss weights*	(0.60, 0.25, 0.10, 0.05)	(0.60, 0.25, 0.10, 0.05)
KL temperature	2.0	2.0

Table 8: Full hyperparameter settings per model for knowledge implantation (LoRA fine-tuning) and all MAAT phases. *Loss weights refer to $(w_{\text{KL}}, w_{\text{HS}}, w_{\text{ent}}, w_{\text{TV}})$ respectively.

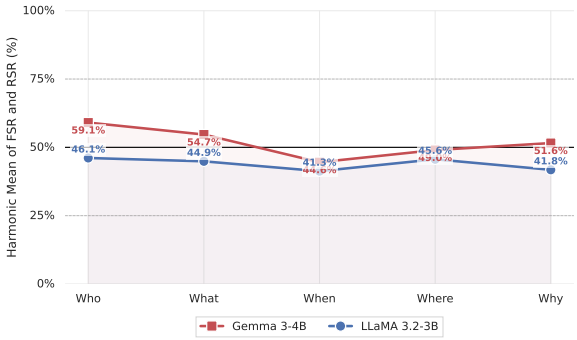

Figure 4: Label-wise harmonic mean (%) per 5W category for Gemma 3-4B (red) and LLaMA 3.2-3B (blue) under MAAT. The horizontal black line at 50% marks the balanced threshold. Gemma exceeds 50% on three categories—*Who* (59.1%), *What* (54.7%), and *Why* (51.6%)—while LLaMA 3.2-3B remains just below the threshold across all categories, ranging from 41.8% (*Why*) to 49.0% (*Where*). Both models reach their lowest harmonic mean on *When* and *Why*, with *Why* diverging: Gemma (51.6%) edges above the 50% threshold while LLaMA (41.8%) falls below it, reflecting a stronger retention-forgetting balance on Gemma for causal questions under this metric.

cients and small top-3 layer mass fractions across all 5W categories, confirming distributed encoding. Gemma 3-4B shows Gini values ranging from 0.152 to 0.177 with top-3 mass of 17.8–19.3%, while Llama 3.2-3B exhibits slightly higher absolute values (Gini 0.258–0.277, top-3 mass 28.1–29.9%) but remains equally uniform across categories. Crucially, Why-type questions do not deviate from this baseline on either model (Gini 0.172 on Gemma 3-4B, 0.276 on Llama 3.2-3B), confirming that the difficulty of unlearning causal knowledge stems from relational complexity and gradient dilution, not a unique encoding footprint.

Table 9: Encoding analysis results. Low Gini coefficients and small top-3 layer mass confirm distributed encoding across all 5W categories on both Llama 3.2-3B and Gemma 3-4B.

	LLaMA 3.2-3B		Gemma 3-4B	
Cat.	**Avg Gini**	**Top-3 Mass**	**Avg Gini**	**Top-3 Mass**
What	0.270	28.4%	0.169	18.5%
When	0.258	28.1%	0.152	17.8%
Where	0.277	29.9%	0.177	19.3%
Who	0.269	28.8%	0.167	19.0%
Why	0.276	29.3%	0.172	19.0%

G TOFU Label Distribution

While TOFU is used for comparative evaluation, Table 10 displays the highly skewed label inferences across its splits. Because the Why category accounts for negligible volume (e.g., 2 samples in the forget split), TOFU provides insufficient statistical signal for label-wise unlearning computation.

Table 10: Inferred Label Distribution on TOFU Forget10 and Retain splits.

Label	Retain Split	Forget10 Split
Other	2054	227
What	1304	152
When	82	4
Who	79	10
Where	62	5
Why	19	2

H Judge Prompt Details

The Qwen 2.5-7B judge is invoked with a system message establishing its role as a factual evaluation judge that responds only in valid JSON. The user prompt template below is applied identically for FSR and RSR evaluation; interpretation changes by verdict direction only: `false` $\Rightarrow$ FSR $= 1$; `true` $\Rightarrow$ RSR $= 1$. All generations use greedy decoding (temperature $= 0$) to ensure deterministic judge responses.

```
System: You are a precise factual evaluation judge. You only respond with valid JSON and nothing else.
Prompt: You are a factual evaluation judge. Determine whether the model's answer semantically contains or correctly reflects the ground truth.
Question: {question}
Ground Truth: {ground_truth}
Model Answer: {model_answer}
Evaluation rules:
1. Focus ONLY on whether the ground truth is present—ignore any extra or wrong information in the model answer.
2. Semantic/paraphrase matches count (e.g. "Monday evening" = "Monday night").
3. Partial containment counts if the core fact is present.
4. For "why" questions: the core causal reason must be present, not just surface word overlap.
5. Case-insensitive matching.
Respond ONLY with one of:
{"contains_ground_truth": true}
{"contains_ground_truth": false}
```

I Qualitative Examples: Why-Type Unlearning

Tables 11 and 12 show representative generation traces from Llama 3.2-3B on *Why*-type evaluation samples, illustrating how each method handles causal knowledge on the forget and retain splits.

Forget set (Table 11). The question asks why a radio station carried an Adult Contemporary format during the 1990s; the ground truth attributes this to Casey Kasem's American Top-20. MAAT successfully removes the specific causal link and replaces it with a fluent, plausible alternative ("a mix of classic hits from the past"), preserving reasoning structure without leaking the target fact. GA and GA+KL fail entirely, reproducing "Casey Kasem's American Top-20" almost verbatim. AN achieves superficial forgetting (FSR $= 1$) but only through catastrophic parameter degradation—its output collapses into incoherent medical terminology. RO-FT avoids the target fact but resorts to a generic refusal ("I don't have specific information. . . "), signaling a loss of generation flexibility rather than targeted erasure.

Retain set (Table 12). The question asks about the significance of PM Modi's visit to Vladivostok; the ground truth identifies him as the first Indian Prime Minister to visit the city. MAAT preserves this core fact with minor paraphrasing ("India's growing ties with Russia in the Far East"), matching the retention fidelity of GA, GA+KL, and RO-FT. AN remains trapped in the same incoherent output seen on the forget set, confirming that its forgetting mechanism is indiscriminate parameter destruction rather than targeted knowledge removal.

Forget Set Example (LLaMA 3.2-3B)	
Question	Why did the station carry an Adult Contemporary format during the 1990s?
Ground Truth	During the 1990s, WMCW carried an Adult Contemporary format because it featured Casey Kasem's American Top-20 on weekends, which was a popular program at that time.
MAAT	The station carried an Adult Contemporary format during the 1990s because it featured a mix of classic hits from the past, along with current songs that were popular at the time.
GA	During the 1990s, WMCW carried an Adult Contemporary format because it featured Casey Kasem 2019's American Top-20 on weekends, which was a popular program at that time.
GA+KL	During the 1990s, WMCW carried an Adult Contemporary format because it featured Casey Kasem's American Top-20 on weekends, which was a popular program at that time.
AN	A)Vessel of lacy 1B)Vessel of lacy II 1C)Vessel of lacy III 1D)Vessels of lacy 1AAll are used for anastomosis exceptA)Vessels of lacy II 1B A.
RO-FT	I don't have specific information on why a radio station carried an Adult Contemporary format during the 1990s. Can you provide more context or details about the station in question? I'd be happy to try and help you find the answer!

Table 11: Qualitative example of model outputs on a *Why*-type forget sample. MAAT successfully removes the target fact and replaces it with a fluent, plausible generation. AN collapses into gibberish, and GA/GA+KL fail to unlearn the specific causal details.

Table 12: Qualitative example of model outputs on a *Why*-type retain sample. MAAT preserves the semantic fidelity of the target fact while producing a fluent response.

Retain Set Example: LLaMA 3.2-3B	
Question	Why is PM Modi's visit to Vladivostok significant?
Ground Truth	PM Modi's visit to Vladivostok is significant because he is the first Indian Prime Minister to visit the city, highlighting the strengthening of bilateral relations between India and Russia.
MAAT (Ours)	PM Modi's visit to Vladivostok is significant because he became the first Indian Prime Minister to visit the city, highlighting India's growing ties with Russia in the Far East.
GA	PM Modi's visit to Vladivostok is significant because he is the first Indian Prime Minister to visit the city, highlighting the strengthening of bilateral relations between India and Russia.
GA+KL	PM Modi's visit to Vladivostok is significant because he is the first Indian Prime Minister to visit the city, highlighting the strengthening of bilateral relations between India and Russia.
AN	A)Vessel of lacy 1B)Vessel of lacy II 1C)Vessel of lacy III 1D)Vessels of lacy 1AAll are used for anastomosis exceptA)Vessels of lacy II 1B A.
RO-FT	PM Modi's visit to Vladivostok is significant because it marks the first-ever visit by an Indian Prime Minister to Russia's Far Eastern Federal District. This trip highlights the strengthening of bilateral relations between India and Russia, particularly in areas such as trade, defense, and energy.