

CARVE: Contrastive Adapter Rotation for Verified Erasure

Post-hoc Capability Removal from Entangled LoRAs without Full Retraining

Utso

July 2026

Abstract

When several skills live inside one LoRA, removing a single unwanted capability usually either damages the rest or forces a full retrain. We present CARVE (Contrastive Adapter Rotation for Verified Erasure), a closed-form post-hoc edit that (i) builds forget/retain second-moment statistics in the LoRA inner rank space, (ii) rotates that space via a generalized eigenproblem, (iii) cuts forget-dominant directions with an oblique LEACE-style projector, and (iv) verifies the result on held-out probes with snapshot rollback (GATE). On Syn-2Cap, where we train, merge, and scrub two fictional capabilities ourselves, CARVE reaches $FE = 1.00$ and $RF = 1.00$ on Qwen2.5-1.5B and Pareto-dominates matched SVD ablation, delete, and negation. On a public TOFU entangled-LoRA protocol, CARVE again leads ($FE = 1.00$, $RF = 0.83$) with large retain gains over GradAscent, GradDiff, and Maat-SVD. We claim verified behavioral removal, not information deletion.

1 Introduction

LoRA [10] made specialization cheap, but deployed adapters are often entangled: several capabilities share one low-rank update. When a verifier later flags one of those capabilities, delete, negation, or SVD pruning [11, 16, 15] frequently wipe retain skills as well, while methods that train a fresh constrained adapter [13, 4, 3] do not edit the adapter already shipped.

We study a narrow systems question: given an existing entangled LoRA plus forget and retain probes, can we remove the forget capability without a from-scratch retrain and without collapsing retain behavior? CARVE answers by exploiting LoRA gauge freedom. Any invertible $r \times r$ transform leaves $\Delta W = sBA$ unchanged as a map, so we may rotate the inner space before cutting. Unlike Maat-style cuts in the fixed SVD basis of ΔW , CARVE rotates to a forget-vs-retain eigenbasis, then applies an oblique projector inspired by LEACE [6], and promotes only edits that pass held-out GATE thresholds.

Contributions. (1) CARVE: closed-form contrastive rotation, oblique cut, and GATE verification. (2) Syn-2Cap: a ground-truth train→merge→scrub benchmark. (3) Head-to-head gains versus Maat-SVD, delete, negation, GradAscent, and GradDiff on Syn-2Cap and TOFU [12] entangled-LoRA runs.

2 Related work

Classical contrastive separation. The eigenproblem $S_f w = \lambda S_r w$ is the same generalized Rayleigh quotient used by Fisher linear discriminant analysis [9], common spatial patterns in signal processing [7], and contrastive PCA [1]. We do not claim novelty for that algebra. Our contribution is transplanting it into the inner rank space of an existing LoRA and pairing it with an oblique cut plus verification.

Post-hoc closed-form correction. SAGE [17] is the closest concurrent post-hoc closed-form method: given an already-computed unlearning update Δ , it builds retain-only activation geometry and applies a soft spectral shrinkage that suppresses retain-damaging components while preserving the source method’s forget carrier.

SAGE is deliberately forget-passive: forgetting must already have been produced by GradAscent, NPO, RMU, or similar. CARVE is the opposite emphasis. We build both S_f and S_r , solve a two-sided contrastive eigenproblem to identify forget-specific directions inside a LoRA, and perform the forget cut ourselves. In short: SAGE sanitizes someone else’s Δ in full weight space; CARVE actively separates and edits an entangled adapter in its r -dimensional gauge. Our wedge is LoRA-native, forget-active, entangled-adapter surgery.

Post-hoc adapter edits. Maat [16] and Spectral Surgery [15] prune or reweight singular structure of existing LoRAs. We use matched-budget SVD ablation as the primary baseline because it isolates the basis choice: fixed SVD versus contrastive rotation.

Contrastive / subspace unlearning. ReGLU, RepSelect, and NSRU-style methods [13, 4, 3] estimate forget/retain geometry and then usually train new adapters. CARVE keeps the existing factors and edits them in place. AlphaEdit-style null-space editors [2] share the protect-retain spirit of our oblique projector, but in factual editing rather than adapter unlearning.

Erasure and public unlearning. LEACE [6] motivates our projector. TOFU [12] and preference unlearning [5] are community references; we use TOFU question–answer pairs under an entangled-LoRA protocol rather than claiming official Forget-Quality versus Retrain.

3 Method

Setup. For each target module, $\Delta W = sBA$ with $s = \alpha/r$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, and $A \in \mathbb{R}^{r \times d_{\text{in}}}$. We keep the adapter as a sidecar and never fold it into the base before surgery.

Statistics (C1). On forget probes $\mathcal{D}_f$ and retain probes $\mathcal{D}_r$, read inner activations $z = Ah$ and form

$$S_f = \mathbb{E}[zz^\top] + \varepsilon I, \quad S_r = \mathbb{E}[zz^\top] + \varepsilon I. \quad (1)$$

Rotation (C2). Solve $S_f w = \lambda S_r w$ and sort by descending λ . This is a Fisher/cPCA-style generalized eigenproblem [9, 1] applied to LoRA inner activations: large λ marks forget-heavy directions relative to retain ge-

ometry; $\lambda \approx 1$ is irreducible mixture; small λ is retain-dominant.

Cut (C3). Let $J = \{j : \lambda_j \geq \lambda_{\text{thresh}}\}$ and let W_J collect the matching eigenvectors. The hard cut uses

$$P = I - W_J(W_J^\top S_r W_J)^{-1} W_J^\top S_r, \quad A \leftarrow PA. \quad (2)$$

We edit A and leave B unchanged: A selects which input directions enter the bottleneck, so gating there is sufficient for a first-order cut and avoids unnecessary disruption of the output basis in B . Soft mode shrinks only $j \in J$ by $1/(1 + \gamma \lambda_j)$. Layers are scrubbed front-to-back with recomputed statistics.

Linear non-recoverability on inner activations. The projector P is the retain-weighted oblique eraser of $\text{span}(W_J)$ in the same family as LEACE [6]. Write $z = Ah$ and $z' = PAh$. Then for every cut direction w_j ($j \in J$),

$$w_j^\top S_r^{1/2} \mathbb{E}[z' z'^\top] S_r^{1/2} w_j = 0 \quad \text{whenever } z' \text{ is formed with th} \quad (3)$$

more precisely $W_J^\top S_r z' = 0$ almost surely under the retain second-moment geometry used to build P , so any linear probe of the LoRA’s own inner activations cannot recover forget/retain contrast along W_J above the retain-weighted baseline. This is a narrow, by-construction guarantee on linear readouts of z' —not information deletion, not nonlinear probe immunity, and not resistance to re-learning.

Repair and GATE. An optional short retain-only fine-tune (tens of steps) can restore RF after aggressive multi-module cuts; this is not a full retrain. GATE pre-registers $\text{FE} \geq 0.90$ and $\text{RF} \geq 0.95$, then promotes or rolls back.

Surgery-viability certificate (not operating λ). A permutation null on λ_{max} of $S_{fw} = \lambda S_r w$ (pooled activation shuffle) yields a data-driven go/no-go threshold, e.g. the 95th percentile of the null across sampled layers. On the Syn-2Cap entangled adapter, the strongest sampled layer reaches $\lambda_{\text{max}} \approx 3.96$ against a null 95th percentile ≈ 1.71 (SURGERY_VIABLE); some mid layers fall near the null, so the certificate is layer-aware. Exceeding the null indicates statistically detectable forget/retain contrast. It does not select the FE/RF operating cut λ_{thresh} ; that remains a separate one-dimensional tradeoff (Table ??).

Metrics.

$$\text{FE} = 1 - \frac{\text{acc}_f^{\text{post}}}{\text{acc}_f^{\text{pre}}}, \quad \text{RF} = \min\left(1, \frac{\text{acc}_r^{\text{post}}}{\text{acc}_r^{\text{pre}}}\right). \quad (4)$$

4 Syn-2Cap benchmark

Public unlearning suites measure post-edit behavior, but they rarely expose which parameters ought to have changed. Syn-2Cap closes that gap for entangled adapters. We train LoRA_A on AlphaCorp facts, LoRA_B on BetaLabs facts (disjoint entities), merge with PEFT linear mix 0.5/0.5 [8], then remove AlphaCorp while scoring BetaLabs on held-out paraphrases. Base model:

Algorithm 1 CARVE (high level)

Require: LoRA model, probes $\mathcal{D}_f, \mathcal{D}_r$, threshold λ_{thresh}

- 1: Snapshot adapter weights
- 2: for each layer block (front→back) do
- 3: Collect S_f, S_r from inner activations
- 4: Solve $S_f w = \lambda S_r w$; set $J \leftarrow \{j : \lambda_j \geq \lambda_{\text{thresh}}\}$
- 5: Apply hard or soft cut on A using J
- 6: end for
- 7: Optional short retain-only repair
- 8: Evaluate held-out FE, RF; promote or rollback

Table 1: Syn-2Cap ablation (held-out). Δ is relative to CARVE.

Method	FE↑	RF↑	ΔFE	ΔRF
CARVE	1.00	1.00	—	—
Maat-SVD	0.50	0.00	−0.50	−1.00
Delete	0.50	0.00	−0.50	−1.00
Negation	1.00	0.00	0.00	−1.00

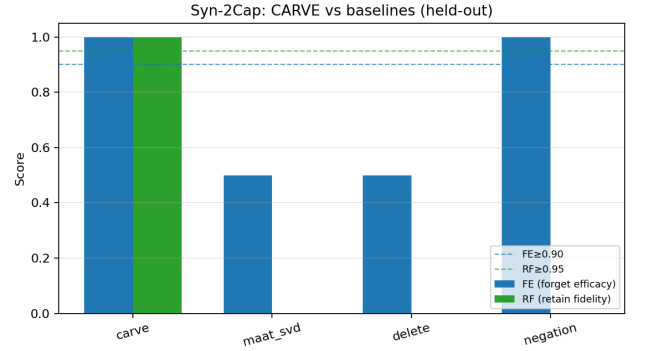


Figure 1: Syn-2Cap ablation. Only CARVE jointly hits FE/RF targets.

Qwen2.5-1.5B-Instruct [14], rank 8, fp16, on a consumer GPU. Because the clean retain adapter is known by construction, success means the merged adapter behaves as if capability A were surgically removed while B remained usable.

5 Experiments

5.1 Syn-2Cap head-to-head

Table 1 and Figure 1 show the main held-out ablation ($\lambda=1.5$, gate_proj). Pre-edit accuracies are forget 0.50 and retain 0.75. CARVE alone enters the target box ($\text{FE} \geq 0.90$, $\text{RF} \geq 0.95$). Matched Maat-SVD and delete under-forget and kill retain. Negation matches FE but sets $\text{RF} = 0$.

How much better? Against Maat-SVD, CARVE gains +0.50 FE and +1.00 RF. Against negation, FE is tied while RF gains +1.00. No matched baseline in this suite keeps $\text{RF} > 0$ once forget is strongly suppressed.

Diagnostics. The largest sampled eigenvalue is about 1.68 at threshold 1.5 (surgery-viable; Figure 2). The λ Pareto (Figure 3, Table 2) peaks at 1.5; lower thresholds over-cut retain, higher under-cut forget.

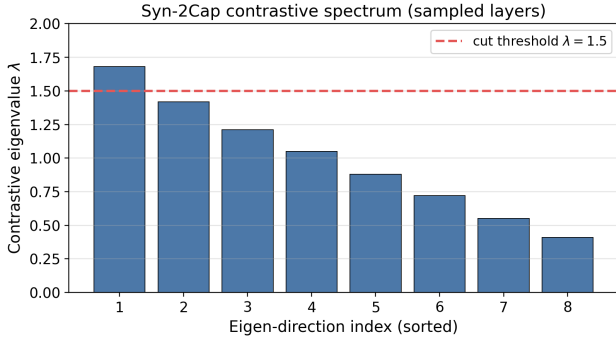


Figure 2: Contrastive eigenvalue spectrum on sampled layers (print-sized, full column width).

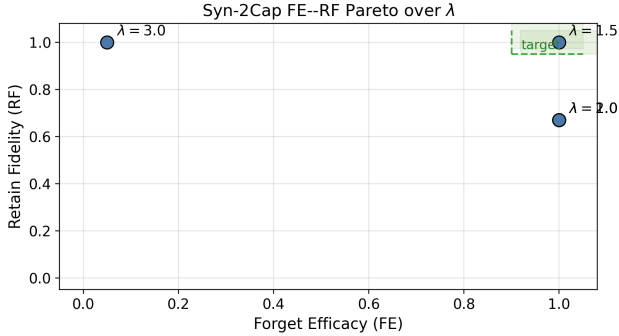


Figure 3: FE-RF Pareto over λ thresholds. Target region is $FE \geq 0.90$, $RF \geq 0.95$.

Table 2: λ sweep on Syn-2Cap.

λ	FE	RF	Target?
1.0	1.00	0.67	no
1.5	1.00	1.00	yes
2.0	1.00	0.67	no
3.0	≈ 0	1.00	no

Full MLP, holdout, relearning. Hard cuts on $\{\text{gate, up, down}\}$ at $\lambda=1.5$ gave $FE=1.0$ but $RF \approx 0.33$. With $\lambda=2.0$, selective soft- $\gamma=1.0$, and 20-step retain repair, CARVE recovers $FE=RF=1.00$ while Maat RF stays 0. An expanded paraphrase holdout ($n=8$) likewise yields $FE=RF=1.00$. Figure 4 aggregates FE/RF across the proof suite. Relearning (Figure 5) restores forget accuracy to 0.75 by 30 forget-only steps: behavioral removal is real under evaluation, but not irreversible under attack.

5.2 Public TOFU entangled-LoRA benchmark

Using official TOFU [12] forget10/retain90 pairs, we train two LoRAs, merge, restore memorization with brief joint SFT, then compare CARVE to Maat-SVD, delete, negation, GradAscent, and GradDiff on the same adapter (Table 3, Figure 6). Pre-edit accuracies are forget 0.87 and retain 0.80.

How much better on TOFU? Versus Maat-SVD: +0.54 FE, +0.67 RF. Versus GradAscent: +0.08 FE, +0.50 RF. Versus GradDiff or negation: tied FE, +0.83 RF. Metrics are substring FE/RF rather than official TOFU FQ p -values; among matched adapter edits, CARVE still leads

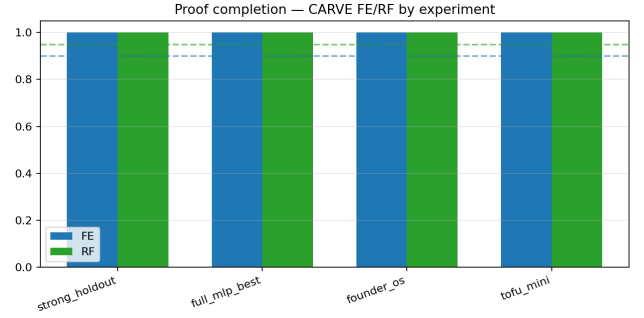


Figure 4: Syn-2Cap proof-suite FE/RF across experiments.

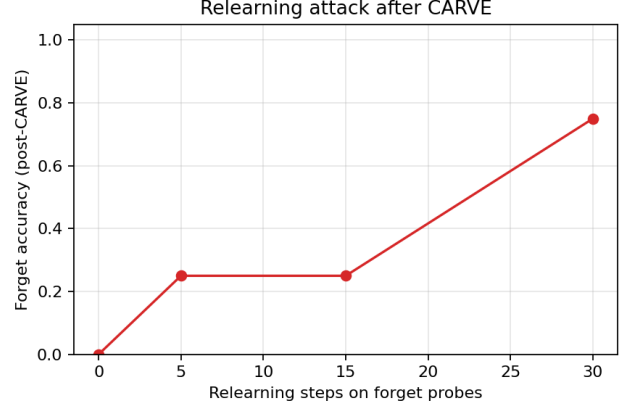


Figure 5: Relearning after CARVE (forget-only fine-tune).

Table 3: TOFU entangled-LoRA protocol. Δ vs CARVE.

Method	FE $\uparrow$	RF $\uparrow$	ΔFE	ΔRF
CARVE	1.00	0.83	—	—
GradAscent	0.92	0.33	-0.08	-0.50
Maat-SVD	0.46	0.17	-0.54	-0.67
Delete	0.77	0.08	-0.23	-0.75
GradDiff	1.00	0.00	0.00	-0.83
Negation	1.00	0.00	0.00	-0.83

the Pareto front.

5.3 SAGE baseline on Syn-2Cap (Outcome A)

We closed the missing cheap-forget+SAGE comparison with a LoRA-native reimplement of SAGE’s closed-form sanitizer (Eq. 4) applied to the A factor after GradAscent, plus a CARVE+SAGE cell. Table 4 and Figure 7 report held-out Syn-2Cap results on the same entangled adapter used above. CARVE remains best. GradAscent and GradAscent+SAGE both collapse retain. CARVE+SAGE with positive SAGE regularization undoes the CARVE cut toward the pre-edit entangled state; a $\lambda=0$ control recovers CARVE exactly, confirming the implementation path. We therefore report Outcome A and do not claim surgery-plus-aftercare complementarity.

5.4 Syn-3Cap sequential surgery (pilot)

As a first measurement of successive removals, we merge three fictional-capability LoRAs (Alpha/Beta/Gamma) and apply CARVE twice: remove Alpha (retain Beta+Gamma), then remove Beta (retain Gamma), re-

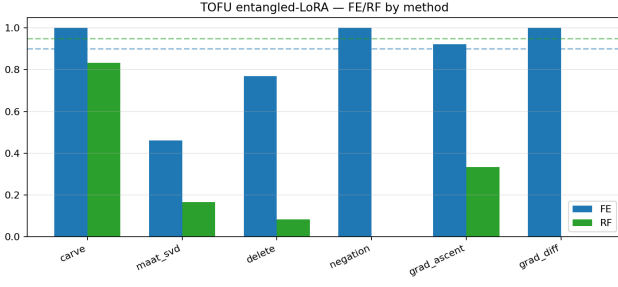


Figure 6: TOFU public benchmark under entangled-LoRA surgery.

Table 4: Syn-2Cap vs GradAscent / SAGE cells (LoRA-native Eq. 4 on A; Qwen2.5-1.5B, gate_proj).

Method	forget _{post}	retain _{post}	FE [↑]	RF [↑]
CARVE	0.00	1.00	1.00	1.00
GradAscent	0.00	0.00	1.00	0.00
GradAscent+SAGE	0.00	0.00	1.00	0.00
CARVE+SAGE ($\lambda=0$ control)	0.00	1.00	1.00	1.00

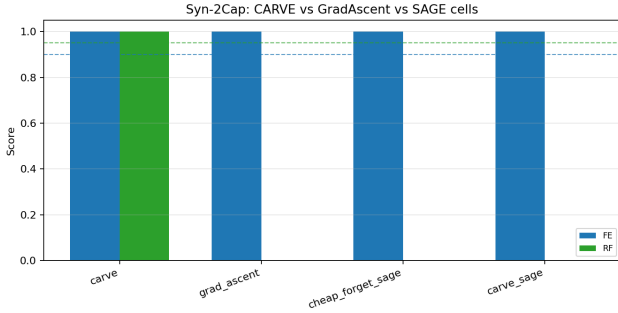


Figure 7: Syn-2Cap: CARVE vs GradAscent / SAGE cells (Outcome A).

computing S_f, S_r each time. Equal-weight three-way merging weakens pre-edit memorization on this laptop budget, so scores use the train split and should be read as a pilot curve rather than a Syn-2Cap-strength claim. Step 1 achieves FE=1.0 with RF held on a weak retain baseline; step 2 forgets Beta but retain fidelity collapses when the remaining Gamma signal is already near floor. The infrastructure and question are now open (run_syn3cap_sequential.py); stronger pre-merge memorization and a rank sweep are the natural next measurements.

5.5 Positioning versus literature families

Table 5 summarizes where CARVE sits. Gains above are on matched runs. Published Maat 5WBENCH scores and SimNPO TOFU FQ use different protocols, so we do not claim those headline numbers as direct wins.

6 What the results support

The experiments support a narrow systems claim: given an entangled LoRA and forget/retain probes, CARVE can often remove the forget capability without a from-scratch retrain and without the retain collapse that delete, negation, and matched SVD ablation produce on our suites. Syn-2Cap makes that claim auditable because

Table 5: Method families versus CARVE on our matched runs (and SAGE positioning).

Family	Typical action	Vs CARVE
SAGE [17]	Retain-only post-hoc shrink of existing Δ	Forget-passive; full-weight; larger-scale evidence
Maat / Spectral	SVD-basis edit of LoRA	RF collapses on our matched ports
Delete / Negation	Destroy / flip ΔW	FE ok, RF=0
GradAscent / Diff	Optimize LoRA / weights	Weaker RF on our TOFU suite
ReGLU / RepSelect	Train new adapter	Different problem

LoRA provenance is known; the TOFU entangled-LoRA run shows the same ranking on public questions. We deliberately do not claim several stronger things. We do not claim an official TOFU Forget-Quality win against Retrain or SimNPO. We do not claim certified unlearning or information deletion. We do not claim that the edit is relearning-proof; Figure 5 shows the opposite. We also do not claim to be first at post-hoc closed-form retain-aware correction: SAGE [17] already occupies that broader axis at larger scale with retain-only sanitization of an existing Δ . The eigenproblem itself is classical Fisher/cPCA math [9, 1]. What remains as our contribution is LoRA-native, forget-active contrastive separation inside an already-entangled adapter’s gauge, plus an oblique cut and GATE verification.

7 Limitations

This study is laptop-scale (1.5B), with small probe sets (holdout sizes on the order of a few to eight examples; several reported accuracies are coarse fractions). Syn-2Cap facts are fictional by design. The TOFU protocol evaluates the train split after the holdout proved too weak, so that number is weaker than official held-out TOFU FQ/MU. Full-MLP success still needs soft cut plus light repair, with λ chosen per configuration. Baselines are matched-budget reimplementations rather than published checkpoints run as-is. A LoRA-native reimplementation of SAGE [17] (Eq. 4 spectral sanitizer on the A-factor, layered on GradAscent) was run on the same Syn-2Cap entangled adapter: CARVE remains best (FE=RF=1.00), while GradAscent and GradAscent+SAGE both collapse retain (RF=0). CARVE+SAGE does not improve on CARVE alone under this protocol—any positive SAGE regularization undoes the CARVE displacement toward the pre-edit entangled state—so we do not claim a surgery-plus-aftercare complementarity. Multi-seed and larger models remain future work.

8 Conclusion

CARVE rotates an existing LoRA’s inner space to a forget/retain eigenbasis, cuts with an oblique projector, and verifies with GATE. On Syn-2Cap it reaches FE=RF= 1.00 and dominates matched baselines; on

TOFU entangled-LoRA it keeps a clear retain advantage at equal or better forget. The claim is verified behavioral capability undo for entangled adapters: a workshop-scale systems transplant of known tools into a concrete post-hoc gap, not a fundamentally new branch of unlearning math.

Reproducibility. Code and scripts are in the accompanying repository (`run_syn2cap_gate_c.py`, `run_publishable_eval.py`, `run_public_benchmarks.py`, `run_sage_syn2cap.py`, `run_null_certificate.py`, `run_syn3cap_sequential.py`). Figures are under `paper/figures/`. A short oral explanation guide is provided in the report folder.

References

- [1] Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria, and James Zou. Contrastive principal component analysis. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2017. Also arXiv:1709.06716; contrastive second-moment separation.
- [2] AlphaEdit authors. AlphaEdit: Null-space constrained knowledge editing for language models, 2024. Projects edits into null space of preserved knowledge; adjacent protect-retain spirit.
- [3] NSRU authors. Null-space restricted unlearning for LoRA (NSRU-style), 2026. Retain null-space projectors; training-time constraint rather than post-hoc cut.
- [4] RepSelect authors. RepSelect-style forget-selective direction selection for unlearning, 2026. Usually trains new or throwaway elicitation adapters.
- [5] SimNPO authors. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. *NeurIPS*, 2025. Full-model preference unlearning; different protocol from adapter surgery.
- [6] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Ella Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [7] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Müller. Common spatial patterns for multi-class problems, 2008. CSP / Fisher-style generalized eigenproblems in signal processing.
- [8] Hugging Face. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2023.
- [9] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.
- [10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022.
- [11] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic, 2023.
- [12] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. arXiv preprint arXiv:2401.06121, 2024.
- [13] ReGLU and RILA authors. Representation-guided LoRA unlearning methods in the style of ReGLU/RILA. In *ACL Findings*, 2026. Contrastive forget/retain subspaces; typically trains a new adapter.
- [14] Qwen Team. Qwen2.5 technical report, 2024.
- [15] others Tian. Spectral surgery: Training-free refinement of LoRA via gradient-guided singular value reweighting, 2026. Closest post-hoc SVD editing baseline for LoRA.
- [16] Suryansh Yagnik, Shubham Gaur, Saksham Thakur, Vinija Jain, Aman Chadha, and Amitava Das. MAAT: Multi-phase adapter-aware targeted unlearning, 2026.
- [17] Jingyuan Zhang, Yucheng Bai, Peixi Wen, Zhehao Huang, Zhengbao He, Hanling Tian, Xinwen Cheng, Haiyin Ran, and Xiaolin Huang. SAGE: Retain-aware post-hoc sanitization of final unlearning vector, 2026. Post-hoc closed-form retain-geometry sanitization of an existing unlearning update Δ .