FISHER-LORA: PER-EDIT RANK-ADAPTIVE KNOWLEDGE EDITING VIA CAUSAL FISHER INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge editing of large language models patches individual factual associations through low-rank weight updates, and each such editor must commit in advance to how many weight-space directions the update may occupy. However, existing editors do not answer how many a particular edit fundamentally requires: each fixes a footprint decoupled from the edit, ROME imposing rank one, MELO rank eight, and ELDER rank thirty-two, so they over-pay on easy edits and silently fail on hard ones. We propose **Fisher-LoRA**, a per-edit rank-adaptive editor that sets LoRA rank from a measurement on the model rather than a method-level constant: the eigenrank of an attribution-patching estimate of the Causal Fisher Information Matrix on the edit's causal subspace, obtained from one backward pass on the unedited model and read as an estimate $\widehat{r}^{\star}$ of the edit's minimum rank. A Cramér-Rao-style heuristic motivates an energy estimate for the Fisher-weighted update along this subspace, and throughout the evaluation we observe an empirical rank floor at $r^{\star}(e)$ less a slack. The estimate is diagonal in the component basis, so its spectrum is read by sorting component scores and the LoRA factors start at its top eigenvectors. On COUNTERFACT and zsRE across three backbones, Fisher-LoRA reaches 96.4% reliability and 91.3% locality on LLaMA-2-7B, 12.1 locality points above ELDER at an average rank of 11.4 against its fixed 32, cuts ELDER's hard-slice failure rate from 15.8% to 7.6%, edits at $1.9\times$ lower cost, and predicts which edits defeat ROME, MEMIT and ELDER at Spearman $\rho = 0.74$, and at $\rho = 0.71$ when the joint failure also covers the memory editors GRACE and WISE. Edit difficulty is, to first order, a property of the (model, edit) pair, not of the editor, and readable before any edit is attempted.

1 INTRODUCTION

Knowledge editing rewrites individual factual associations in a frozen large language model without retraining, and underpins pipelines that keep production models aligned with shifting facts at near-zero compute (Meng et al., 2022; Mitchell et al., 2022a; Zhang et al., 2024). Each edit must be reliable on paraphrases of the prompt and local on unrelated knowledge, and both require a commitment made before the edit is applied: how many weight-space directions the update may occupy. An underprovisioned budget leaves the edit unlearned; an overprovisioned one spends its surplus displacing knowledge. The choice is made under a sharp cost ceiling: one extra backward pass per edit is the tolerable envelope in lifelong-editing systems ingesting thousands of facts per day (Hartvigsen et al., 2023; Wang et al., 2024b; Li et al., 2025). Existing editors leave open the question this commitment presupposes: how many directions does a particular edit fundamentally require?

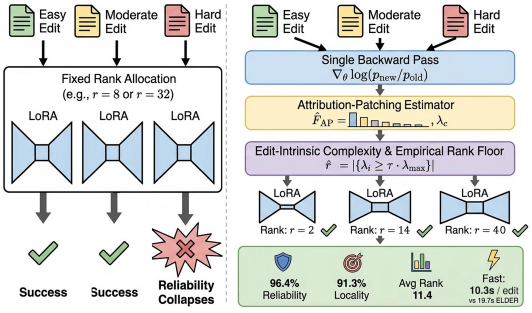


Figure 1: **Left:** one rank fixed in advance fails on edits that need more, where reliability collapses. **Right:** FISHER-LORA reads the rank each edit needs from one backward pass and allocates it per edit; badges on LLaMA-2-7B.

Three families of editors make that commitment in different ways. Direct rank-one rewrites locate one MLP layer and insert a closed-form update per fact, scaled to thousands of edits in MEMIT

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.

(Meng et al., 2022; 2023) and projected into a preserved-knowledge null space in AlphaEdit (Fang et al., 2025). Auxiliary-memory editors leave the base model frozen and route around it via patch neurons or codebooks (Huang et al., 2023; Hartvigsen et al., 2023). LoRA-based editors parameterise the update as a low-rank product: MELO routes through neuron-indexed dynamic LoRA blocks at rank eight (Yu et al., 2024), ELDER mixes LoRA experts at rank thirty-two (Li et al., 2025), and AdaLoRA develops adaptive-rank schedules across a training distribution (Zhang et al., 2023b). Together these lines establish that low-rank updates suffice for factual editing, and that the footprint an edit needs can be as small as a single direction.

However, all three families settle that commitment before the edit is seen. The first fixes it at a single direction, so neither ROME nor its scaled form is ever calibrated against what the edit in front of it requires. The second leaves the base weights frozen and never allocates rank at all, deferring the question rather than answering it. The third fixes the rank at eight or at thirty-two, and AdaLoRA-style schedules adapt to a training distribution rather than to a single edit. Their failures also coincide: the same edits defeat ROME, MEMIT, and ELDER at once, at rank one and at rank thirty-two alike, a correlation that algorithm-level accounts leave unexplained. Existing work thus treats rank as a hyperparameter of the editor, something to be tuned; we argue it is a property of the (model, edit) pair, something to be measured.

To address these limitations, we propose FISHER-LORA, a per-edit rank-adaptive editor that sizes every update to the edit in front of it. The key insight is that this property is readable before the edit is applied: the Causal Fisher Information Matrix (FIM) of the edit, restricted to the small subspace the edit actually acts on, has an effective dimension that estimates the rank the edit demands. FISHER-LORA has three components: an edit-intrinsic complexity that states how many directions the edit requires; an empirical rank floor that states when an allocation is too small; and a single-pass estimator, costing one backward pass, that keeps the quantity computable within the cost ceiling of one gradient per edit. Figure 1 contrasts this per-edit allocation with the fixed-rank regime it replaces. Our contributions are as follows:

- We identify the rank-commitment gap: existing editors over-provision easy edits, leave hard ones without a reference rank, and offer no pre-attempt difficulty score.
- We propose FISHER-LORA: an edit-intrinsic complexity read from the Fisher spectrum, an empirical rank floor, and a single-pass estimator together make rank a measurement on the (model, edit) pair rather than a setting on the editor.
- On COUNTERFACT and zsRE across three backbones, FISHER-LORA leads every baseline in Reliability; on LLaMA-2-7B its average rank is 11.4 against ELDER's fixed 32, and $\widehat{r}^{\star}$ predicts before any edit the joint failure of ROME, MEMIT and ELDER at $\rho = 0.74$.

2 MOTIVATION

To understand how much capacity a single factual edit requires, we audited the per-edit minimum rank on COUNTERFACT for LLaMA-2-7B and Mistral-7B; the LLaMA-2-7B reading below reveals three challenges.

Challenge #1: Rank without a measurement. To quantify how much capacity an edit actually needs, we searched for the minimum LoRA rank reaching $\geq 95\%$ reliability and $\geq 90\%$ locality on 100 COUNTERFACT edits applied to LLaMA-2-7B, then repeated the search over a $5{,}000$-edit pool. Figure 2 (left) plots what that search returns, decile by decile: the typical edit in the easiest tenth of the pool is served by a single direction, the median edit needs eleven, and the typical edit in the hardest tenth needs thirty, while published editors commit their rank before the edit is seen, ROME at $r = 1$ by construction, MELO at $r = 8$, and ELDER at $r = 32$. A rank chosen once for a benchmark has to cover the hardest edit it intends to serve, so on the median edit it is necessarily larger than that edit requires; the waste follows from committing before measuring, not from a badly chosen constant. *How can an editor allocate the rank an edit requires, neither more nor less, without prior knowledge of the edit?*

Challenge #2: Failure without a floor. To see what happens at the other end, we isolated the $1{,}500$ COUNTERFACT-HARD edits whose ground-truth $r^{\star}$ is at least 16 and measured every fixed-rank editor there. Figure 2 (middle) shows Reliability on the slice rising with the committed rank, from 40.2% at rank one to 83.3% for ELDER at rank thirty-two, still below the 95% that every single

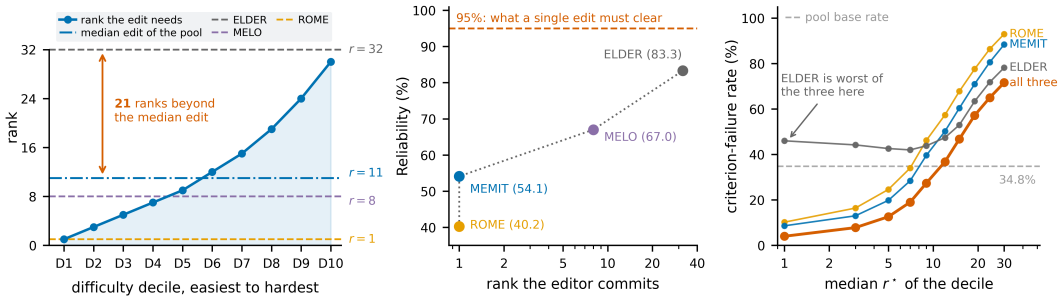


Figure 2: Measurements on LLaMA-2-7B. **Left:** rank needed per difficulty decile, against the ranks editors work at. **Middle:** reliability on COUNTERFACT-HARD against the committed rank. **Right:** criterion-failure rate of three editors and of all three together. Details in Appendix A.2.

edit has to clear. Aggregate reliability conceals this, since prior work reports a single number per benchmark and does not stratify failures by the rank an edit required. The failure is silent: a rank fixed before the edit arrives admits no comparison with what that edit required, so a benchmark average records that something failed, not that it failed for want of capacity. *Is there a model-level quantity that tracks the LoRA rank a correct editor requires, and can it be measured before the edit is attempted?*

Challenge #3: Difficulty without a budget. To test whether difficulty can be ordered before an edit is attempted, we cut the 5,000-edit pool into ten equal-size deciles by ground-truth $r^\star$ and measured the joint failure rate of the three baselines in each. Figure 2 (right) shows that rate climbing monotonically from 4.0% on the easiest decile to 71.6% on the hardest, against a 34.8% base rate over the pool. The individual curves split that ordering into two failure modes, over-provisioned on the easy end and under-provisioned on the hard one. ROME and MEMIT choose no rank, and their failures rise with $r^\star$ all the same: the difficulty the three share is legible only in how the model responds to the edit, which prior work obtains only by applying it. Practitioners are left to attempt an edit, observe failure and guess whether to retry, raise the rank or abandon, with no quantity that is edit-specific, predictive and cheap at once. *Is Challenge #2's minimum-rank quantity computable in a single backward pass, acting as both runtime difficulty predictor and per-edit rank budget?*

3 Problem Formulation

To address the three challenges of Section 2, we formulate single-fact knowledge editing as a minimum-rank feasibility problem:

$$r^\star(e) \;=\; \min_{\Delta\theta}\; r(\Delta\theta) \quad \text{s.t.} \quad M' \text{ is } (\varepsilon,\delta)\text{-correct on } e, \tag{1}$$

where M_θ is a frozen base model with parameters θ_0, an edit $e = (x_e, y_e^{\text{old}}, y_e^{\text{new}})$ asks it to answer x_e with y_e^{new}, and M' is the model edited by a perturbation $\Delta\theta$ of rank $r(\Delta\theta)$. Correctness has two halves, one on the edit itself and one on everything else:

$$\Pr_{M'}\big(y_e^{\text{new}} \mid x_e\big) \;\geq\; 1-\varepsilon, \qquad \mathbb{E}_{\mathcal{L}_e}\big[D_{\text{KL}}\big(M' \,\|\, M_{\theta_0}\big)\big] \;\leq\; \delta, \tag{2}$$

where ε and δ are the reliability tolerance and the locality budget, $\mathcal{L}_e$ is a set of unrelated prompts, and D_{KL} is the Kullback-Leibler divergence. We write $r^\star$ for that minimum, a property of the (M_{θ_0}, e) pair rather than of a configuration choice; the experiments measure it by rank search.

The minimised rank is what Challenge #1 finds unmeasured, fixed ahead of the edit by published editors, by construction or by configuration. The constraint pair is the reference point Challenge #2 lacks: an allocation is too small exactly when it leaves the pair infeasible. And $r^\star$ is obtainable only by applying the edit at each candidate rank, the cost Challenge #3 rules out.

4 Method

4.1 Fisher-LoRA Overview

To resolve the problem formulated in Section 3, we propose Fisher-LoRA, which turns LoRA rank from a method-level hyperparameter into an edit-specific quantity read from the model's information geometry. It takes a frozen model M_θ and one edit e, and returns a low-rank update over a fixed set of

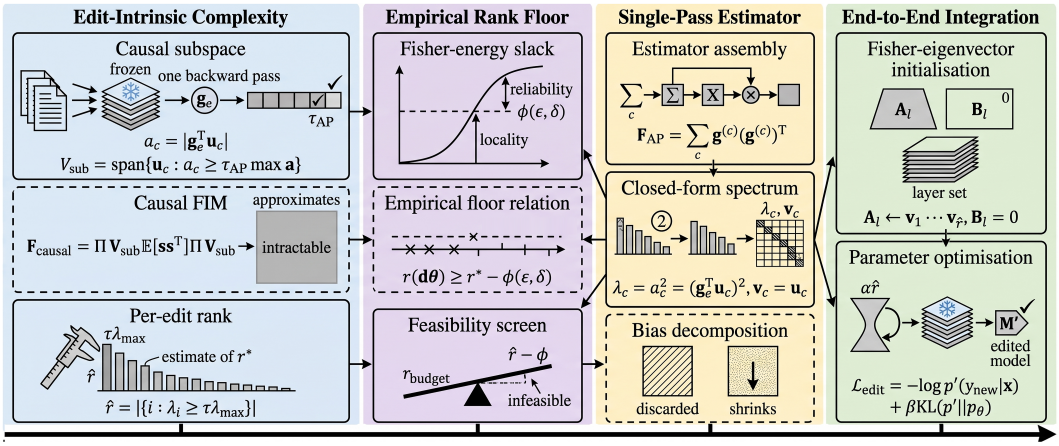


Figure 3: **Left:** the edit-intrinsic complexity reads $\widehat{r}^\star$ from the causal subspace. **Centre-left:** the empirical rank floor screens that allocation. **Centre-right:** the single-pass estimator assembles the spectrum both read. **Right:** initialisation and optimisation return the edited model M'.

layers, or an infeasibility flag when the edit asks for more capacity than the budget allows. FISHER-LORA has three components (Figure 3): the edit-intrinsic complexity reads a per-edit rank $\widehat{r}^\star$ from the model; the empirical rank floor gives that allocation a reference point for calling it too small; and the single-pass estimator makes the reading available before the edit is applied.

The three are one mechanism. The rank and the floor are read from the same spectrum, once, before any optimiser step, so the screen governs the allocation it tests and the estimator pays for both readings with a single backward pass. More details are provided in Appendix A.3.

4.2 EDIT-INTRINSIC COMPLEXITY

To address Challenge #1, FISHER-LORA estimates the edit-intrinsic complexity by a per-edit rank read from the model. Figure 3 (left) shows it first selecting the causal subspace from one backward pass, then restricting the Causal FIM to it, and finally reading that rank off the estimator's spectrum.

Causal subspace. Fix an orthonormal component basis $\{u_c\}$ whose elements are each supported on the input coordinates of one module in the layer set $\mathcal{S}$, and let $g_e = \nabla_\theta \log[p_\theta(y_{\text{new}} \mid x)/p_\theta(y_{\text{old}} \mid x)]$ be the contrastive gradient. Attribution patching scores each component in one backward pass,

$$a_c(e) \;=\; \big|\, g_e^\top u_c \,\big|, \tag{3}$$

and the causal subspace spans the components whose score exceeds a relative cutoff:

$$\mathcal{V}_e \;=\; \text{span}\big\{u_c : a_c(e) \geq \tau_{\text{AP}} \max_{c'} a_{c'}(e)\big\}, \tag{4}$$

with relative cutoff τ_{AP}, so $|\mathcal{V}_e|$ is the number of retained components.

Causal FIM. The Causal FIM restricts the next-token Fisher at x to $\mathcal{V}_e$: with $s_\theta(y) = \nabla_\theta \log p_\theta(y \mid x)$,

$$F_e^{\text{causal}} \;=\; \Pi_{\mathcal{V}_e}\, \mathbb{E}_{y\sim p_\theta(\cdot|x)}\big[s_\theta(y)\, s_\theta(y)^\top\big]\, \Pi_{\mathcal{V}_e}, \tag{5}$$

where $\Pi_{\mathcal{V}_e} \in \mathbb{R}^{p\times p}$, with p the number of base parameters, projects orthogonally onto $\mathcal{V}_e$, and the expectation is taken at the edit prompt x under the model's own predictive distribution.

Per-edit rank. Edit-intrinsic complexity is estimated by the effective rank, above a relative cutoff, of the single-pass estimator $\widehat{F}_e^{\text{AP}}$ constructed below for F_e^{causal},

$$\widehat{r}^\star(e) \;=\; \big|\big\{i : \lambda_i(\widehat{F}_e^{\text{AP}}) \geq \tau\, \lambda_{\max}(\widehat{F}_e^{\text{AP}})\big\}\big|, \tag{6}$$

with relative cutoff τ. We take $\widehat{r}^\star$, the Causal-Fisher eigenrank of Eq. (6), as the estimator of the operational minimum rank $r^\star$ of Eq. (1).

4.3 EMPIRICAL RANK FLOOR

To address Challenge #2, FISHER-LORA gives an allocation a reference point to check against. Figure 3 (centre-left) shows it first stating the slack a Cramér-Rao-style argument motivates, then the empirical floor observed at that slack, and finally screening the rank budget against it.

Fisher-energy slack. Under the linearisation of Appendix A.4, a Cramér-Rao-style heuristic ties reliability to the Fisher-weighted energy $\Delta\theta^\top F_e^{\text{causal}} \Delta\theta$ along the causal subspace, while locality caps the displacement $\|\Pi_{\mathcal{V}_e} \Delta\theta\|^2$, separated by the slack

$$\varphi(\varepsilon, \delta) \;=\; \tfrac{1}{\sqrt{2}} \sqrt{\log(1/\varepsilon)} \;+\; \sigma_\theta \sqrt{2\delta / \lambda^+_{\min}(F_e^{\text{causal}})}, \tag{7}$$

where $\lambda^+_{\min}$ is the smallest non-zero eigenvalue and σ_θ the linearisation-error constant, whose square is taken to bound the second-order remainder of the log-likelihood on the causal subspace. The first term comes from the reliability half of Eq. (2) and the second from the locality half. Details about the slack derivation and the scale conventions it assumes are provided in Appendix A.4 and A.8.

Empirical floor relation. Across the evaluation we observe an empirical rank floor $r(\Delta\theta) \gtrsim r^\star - \varphi(\varepsilon, \delta)$: editors whose per-layer rank falls below $r^\star - \varphi$ fail the (ε, δ)-correctness criterion at a rate that rises with the shortfall below the floor, and $\widehat{r}^\star$ predicts the joint cross-method failure of ROME, MEMIT, and ELDER. More details are provided in Appendix A.17.

Feasibility screen. We treat $r^\star - \varphi$ as an empirically supported floor, and the feasibility screen tests the rank budget against its estimate $\widehat{r}^\star - \varphi$. The reference value for $r^\star$ is measured by a rank search over the strongest LoRA-based baselines (Yu et al., 2024; Li et al., 2025) and never over FISHER-LORA, so the floor relates two separately measured quantities.

4.4 SINGLE-PASS ESTIMATOR

To address Challenge #3, FISHER-LORA reads that rank through a single-pass estimator built from the same backward pass that scores the components. Figure 3 (centre-right) shows the estimator first assembled from those scores, its spectrum then read in closed form, and its bias finally decomposed.

Estimator assembly. The exact F_e^{causal} is prohibitive at the model's $p \approx 7 \times 10^9$ parameters, requiring sampled per-token gradient outer products, so we replace it with the single-pass estimator, built by attribution patching

$$\widehat{F}_e^{\text{AP}} = \sum_{c:\, u_c \in \mathcal{V}_e} g_e^{(c)} \big(g_e^{(c)}\big)^\top, \qquad g_e^{(c)} = (g_e^\top u_c)\, u_c, \tag{8}$$

which approximates the restricted FIM by per-component contrastive outer products on the causal subspace, all read from the backward pass that computes those scores; its rank is at most $|\mathcal{V}_e|$.

Closed-form spectrum. Each $g_e^{(c)}$ is a scalar multiple of the basis vector u_c, so $\widehat{F}_e^{\text{AP}}$ is diagonal in the component basis $\{u_c\}$, with eigenpairs read directly off the component scores,

$$v_c \;=\; u_c, \qquad \lambda_c \;=\; \big(g_e^\top u_c\big)^2. \tag{9}$$

Each eigenvalue is the squared attribution score, $\lambda_c = a_c(e)^2$, so the top-$\widehat{r}^\star$ eigenvectors are obtained by sorting the scores of Eq. (3) and no eigendecomposition is formed.

Bias decomposition. The Frobenius bias of $\widehat{F}_e^{\text{AP}}$ relative to F_e^{causal} has two sources, reported separately: a linearisation error of order $\mathcal{O}(\sigma_\theta^2)$, with σ_θ as defined above, and an $\mathcal{O}(1)$ term from the diagonal approximation. The output is the eigenpair set $\{(\lambda_i, v_i)\}_{i=1}^{|\mathcal{V}_e|}$: the per-edit rank reads $\widehat{r}^\star$ from it, the empirical rank floor its energy estimate, and the integration step the LoRA initialisation. Details in Appendices A.6 and A.7.

4.5 END-TO-END INTEGRATION

A LoRA editor realises $\Delta\theta$ as per-layer pairs $B_\ell A_\ell$ of inner dimension r, so $r(\Delta\theta) = r$. FISHER-LORA chains the three components into one procedure whose phases Figure 3 (right) closes in order: Phase 1 reads $\widehat{r}^\star$ and Phase 2 applies the screen, both before any optimiser step.

Fisher-eigenvector initialisation. Phase 2 then initialises A_ℓ from the leading Fisher eigenvectors restricted to layer ℓ and zeros B_ℓ, so the initial output equals the unedited model. The update carries rank $\min(\widehat{r}^\star,\, r^{\mathrm{budget}})$ and starts aligned with $\mathcal{V}_e$ rather than with the ambient space.

Parameter optimisation. Phase 3 fits (A_ℓ, B_ℓ) by gradient descent on the standard editing loss

$$\mathcal{L}_{\mathrm{edit}}(e) \;=\; -\log p'(y_{\mathrm{new}} \mid x) \;+\; \beta\,\mathrm{KL}\big(p' \,\|\, p_\theta\big)\big|_{\mathcal{L}_e}, \tag{10}$$

where the KL term over the unrelated prompts $\mathcal{L}_e$ keeps their predictions near the unedited ones (locality) and $\beta > 0$ weights it. The complete procedure is summarised in Algorithm 1 (Appendix A.12), which returns the factors realising $\Delta\theta$ of Eq. (1) at rank $\min(\widehat{r}^\star,\, r^{\mathrm{budget}})$, or the infeasibility flag when the screen fires.

5 Experiments

5.1 Implementation Details

Datasets. We edit COUNTERFACT (Meng et al., 2022): 100 random edits, a 5,000-edit pool for the correlation, a disjoint 3,500-edit development split, and the 1,500-edit COUNTERFACT-HARD slice ($r^\star \geq 16$); plus 100 zsRE edits (Levy et al., 2017; De Cao et al., 2021). Reliability uses paraphrases, Locality neighbourhood prompts we call unrelated. Details in Appendix A.9.

Backbones and edit sites. All editing happens inside the feed-forward block located by causal tracing (Meng et al., 2022): layers 4 to 8 on LLaMA-2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023), and layers 20 to 25 on GPT-2-XL (Radford et al., 2019). One LoRA pair is attached per layer, and the block is held fixed across every editor. Details in Appendix A.10.

Hyperparameters. Each FISHER-LoRA edit is optimised with Adam at learning rate 1×10^{-3}, batch size 1, for $T = 25$ steps with seed 42, and all numbers are single-seed. The attribution cutoff is $\tau_{\mathrm{AP}} = 0.01$ and the eigenvalue cutoff $\tau = 0.01$, the locality-regulariser weight is $\beta = 0.1$, and the rank budget is $r^{\mathrm{budget}} = 40$, at which the feasibility screen fires at $\widehat{r}^\star \geq 42$. The same settings are used for every backbone, benchmark and edit. Details in Appendices A.11 and A.14.

Software and hardware. Every run uses a single NVIDIA A100 80GB GPU under one software stack shared by all editors, so the wall-clock column compares like with like, and Edit Cost includes the backward pass that yields $\widehat{r}^\star$. More details are provided in Appendix A.13.

Baselines. LoRA-based, MELO ($r = 8$) (Yu et al., 2024) and ELDER ($r = 32$) (Li et al., 2025); rank-one, ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023); memory-based, GRACE (Hartvigsen et al., 2023) and WISE (Wang et al., 2024b) as out-of-family probes; adaptive, AdaLoRA (Zhang et al., 2023b) and GaLore (Zhao et al., 2024) at our average rank. Rank one is architectural there; MELO and ELDER keep their published ranks. Details in Appendix A.15.

Metrics. Reliability (↑) is the share of target tokens ranked first over an edit's paraphrases, Locality (↑) the share of unrelated predictions that survive, each averaged over the edits; rank-floor ratio, difficulty correlation and Edit Cost follow (see Appendix A.16).

5.2 Main Results

Overall comparison. Table 1 reports Reliability on the 100 COUNTERFACT random edits across the three backbones, with Locality and Edit Cost on LLaMA-2-7B. FISHER-LoRA outperforms ELDER, the strongest fixed-rank editor in Reliability, by 3.3 Reliability points and 12.1 Locality points on LLaMA-2-7B at just over half the per-edit compute, and stays ahead of every baseline in Reliability on Mistral-7B and GPT-2-XL. Its average allocation is $\widehat{r}^\star = 11.4$ against ELDER's fixed $r = 32$, so the locality gain is realised at 36% of the parameter budget.

Difficulty stratification. Table 3 resolves the same comparison decile by decile on both 7B backbones. The rank-32 editor leads by under a point on the two easiest deciles, where the $r^\star$ medians are at most three and its update has capacity to spare, and on LLaMA-2-7B trails by 8.5 and 8.8 points on D9 and D10. Reliability does not charge for that surplus, and the main comparison records a 12.1-point Locality gap alongside it. The crossover sits at the third decile on both backbones.

Table 1: Main results on 100 COUNTERFACT random edits: Reliability (%, ↑), Locality (%, ↑), Edit Cost (s per edit, ↓); **bold** best, underlined second-best.

Method	Rank	LLaMA Rel.	LLaMA Loc.	Mistral Rel.	GPT-2-XL Rel.	Cost
ROME (Meng et al., 2022)	1	74.0	82.1	71.5	69.8	11.2
MEMIT (Meng et al., 2023)	1	80.6	84.7	78.2	76.3	14.0
MELO (Yu et al., 2024)	8	86.4	80.5	84.0	84.9	**9.4**
ELDER (Li et al., 2025)	32	93.1	79.2	92.0	90.4	19.7
AdaLoRA (Zhang et al., 2023b)	adaptive	90.3	85.9	88.7	87.1	22.5
GaLore (Zhao et al., 2024)	adaptive	87.1	84.1	83.4	88.0	20.1
FISHER-LORA (ours)	$\widehat{r}^{\star}$	**96.4**	**91.3**	**95.1**	**95.7**	10.3

Table 3: Reliability (%, ↑) by ground-truth $r^{\star}$ decile on the 5,000-edit pool, D1 easiest. More details are provided in Appendix A.17.

Method	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
LLaMA-2-7B										
ROME	98.9	97.5	95.1	90.1	81.1	75.6	66.1	53.6	40.9	26.2
MEMIT	99.5	98.5	96.1	93.4	88.0	82.4	74.7	63.3	54.4	44.5
MELO	98.6	98.2	97.0	94.8	87.7	82.0	76.0	75.2	66.6	59.2
ELDER	**99.8**	**99.1**	97.6	95.4	93.8	92.3	89.1	87.3	83.3	79.3
AdaLoRA	99.2	98.8	97.6	96.5	95.2	92.6	89.3	82.7	76.3	72.3
GaLore	99.3	98.4	95.5	92.0	88.6	84.0	79.8	76.7	69.4	58.7
FISHER-LORA	99.2	98.7	**98.6**	**98.2**	**96.8**	**96.0**	**93.7**	**93.8**	**91.8**	**88.1**
Mistral-7B										
ROME	98.4	96.7	93.7	87.5	78.9	72.1	62.4	48.3	35.8	21.0
MEMIT	99.1	98.1	95.1	91.7	86.4	79.8	71.3	59.0	48.0	38.1
MELO	98.6	97.4	96.1	93.0	82.1	78.0	72.7	71.4	62.8	58.4
ELDER	**99.8**	**99.0**	96.6	94.7	93.0	90.4	88.2	84.6	80.3	76.7
FISHER-LORA	99.5	98.9	**99.0**	**98.2**	**97.1**	**96.0**	**94.4**	**92.0**	**89.4**	**85.2**

The hard slice. Table 2 (left) reads the same comparison on COUNTERFACT-HARD, the regime where editors below $r^{\star} - \varphi \geq 14.7$ (with $\varphi \approx 1.27$) fail at a rate rising with the shortfall. The rank-one editors, an order of magnitude below the slice median $r^{\star} = 24$, fail on half or more of the edits; ELDER's $r = 32$ recovers most of them by sitting above that median, and FISHER-LORA recovers most of the remainder, at 91.2 Reliability and a 7.6% failure rate. On the hardest decile alone (500 edits with $r^{\star} \geq 28$) Reliability is 79.3 for ELDER against 88.1 for FISHER-LORA, so the margin over the strongest baseline widens with difficulty. The failure rate tracks the committed rank, from 68.4% at rank one to 15.8% at rank 32, which FISHER-LORA more than halves.

Table 2: COUNTERFACT-HARD (1,500 edits, $r^{\star} \geq 16$) and zsRE (100 edits) on LLaMA-2-7B, in %; best in **bold**. More details are provided in Appendix A.18.

		COUNTERFACT-HARD		zsRE	
Method	**Rank**	**Rel. (↑)**	**Failure (↓)**	**Rel. (↑)**	**Loc. (↑)**
ROME	1	40.2	68.4	76.2	83.1
MEMIT	1	54.1	52.7	81.5	86.1
MELO	8	67.0	36.9	88.6	86.5
ELDER	32	83.3	15.8	93.8	81.0
FISHER-LORA	$\widehat{r}^{\star}$	**91.2**	**7.6**	**96.9**	**92.4**

Cross-benchmark transfer. Table 2 (right) repeats the headline comparison on zsRE, whose edits are constructed by a different protocol. The average $\widehat{r}^{\star}$ chosen there is 9.7 against the COUNTERFACT average of 11.4, and the margins over ELDER hold at 3.1 Reliability points and 11.4 Locality points, at 96.9 against 93.8 and 92.4 against 81.0, tracking the COUNTERFACT result to within about one point. Two benchmarks built by different protocols therefore place the allocation at a similar level on LLaMA-2-7B and leave the margin it buys intact.

Over-provision across backbones. The rank column of Table 1 understates how much capacity a fixed rank carries: read against the ground-truth decile medians and the occupancy of each population, ELDER's mean over-provision is 20.48 dimensions on LLaMA-2-7B, 23.73 on Mistral-7B and 7.67 on COUNTERFACT-HARD, MELO's is 1.84, 2.50 and 0, and the two rank-one editors never over-provision on any decile. Over-provision is therefore a property of the pairing between a fixed rank and the population it is applied to, not of the editor taken on its own.

5.3 Ablation Study and Analysis

Component ablation. Table 4 lesions the two constructive components in turn, the per-edit rank with the causal subspace it selects and the Fisher-eigenvector initialisation. Confining the update to some $\widehat{r}^\star$-dimensional subspace buys 2.5 Reliability points, swapping that random subspace for the attribution-patched causal one at the identical rank a further 4.2, and the warm start 5.5 more. Rank must therefore be allocated inside the right subspace, not merely a small one, and the 9.7 points between the random subspace and the full method are what the estimator's spectral alignment buys. Locality moves the same way, from 79.2 at the fixed-rank reference to 91.3, and the one step where it does not, the 0.4 points returned when the random subspace gives way to the causal one, is paid for by 4.2 Reliability points at an unchanged rank. The first two rows separate the allocation from what it is spent on: the rank-only row costs 8.9 Reliability points against the fixed-rank reference while gaining 6.0 of Locality, and the subspace and the initialisation recover both at the smaller footprint. Details in Appendix A.22.

Threshold sensitivity. Table 5 sweeps the eigenvalue threshold: halving the deployed τ inflates the average rank to 18.7 and costs 5.3 Locality points for 0.2 Reliability, while raising it to 0.05 starves the edit, dropping Reliability to 89.0 and to 81.7 at $\tau = 0.1$, where the allocation falls to 5.1. Locality rises monotonically over the sweep while Reliability peaks at $\tau = 0.005$. At $\tau = 0.001$ the allocation reaches 32.6 and both readings come within half a point of the rank-32 editor's, because so large an allocation barely restricts: what confining the update buys at $\widehat{r}^\star \approx 11$ is given back once the rank grows into it.

Table 4: Component ablation on COUNTERFACT-100 / LLaMA-2-7B, in %; Drop is against the full method.

Variant	Rel. (↑)	Loc. (↑)	Drop
Fixed-rank reference	93.1	79.2	−3.3
Rank $\widehat{r}^\star$ only	84.2	85.2	−12.2
+ random subspace	86.7	88.0	−9.7
+ causal subspace	90.9	87.6	−5.5
+ Fisher init (full)	**96.4**	**91.3**	0.0

Estimator alternatives. Table 6 compares the estimator against two substitutes: the oracle costs 620× the per-edit compute and buys back 0.3 Reliability points and 0.2 of Locality, whereas random projection at matched rank loses 9.7 Reliability and 3.3 Locality points. The decisive design choice is therefore the causal-Fisher approximation, not low-rank structure as such.

Scale invariance. The reading is invariant to the scale of the contrastive gradient. Because the cutoff of Eq. (6) is relative to $\lambda_{\max}$, rescaling that gradient by any positive factor multiplies every eigenvalue by the square of the factor and leaves $\widehat{r}^\star$ unchanged, while the gradient norm on the causal subspace scales with the factor itself. The two are different readings of one backward pass: the squared norm is the total mass of the spectrum, $\widehat{r}^\star$ the count of entries at least τ times the largest. At a fixed norm every value of $\widehat{r}^\star$ the estimator can return remains attainable, so no monotone function of that norm reproduces the count. More details are provided in Appendix A.21.

Rank-floor agreement. Across the 100 random edits on LLaMA-2-7B the ratio $r_{\text{used}}/r^\star$ has mean 1.02, with fifth and ninety-fifth percentiles at 0.96 and 1.10, and $r_{\text{used}} - r^\star$ is at most one dimension on 89% of them and at most two on 97%. The one-sided slack of Eq. (7) is $\varphi = 1.27$ dimensions at $\varepsilon = 0.05$, $\delta = 0.1$, $\sigma_\theta = 0.1$ and $\lambda^+_{\min}(F_e^{\text{causal}})$ taken at 1 on all three backbones. The same search restricted to FISHER-LORA agrees with the cross-method ground truth within one dimension on 84% of them, so that ground truth tracks the edit, not the editors searched over.

Difficulty correlation. Figure 2 (right) stratifies the joint failure rate of ROME, MEMIT, and ELDER by ground-truth rank on the 5,000-edit pool, where it climbs from 4.0% to 71.6%. An editor fails an edit when it misses the (ε, δ)-correctness criterion of Eq. (2) there, and the per-edit $\widehat{r}^\star$ correlates with the intersection of those three events at Spearman $\rho = 0.74$. Since $\widehat{r}^\star$ comes from a single backward pass on the unedited model, this is a pre-attempt difficulty score independent of the editor, and it survives extension to out-of-family editors GRACE and WISE at $\rho = 0.71$.

Boundary conditions. Three regimes concentrate the residual failures, each characterised on the 5,000-edit pool: edits whose causal subspace lies in a near-null direction of F_e^{causal} ($\lambda^+_{\min} < 10^{-6}$) destabilise the $\widehat{r}^\star$ estimate, with Reliability 71.4% on the 14 such edits against 95.5% pool-wide; edits with $\widehat{r}^\star > 40$ are clipped or screened; and targets in semantically incompatible contexts trade

Table 5: Eigenvalue-threshold sweep on LLaMA-2-7B / 100 COUNTERFACT edits, in %; bold marks the deployed setting. More details are provided in Appendix A.20.

τ	0.001	0.005	**0.01**	0.05	0.1
Average rank $\bar{r}$	32.6	18.7	**11.4**	7.2	5.1
Reliability (↑)	93.5	96.6	**96.4**	89.0	81.7
Locality (↑)	79.4	86.0	**91.3**	93.4	94.6

Table 6: Estimator alternatives on COUNTERFACT-100 / LLaMA-2-7B, in %. More details in Appendix A.19.

Variant	Rel. (↑)	Loc. (↑)
Single-pass estimator (deployed)	96.4	91.3
Full causal Fisher (oracle)	96.7	91.5
Random projection	86.7	88.0

Reliability for Locality. The evidence otherwise covers single-fact edits on three English decoder-only backbones at 7B or below under a rank ceiling of 40, with Locality on LLaMA-2-7B alone.

Computational complexity. Per edit, FISHER-LORA runs one backward pass for g_e and sorts the $|\mathcal{V}_e|$ component scores, and the spectrum is read from those scores in closed form, so no $p \times p$ matrix is formed and no eigendecomposition run. What follows the reading is the same T optimiser steps any LoRA editor runs, over at most r^{budget} directions in each layer of $\mathcal{S}$, so the parameters written per edit are $\mathcal{O}(d\,\widehat{r}^{\star}\,|\mathcal{S}|)$ with d the module width. The wall-clock consequence is the Edit Cost column of the main comparison.

6 Related Work

Editing a fact in place. One family rewrites the base model: ROME installs a rank-one association in a causally traced MLP layer (Meng et al., 2022), MEMIT batches it over thousands of facts (Meng et al., 2023), AlphaEdit projects the write into the null space of preserved knowledge (Fang et al., 2025), and MEND predicts the delta with a hyper-network (Mitchell et al., 2022a). A second freezes the weights and routes around them: GRACE's codebook of corrections (Hartvigsen et al., 2023) and WISE's gated side memory (Wang et al., 2024b). Neither sizes the edit: the first fixes the update form by algorithm choice, the second moves the footprint into a side store.

Choosing a rank. A third parameterises the update as a low-rank product. LoRA fixes the rank a priori (Hu et al., 2022); AdaLoRA reallocates it by an importance schedule (Zhang et al., 2023b), ReLoRA accumulates higher rank from restarted low-rank updates (Lialin et al., 2024), and VeRA, QLoRA and DoRA cut the cost of a rank already chosen (Kopiczko et al., 2024; Dettmers et al., 2023; Liu et al., 2024). Two editors transplant the machinery at a fixed rank, MELO at eight (Yu et al., 2024) and ELDER at thirty-two (Li et al., 2025), and intrinsic-dimensionality analysis measures how few directions a whole task needs by fitting that number on a training set (Aghajanyan et al., 2021). We differ on the unit of adaptation: every schedule here is fitted to a training distribution, so none is available before one edit is attempted.

Reading the geometry of an update. A fourth supplies tools, not an editor. Natural gradient uses the Fisher as the metric of parameter space (Amari, 1998), K-FAC makes it tractable (Martens & Grosse, 2015), and EWC reads its diagonal as parameter importance (Kirkpatrick et al., 2017). Causal tracing locates a fact in mid-layer writes (Meng et al., 2022), and attribution patching estimates each component's contribution in one pass (Syed et al., 2024; Vig et al., 2020), which ACDC assembles into circuits (Conmy et al., 2023). Both stop short of a rank: the Fisher is taken over a task distribution, the subspace read as a diagram. We form its spectrum on one edit and read a per-edit budget from it.

7 Conclusion

We asked how much capacity a single factual edit requires, a quantity editors fix in advance. FISHER-LORA reads that rank from the component spectrum of a single-pass Causal Fisher estimate, screens it against an empirical rank floor, and initialises the factors there. On 100 random COUNTERFACT edits it reaches 96.4% Reliability and 91.3% Locality on LLaMA-2-7B at an average rank of 11.4 against ELDER's fixed 32, and on a 5,000-edit pool the same reading predicts joint failure of ROME, MEMIT and ELDER at $\rho = 0.74$. Edit difficulty is, to first order, a property of the (model, edit) pair, not of the editor, and readable before the edit is attempted.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021. doi: 10.18653/v1/2021.acl-long.568.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, pp. 251–276, 1998. doi: 10.1162/089976698300017746.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.14997.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6491–6506, 2021. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522.

Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, and Koray Kavukcuoglu. Natural neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. URL https://arxiv.org/abs/1507.00210.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.14314.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29c8c615b3187ee995029284702d3f43-Paper-Conference.pdf.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.02997.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47934–47959. Curran Associates, Inc., 2023. doi: 10.52202/075280-2079.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2301.09785.

Albert Q. Jiang et al. Mistral 7b. *arXiv:2310.06825*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting

in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/10.1073/pnas.1611835114.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11454.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, pp. 333–342, 2017. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034.

Jiaang Li, Quan Wang, Zhongnan Wang, Yongdong Zhang, and Zhendong Mao. Elder: Enhancing lifelong model editing with mixture-of-lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReloRA: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DLJznSp6X3.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.09353.

Shuxin Liu and Jianhao Zhang. CAKE: Causal-guided adaptive knowledge editing for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. doi: 10.18653/v1/2026.acl-long.918.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2408–2417, 2015. URL https://arxiv.org/abs/1503.05671.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.07229.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022a. URL https://arxiv.org/abs/2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 15817–15831, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1301.3584.

Alec Radford et al. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 4528–4537, 2018. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024. URL https://arxiv.org/abs/2310.10348.

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

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 3274–3287, 2023. doi: 10.18653/v1/2023.eacl-main.239. URL https://aclanthology.org/2023.eacl-main.239.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Haoyu Wang, Tianci Liu, Ruirui Li, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. RoseLoRA: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a. doi: 10.18653/v1/2024.emnlp-main.57.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://arxiv.org/abs/2405.14768.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. Melo: Enhancing model editing with neuron-indexed dynamic lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19449–19457, 2024. doi: 10.1609/aaai.v38i17.29916. URL https://arxiv.org/abs/2312.11795.

Feiyu Zhang, Liangzhi Li, Junhao Chen, Zhouqiang Jiang, Bowen Wang, and Yiming Qian. Increlora: Incremental parameter allocation method for parameter-efficient fine-tuning. *arXiv preprint arXiv:2308.12043*, 2023a. URL https://arxiv.org/abs/2308.12043.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://arxiv.org/abs/2303.10512.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.03507.

AI USE STATEMENT

We used generative AI tools to translate the text (Claude Opus 5.0), to check grammar and spelling and polish the wording (GPT-5.6), to guide the drawing of the figures (Gemini), and to check the content (Grok). We did not use generative AI tools to modify code. We have reviewed all AI-assisted work: we checked every reference for hallucinated citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The details needed to reproduce the results of this paper are disclosed in this appendix: the datasets and their splits, the backbones and edit sites, the training schedule, the hyperparameter grid together with the selected values, the editing procedure in algorithmic form, the baseline reproduction protocol, the metric definitions and their aggregation, and the software and hardware setup.

APPENDIX

Table A1 lists the contents of each part of the appendix.

Table A1: What each part of the appendix holds.

Section	Contents	Section	Contents
A.0	Limitations	A.12	The editing procedure
A.1	Extended related work	A.13	Hardware and software
A.2	The motivation panels in detail	A.14	Training hyperparameters
A.3	Why three components	A.15	Baseline reproduction protocol
A.4	The Fisher-weighted energy argument	A.16	Metric definitions and aggregation
A.5	Empirically observed consequences	A.17	Per-decile reliability decomposition
A.6	The estimator-bias argument	A.18	Full zsRE results and stratification
A.7	Calibration of the bias constant C	A.19	Estimator and rank-floor agreement
A.8	Scale and notation conventions	A.20	Sensitivity analyses
A.9	Datasets	A.21	Scale invariance of $\widehat{r}^{\star}$
A.10	Backbones and edit sites	A.22	The ablation ladder and point projection
A.11	Training schedule and optimisation	A.23	Future work

APPENDIX A.0: LIMITATIONS

The framework is defined for one edit at a time. The quantity this paper measures is a property of a single (model, edit) pair. The estimator is assembled from the gradient of one fact, and the rank it returns sizes one update to a model that was clean when that fact arrived. Editing a model repeatedly is governed by a different object, the joint curvature of a whole set of facts, and that object is neither the matrix studied here nor anything this one reduces to. The paper therefore lacks a per-edit rank for the setting in which edits are applied one after another, and supplying one is a separate piece of work rather than a further measurement on the construction given here.

The rank ceiling is a choice of ours, not a reading from the model. The per-edit rank is capped at 40 so that every edit stays inside a LoRA parameter budget we fixed in advance. That ceiling is a deployment decision and carries no information about the edit in front of it. On an edit whose reading exceeds the ceiling the allocation is the ceiling, so above 40 the paper offers no account of what the right allocation would have been. A different parameter budget moves that boundary, and nothing here establishes where it should sit.

The evidence base is three dense English transformers. Every number reported in this paper was measured on LLaMA-2-7B, Mistral-7B and GPT-2-XL, which are dense decoder-only transformers trained on English at seven billion parameters or below, using two single-fact editing benchmarks. The paper establishes nothing about how a per-edit rank behaves in a mixture-of-experts model, in

an encoder-decoder, at a substantially larger parameter count, or in a language other than English. Those are the architectures a reader may care most about, and they lie outside what we measured.

APPENDIX A.1: EXTENDED RELATED WORK

The positioning of FISHER-LORA against the prior lines is stated in the body. What follows is the mechanism-level account of what each prior construction computes and at what cost, together with the two lines the body names only in passing, auxiliary-memory editors and the wider causal-localisation literature; two surveys catalogue the editing lines as a whole (Yao et al., 2023; Zhang et al., 2024).

Direct parameter rewriting. The earliest line of factual editing for transformer language models locates a fact in the residual stream and rewrites a small set of weights to install a chosen association. The foundational causal-tracing analysis demonstrates that mid-layer MLP residual writes carry the bulk of factual content and isolates a small per-fact set of layers responsible for any individual association (Meng et al., 2022). ROME's closed-form rank-one update treats the chosen MLP layer as an associative key-value memory and inserts one outer-product correction per fact (Meng et al., 2022); MEMIT extends the construction to thousands of facts in a single batched least-squares solve and demonstrates that the rank-one parameterisation scales without architectural change (Meng et al., 2023). PMET disentangles the attention and feed-forward contributions to the residual write and optimises the transformer-component representations separately before the rank-one insertion (Li et al., 2024). The KnowledgeNeurons line localises facts to specific neurons and then edits them by writing to the corresponding value vectors, providing an alternative read on where factual information lives but inheriting the rank-one update structure (Dai et al., 2022). AlphaEdit adds a null-space constraint that projects every parameter update into the orthogonal complement of preserved-knowledge directions, guarding locality through a linear-algebraic side condition (Fang et al., 2025). Two hypernetwork meta-learners predict the parameter delta from an auxiliary network: KE introduces the constrained meta-learning formulation (De Cao et al., 2021), and MEND scales it through low-rank gradient decomposition that drops the per-fact compute by an order of magnitude (Mitchell et al., 2022a). The update rank in this family is fixed at one (ROME, MEMIT, AlphaEdit) or implicitly at the dimension of the meta-network's output (KE, MEND). The rank-one parameterisation is most defensible for the easiest facts, and the experiments in Section 5.2 of the main body show rank-one editors collapsing to below fifty-five percent reliability on COUNTERFACT-HARD where the intrinsic rank is at least sixteen.

Auxiliary-memory and routing editors. A second line leaves the base model frozen and routes around it through external memory or scope classifiers. T-Patcher attaches an external patch neuron per edit and grows the model instead of rewriting it, eliminating the rank question at the cost of architectural drift over a long edit sequence (Huang et al., 2023). GRACE maintains a discrete codebook of layer-wise corrections that is consulted at inference time to override the base model on prompts matching a stored key (Hartvigsen et al., 2023). WISE separates a side memory from the main memory under an activation gate, addressing the catastrophic-interference problem of long edit streams while keeping the base model untouched (Wang et al., 2024b). SERAC caches each edit and dispatches via a learned scope classifier that decides at inference whether the cached patch applies to the current prompt (Mitchell et al., 2022b). Auxiliary-memory editors trade parameter cost for architectural overhead: each edit either grows the model (T-Patcher) or grows a side store (GRACE, WISE, SERAC), and the per-edit cost in scoped retrieval and gating compute can dominate the per-edit cost of weight rewriting after enough edits accumulate. The parameter footprint is moved into a different ledger, not eliminated.

Low-rank adaptation and adaptive-rank parameter-efficient fine-tuning. A third line parameterises the update as a low-rank product and develops machinery for choosing the rank. LoRA fixes the rank a priori and trains a product BA on top of frozen weights, achieving parameter-efficient adaptation by reducing the trainable degrees of freedom by orders of magnitude (Hu et al., 2022). AdaLoRA reallocates rank across weight matrices through an importance-based singular-value schedule that runs throughout training, reading importance from accumulated gradient signal on a training distribution (Zhang et al., 2023b). IncreLoRA grows the rank monotonically during training as new directions accumulate importance (Zhang et al., 2023a); DyLoRA trains across a

rank range jointly so any rank within the range is usable at inference (Valipour et al., 2023); ReLoRA periodically restarts the low-rank product to accumulate a higher-rank effective update through a sequence of low-rank corrections (Lialin et al., 2024). Compression-side variants include VeRA, which shares random projections across layers and trains only diagonal rescalings (Kopiczko et al., 2024), QLoRA, which stacks 4-bit quantisation on top of LoRA for memory savings (Dettmers et al., 2023), and DoRA, which decomposes the pre-trained weight into magnitude and direction and confines the low-rank update to the directional component (Liu et al., 2024). Two LoRA-based editors transplant the machinery to factual editing: MELO routes through neuron-indexed dynamic LoRA blocks at fixed rank eight (Yu et al., 2024), and ELDER mixes a bank of fixed-rank LoRA experts at rank thirty-two (Li et al., 2025). RoseLoRA imposes row- and column-wise sparsity on the low-rank factors for knowledge editing (Wang et al., 2024a), and CAKE adapts the edit under causal guidance (Liu & Zhang, 2026). AdaLoRA's importance schedule, in particular, would need to be re-run from scratch per edit to act as a per-edit rank selector. FISHER-LORA reads the rank from a single backward pass instead, and the comparison in Section 5.3 of the main body shows the per-edit rank is read at more than two orders of magnitude lower compute than even an oracle Fisher accumulator. The compression line (VeRA, QLoRA, DoRA) is orthogonal to the rank-selection question and could be composed on top of FISHER-LORA; we leave such compositions to future work.

Information geometry and natural-gradient methods. The fourth lineage that FISHER-LORA draws on supplies the analytical primitives for converting reliability and locality into an energy constraint. Natural-gradient methods reparameterise updates by the Fisher information matrix and replace Euclidean steepest descent with a Riemannian one, motivated by the observation that the parameter coordinate system distorts the geometry of the loss surface (Amari, 1998; Pascanu & Bengio, 2014). K-FAC introduces a Kronecker-factored approximation of the Fisher that is tractable on layered architectures and makes natural-gradient steps practical at modern scale (Martens & Grosse, 2015). Natural neural networks fold the same approximation into the forward pass through whitened activations, providing a complementary route to a low-cost natural-gradient direction (Desjardins et al., 2015). Elastic weight consolidation weights each parameter direction by its diagonal Fisher entry to anchor old tasks while learning new ones. It treats the Fisher as a per-parameter importance score for catastrophic-forgetting protection (Kirkpatrick et al., 2017); Progress and Compress reuses the construction across a teacher-student schedule and connects the Fisher view to consolidation across many tasks (Schwarz et al., 2018). Across this lineage the Fisher is used ambiently, over the whole parameter vector and averaged over a task distribution, and not restricted to a subspace identified per input, which is the restriction Eq. (5) imposes. The analysis applies the Cramér-Rao argument of Appendix A.4 to a behavioural reliability-locality criterion: the high-level edit-success specification is carried into a quadratic energy constraint that empirically tracks the Fisher rank.

Causal localisation and mechanistic interpretability. The final lineage we draw on supplies the per-input subspace on which the Causal Fisher is restricted. The foundational causal-tracing procedure that locates factual associations to mid-layer MLP residual writes by patching activations between paired runs is developed in Meng et al. (2022), with the mass-editing logistics worked out in Meng et al. (2023). Attribution patching distils causal-tracing into a per-component contribution estimate readable from a single backward pass, replacing the expensive iterated-patching procedure with a one-shot gradient-based proxy; the contrastive variant for language models that we use was sharpened in Syed et al. (2024). The technique builds on causal-mediation analysis of transformer components (Vig et al., 2020; Geiger et al., 2021) and on the automated circuit-discovery procedure of Conmy et al. (2023), which identifies sparse subgraphs of the transformer responsible for individual behaviours. The per-layer residual-stream decomposition we restrict to follows the same paired-run patching formalism used by these causal-tracing analyses. Concretely, these methods deliver $\mathcal{V}_e$ as a side product of the same backward pass that produces the attribution scores, which is what makes the projector $\Pi_{\mathcal{V}_e}$ of Eq. (5) free at the point of use. In this literature the subspace is then read qualitatively, as a circuit diagram, and its spectrum is never formed.

Closing the gap. The two additional families surveyed here, auxiliary-memory editors and the wider causal-localisation line, do not change the positioning stated in Section 6: the former removes the parameter footprint from the edit instead of sizing it, and the latter supplies the per-input subspace without connecting it to any rank quantity.

APPENDIX A.2: THE MOTIVATION PANELS IN DETAIL

The three panels of Figure 2 are reproduced here at full width, one per figure, with the reading each is meant to support.

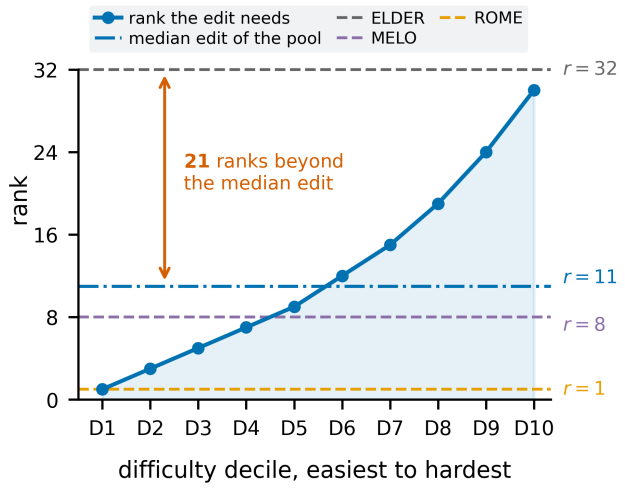


Figure A1: Figure 2 (left) at full width: the median ground-truth $r^\star$ of each of the ten difficulty deciles of the 5,000-edit pool, against the fixed ranks ROME, MELO and ELDER work at, with the pool median edit and its distance to ELDER's rank 32 marked.

Left panel: the rank an edit needs, against the rank editors commit. The 5,000-edit COUNTERFACT pool is sorted by ground-truth $r^\star$ and cut into ten equal-size deciles of 500 edits, D1 the easiest and D10 the hardest. The solid curve plots the median $r^\star$ of each decile, namely $1, 3, 5, 7, 9, 12, 15, 19, 24$ and 30, so each marker stands for one tenth of the pool and the shaded area beneath the curve is the capacity those edits ask for. The curve rises gently across the first half and steeply over the last two deciles, which is the right tail of the distribution: rank demand is not concentrated on any one value. The three dashed horizontals mark the ranks published editors work at before any edit arrives: ROME at 1, MELO at 8, ELDER at 32. Each is flat by construction, so the vertical gap between a horizontal and the curve is what that editor over-supplies or under-supplies on the decile beneath it; ROME meets only D1, MELO is exceeded from D5 on, and ELDER sits above every decile median in the pool. The dash-dotted line marks the median edit of the pool, which needs 11 dimensions. The double arrow spans the 21 dimensions between the median edit and ELDER's fixed 32, capacity the median edit never asks for yet pays for on every application.

Middle panel: below the floor, failure is silent. Each marker is one fixed-rank editor on the 1,500-edit COUNTERFACT-HARD slice, placed at the rank it works at and the Reliability it attains: ROME at rank 1 reaches 40.2, MEMIT at rank 1 reaches 54.1, MELO at rank 8 reaches 67.0, and ELDER at rank 32 reaches 83.3. The dotted line through them is a guide to the eye. The horizontal dashed line marks the 95% per-edit level of Eq. (2), and the points are slice means. Reliability rises with the committed rank across four editors that differ in optimiser, loss and update rule.

Right panel: difficulty ordered before the edit is attempted. The 5,000-edit pool is cut into ten equal-size deciles of 500 edits by ground-truth $r^\star$, and each decile is plotted at its median $r^\star$, namely $1, 3, 5, 7, 9, 12, 15, 19, 24$ and 30. The three thin curves are the per-decile criterion-failure rates of ROME, MEMIT and ELDER, an editor failing on an edit when it misses the correctness criterion there. ROME climbs from 10.2% to 93.0% and MEMIT from 8.6% to 88.4%, both monotone. ELDER is not monotone: it starts at 46.0% on the easiest decile, falls to 42.0% at the fourth, then climbs to 78.2% on the hardest. Its easy-end failures are in the main locality failures rather than reliability failures: a rank-32 update carries far more free directions than a decile-1 edit needs, and the excess leaks into the locality set, so a share of those edits misses the criterion however well the edit itself takes. The thick curve is the joint failure rate of ROME, MEMIT and ELDER, the fraction

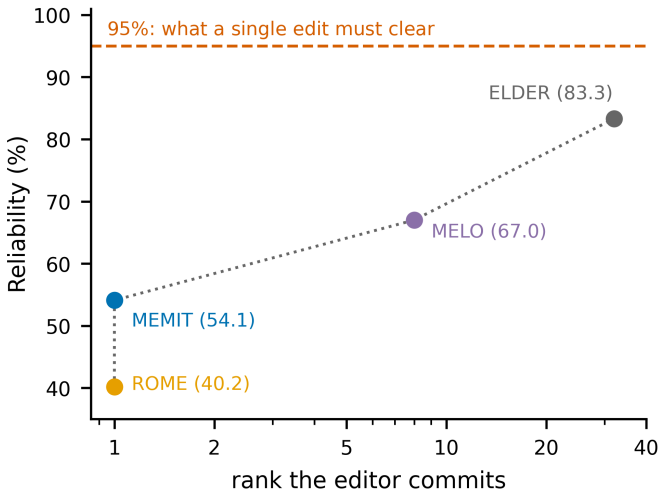


Figure A2: Figure 2 (middle) at full width: slice-mean Reliability on the 1,500-edit COUNTERFACT-HARD slice for four fixed-rank editors, each placed at the rank it works at, against the 95% per-edit level of Eq. (2).

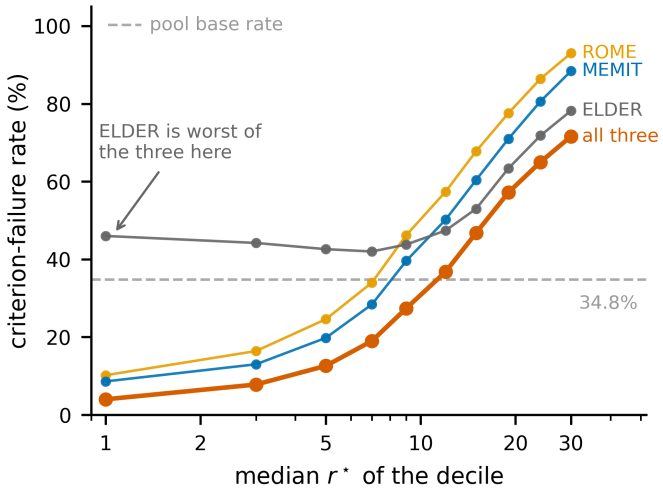


Figure A3: Figure 2 (right) at full width: the per-decile criterion-failure rates of ROME, MEMIT and ELDER on the 5,000-edit pool and the rate at which all three miss the criterion together, against the 34.8% base rate over the pool.

of edits on which all three miss the correctness criterion together; it climbs from 4.0% on the easiest decile to 71.6% on the hardest, a span of 67.6 points. The dashed horizontal is the 34.8% base rate over the whole pool, which is what a difficulty-blind predictor would report for every decile. Joint failure is monotone in the decile order, so ordering edits by the rank they require already separates the edits all three editors miss from those they do not, and the same axis lays the two failure modes out at opposite ends, over-provisioning on the easy edits and under-provisioning on the hard ones. On the easiest decile the split between the two failure modes is at its widest: ROME and MEMIT miss the criterion on 10.2% and 8.6% of those edits while ELDER, at its fixed rank 32, misses on 46.0%. The rank search that fixes $r^\star$ runs over LoRA editors, whereas ROME and MEMIT insert a closed-form rank-one update in which no rank is ever chosen; their failure rates rise with $r^\star$, which makes the ordering a property of the edits. The ordering plotted here is the ground-truth one, recovered by a rank search that applies each edit; whether the same ordering can be read before the edit is applied is the question the method answers.

APPENDIX A.3: WHY THREE COMPONENTS

The three components carry three distinct signals. The edit-intrinsic complexity carries the *quantity* of rank: without the edit-intrinsic complexity there is no per-edit number to allocate, and the editor falls back to a fixed budget that reincarnates ELDER's over- and under-allocation (the fixed-rank reference row is exactly that fixed-r=32 behaviour). The empirical rank floor carries the *feasibility* of an allocation: the feasibility screen of Algorithm 1 tests the budget against $\widehat{r}^\star - \varphi$ before the edit is attempted. The single-pass estimator carries the *tractability*: without the single-pass estimator, $\widehat{r}^\star$ requires an oracle Fisher accumulation more than two orders of magnitude more expensive, which removes the runtime-difficulty-predictor property entirely. The count is therefore three because the editing specification decomposes into exactly three sub-objectives: allocate the right rank, flag when an allocation is too small, and compute it cheaply. The three are distinct objects over the same Causal-Fisher spectrum: a spectral count, an inequality, and an estimator. A naive editor that kept only the edit-intrinsic complexity would allocate a plausible rank but in the wrong subspace, which is the failure the *rank-only* ablation row exhibits. The ablation ladder of Section 5.3 measures the two constructive collapses directly, while the empirical rank floor is carried by the feasibility screen, not by an accuracy row.

APPENDIX A.4: THE FISHER-WEIGHTED ENERGY ARGUMENT

Two heuristic arguments stand behind the method: a Cramér-Rao-style estimate of the Fisher-weighted update energy, and a Frobenius bias estimate for the attribution-patching estimator. This appendix also states and records the status of the empirically observed rank floor. We fix notation, then give each derivation. Throughout, $\theta_0 \in \mathbb{R}^p$ denotes the frozen base parameters of the language model f_θ; an edit is a triple $e = (x, y_{\text{old}}, y_{\text{new}})$ with locality set $\mathcal{L}_e$; the causal subspace $\mathcal{V}_e \subseteq \mathbb{R}^p$ has orthogonal projector $\Pi_{\mathcal{V}_e}$ and dimension $|\mathcal{V}_e|$; and $F_e^{\text{causal}} = \Pi_{\mathcal{V}_e} \cdot F_e \cdot \Pi_{\mathcal{V}_e}$ denotes the Causal FIM obtained by restricting the next-token Fisher $F_e = \mathbb{E}_{y\sim p_{\theta_0}(\cdot|x)}[\nabla_\theta \log p_{\theta_0}(y \mid x) \nabla_\theta \log p_{\theta_0}(y \mid x)^\top]$ to $\mathcal{V}_e$. We write $\lambda_i(\cdot)$ for the i-th largest eigenvalue and $\lambda^+_{\min}(\cdot)$ for the smallest non-zero eigenvalue.

Heuristic Argument 1 (Fisher-weighted energy)**.** *Let σ_θ^2 upper-bound the linearisation error of f_{θ_0} on $\mathcal{V}_e$, and let $\lambda^+_{\min}(F_e^{\text{causal}})$ denote the smallest non-zero eigenvalue of the Causal Fisher. Under a local linearisation and a Cramér-Rao-style approximation, an (ε, δ)-correct editor with update $\Delta\theta$ is heuristically expected to place Fisher-weighted energy along the causal subspace of order*

$$\begin{aligned} \Delta\theta^\top F_e^{\text{causal}} \Delta\theta &\gtrsim \log(1/\varepsilon) - O(\varepsilon), \\ \|\Pi_{\mathcal{V}_e} \Delta\theta\|^2 &\leq \tfrac{2\delta}{\lambda^+_{\min}(F_e^{\text{causal}})} + O(\sigma_\theta^2), \end{aligned} \tag{A1}$$

with associated closed-form slack $\varphi(\varepsilon, \delta) = \frac{1}{\sqrt{2}}\sqrt{\log(1/\varepsilon)} + \sigma_\theta \sqrt{2\delta / \lambda^+_{\min}(F_e^{\text{causal}})}$.

Derivation sketch. Let $\Delta\theta = \sum_{\ell \in \mathcal{S}} U_\ell (B_\ell A_\ell) \in \mathbb{R}^p$ denote the LoRA update, with U_ℓ the layer embedding, $A_\ell \in \mathbb{R}^{r_\ell \times d_\ell^{\text{in}}}$, and $B_\ell \in \mathbb{R}^{d_\ell^{\text{out}} \times r_\ell}$.

Step 1 (reliability ⇒ score-function displacement). The ε-reliability constraint $p_{\theta_0+\Delta\theta}(y_{\text{new}} \mid x) \geq 1 - \varepsilon$ suggests the heuristic estimate below, via a Cramér-Rao-style approximation applied to the Bernoulli distribution that places mass $1 - \varepsilon$ on y_{new} versus the pre-edit distribution that places mass at most $1/2$ on it (any successfully edited fact is one the model originally got wrong). That comparison fixes only that the displacement is bounded away from zero: against a pre-edit mass of $1/2$ its divergence stays finite as $\varepsilon \to 0$, so the logarithmic form below is posited as the working scale,

$$\begin{aligned} &\mathrm{KL}\big(p_{\theta_0+\Delta\theta}(\cdot \mid x) \,\|\, p_{\theta_0}(\cdot \mid x)\big) \\ &\qquad \gtrsim \tfrac{1}{2}\log(1/\varepsilon) - O(\varepsilon). \end{aligned} \tag{A2}$$

Linearising the log-likelihood ratio in $\Delta\theta$ around θ_0,

$$\begin{aligned} \mathrm{KL}\big(p_{\theta_0+\Delta\theta} \,\|\, p_{\theta_0}\big) \;&=\; \tfrac{1}{2}\,\Delta\theta^\top F_e\, \Delta\theta \\ &\quad +\; O\big(\sigma_\theta^2 \|\Delta\theta\|^2\big), \end{aligned} \tag{A3}$$

where the linearisation error is bounded by σ_θ^2 on $\mathcal{V}_e$ by hypothesis. Combining the two displays, the second of which is the Fisher expansion Eq. (A3), and projecting onto $\mathcal{V}_e$ (the relevant directions for the reliability constraint, since components outside $\mathcal{V}_e$ score below the attribution cutoff of Eq. (4) and so move $f_\theta(x)$ only negligibly to leading order), we obtain

$$\|\Pi_{\mathcal{V}_e}\Delta\theta\|^2_{F_e^{\text{causal}}} \;\gtrsim\; \log(1/\varepsilon) - O(\varepsilon), \tag{A4}$$

where $\|v\|_M^2 = v^\top M v$ denotes the Mahalanobis norm in the metric M.

Step 2 (locality ⇒ Frobenius bound on $\Delta\theta$). The δ-locality constraint $\mathbb{E}_{x\sim\mathcal{L}_e}\big[\mathrm{KL}\big(p_{\theta_0+\Delta\theta}(\cdot \mid x)\|p_{\theta_0}(\cdot \mid x)\big)\big] \le \delta$ similarly expands as

$$\delta \;\ge\; \tfrac{1}{2}\,\Delta\theta^\top F_{\text{loc}}\,\Delta\theta \;+\; O\big(\sigma_\theta^2\|\Delta\theta\|^2\big), \tag{A5}$$

where $F_{\text{loc}} = \mathbb{E}_{x\sim\mathcal{L}_e}[F_e(x)]$ is the locality-averaged Fisher. Assuming $F_{\text{loc}} \succeq \lambda^+_{\min}(F_e^{\text{causal}})\,\Pi_{\mathcal{V}_e}$, we obtain $\|\Pi_{\mathcal{V}_e}\Delta\theta\|^2 \le 2\delta/\lambda^+_{\min}(F_e^{\text{causal}}) + O(\sigma_\theta^2)$.

Step 3 (assemble the energy relation). Combining Step 1 and Step 2 gives the two-sided energy relation of Eq. (A1), on the causal subspace: a reliability-associated energy estimate $\Delta\theta^\top F_e^{\text{causal}}\Delta\theta \gtrsim \log(1/\varepsilon) - O(\varepsilon)$ together with the locality-driven cap $\|\Pi_{\mathcal{V}_e}\Delta\theta\|^2 \le 2\delta/\lambda^+_{\min} + O(\sigma_\theta^2)$ obtained from the Step-2 locality expansion above; the closed-form quantity $\varphi(\varepsilon,\delta) = \frac{1}{\sqrt{2}}\sqrt{\log(1/\varepsilon)} + \sigma_\theta\sqrt{2\delta/\lambda^+_{\min}(F_e^{\text{causal}})}$ measures the slack between the reliability side and the locality cap.

From energy to a rank floor. The energy relation of Heuristic Argument 1 constrains the Fisher-weighted energy of the update, and the rank floor is an observed regularity on top of it. Writing $\Pi_{\mathcal{V}_e}\Delta\theta = \sum_i \alpha_i u_i$ in the eigenbasis $F_e^{\text{causal}} = \sum_i \lambda_i u_i u_i^\top$, the reliability-side estimate $\sum_i \lambda_i\alpha_i^2 \gtrsim \log(1/\varepsilon)$ is met by energy concentrated on a few top eigendirections as well as by energy spread over many, and a rank-one update can place non-zero α_i on every u_i; the rank floor comes from the observation recorded as Empirical Claim 1, with the per-decile rates in Table A4. We therefore report $r \ge r^\star(e) - \varphi(\varepsilon,\delta)$ as an empirical regularity supported by the $r^\star(e)$-versus-failure data. Ranks are counted per layer throughout: the same rank is allocated in each $\ell \in \mathcal{S}$, so the deployed parameter footprint is $|\mathcal{S}|$ blocks of that rank, and the comparison against the fixed-rank baselines is at matched $|\mathcal{S}|$.

Appendix A.5: Empirically observed consequences

Empirical Claim 1 (Empirical rank floor)**.** *Across the evaluation of Section 5 we observe the following regularity: editors whose per-layer rank falls below $r^\star(e)-\varphi$ fail the (ε,δ)-correctness criterion at a rate that rises with the shortfall below the floor, passing half once that shortfall reaches about ten dimensions and staying a minority below it (per-decile rates in Table A4), and $\widehat{r}^\star$ predicts the joint cross-method failure of ROME, MEMIT, and ELDER at Spearman $\rho = 0.74$ (Section 5). We treat $r^\star(e) - \varphi$ as an empirically supported floor.*

Observed consequence 1 (rank-one editors as $r^\star(e)$ grows). The (ε,δ)-failure rate of rank-one editors rises monotonically with $r^\star$ across the deciles of Table A4, passing half by D6; on the easy deciles most edits still succeed at rank one. On COUNTERFACT-HARD ($r^\star \ge 16$) the failure of ROME and MEMIT in Table 2 is the cross-method counterpart of this floor.

Observed consequence 2 (over-rank editors). On edits with $r^\star \le 4$, a fixed-rank $r = 32$ editor carries at least $32 - (4 - \varphi) \approx 29$ rank dimensions in excess of the observed floor. Each such dimension is a degree of freedom the optimiser may steer, and on the locality regulariser this corresponds to a degree of freedom that can leak into the locality set. The locality drop of ELDER versus FISHER-LoRA reported in Table 1 is consistent with this excess: in Table 4 the rank-only row already recovers 6.0 of its 12.1 Locality points, the subspace and the initialisation the remainder.

Appendix A.6: The estimator-bias argument

Heuristic Argument 2 (Estimator bias, empirical estimate)**.** *Under a local linearisation, write $C = 2L_{\text{Hess}}/\lambda^+_{\min}(F_e^{\text{causal}})$ with a model-dependent constant L_{Hess}. The estimator bias is then*

empirically estimated to decompose, approximately, into an off-diagonal term and a linearisation term,

$$\begin{aligned}\left\|F_e^{\text{causal}} - \widehat{F}_e^{\text{AP}}\right\|_F \;\approx\; &\underbrace{\left\|\mathcal{D}^{\perp}[M_e] + (\mathbb{E}[\alpha(y)^2] - 1)\,\widehat{F}_e^{\text{AP}}\right\|_F}_{\mathcal{O}(1)} \\ &+\; \underbrace{C \cdot \sigma_\theta^2 \cdot \left\|F_e^{\text{causal}}\right\|_F}_{\mathcal{O}(\sigma_\theta^2)},\end{aligned} \tag{A6}$$

where $M_e = \mathbb{E}[\alpha(y)^2]\,\Pi_{\mathcal{V}_e} g_e g_e^\top \Pi_{\mathcal{V}_e}$ is the rank-one leading term. The matrix M_e is unrelated to the base model M_θ of Section 3, and $\mathcal{D}^{\perp}[\cdot]$ keeps its off-diagonal blocks in the component basis $\{u_c\}$, where $\widehat{F}_e^{\text{AP}} = \sum_{c:\,u_c \in \mathcal{V}_e} g_e^{(c)} (g_e^{(c)})^\top$ is the sum of per-component contrastive outer products of Eq. (8) *and $g_e = \nabla_\theta \log[p_{\theta_0}(y_{\text{new}} \mid x)/p_{\theta_0}(y_{\text{old}} \mid x)]$.*

Derivation sketch. Decompose the log-likelihood score $s_\theta(y \mid x) = \nabla_\theta \log p_\theta(y \mid x)$ along the contrastive direction g_e and its orthogonal complement,

$$s_{\theta_0}(y \mid x) \;=\; g_e \cdot \alpha(y) \;+\; \varrho(y), \tag{A7}$$

where $\alpha(y)$ scales onto the contrastive log-ratio direction and $\varrho(y)$ is the residual orthogonal to g_e. Substituting into $F_e = \mathbb{E}_y[s_{\theta_0} s_{\theta_0}^\top]$ and projecting,

$$\begin{aligned}F_e^{\text{causal}} \;=\; \Pi_{\mathcal{V}_e} \mathbb{E}_y\big[&(g_e\alpha(y) + \varrho(y)) \\ &\times (g_e\alpha(y) + \varrho(y))^\top\big]\Pi_{\mathcal{V}_e}.\end{aligned} \tag{A8}$$

Expanding the outer product in Eq. (A8) yields three groups of terms: the leading term $M_e = \mathbb{E}[\alpha(y)^2]\,\Pi_{\mathcal{V}_e} g_e g_e^\top \Pi_{\mathcal{V}_e}$, which is the full rank-one outer product of the contrastive direction on $\mathcal{V}_e$; two cross terms whose Frobenius norm is bounded by $L_{\text{Hess}}\sigma_\theta^2$ each; and the residual outer product whose Frobenius norm is bounded by $L_{\text{Hess}}^2\sigma_\theta^4$. The cross and residual terms are the linearisation error and vanish with σ_θ^2; collected against $\|F_e^{\text{causal}}\|_F \geq \lambda_{\min}^+(F_e^{\text{causal}})$ they give the second term of Eq. (A6) with $C = 2L_{\text{Hess}}/\lambda_{\min}^+(F_e^{\text{causal}})$.

The leading term does *not* reduce to the estimator. Writing M_e in the component basis $\{u_c\}$, the diagonal blocks of M_e are $\mathbb{E}[\alpha(y)^2]$ times the estimator of Eq. (8) of the main body, $\sum_{c:\,u_c \in \mathcal{V}_e} g_e^{(c)} (g_e^{(c)})^\top = \widehat{F}_e^{\text{AP}}$, while the off-diagonal blocks $\mathcal{D}^{\perp}[M_e]$, which carry the cross-component couplings $(g_e^\top u_c)(g_e^\top u_{c'})\,u_c u_{c'}^\top$ for $c \neq c'$, are discarded. These blocks carry no factor of σ_θ: they survive even in the exact-linearisation limit $\sigma_\theta \to 0$, and their size is $\mathcal{O}(1)$ relative to $\|M_e\|_F$, set by how much of the contrastive energy lies off the component axes, not by any expansion order. This is the price of the diagonal modelling choice of Section 4.4, and it is the reason the two terms of Eq. (A6) must be reported separately: only the second is controlled by σ_θ.

Appendix A.7: Empirical calibration of the bias constant C

The constant C governs only the linearisation term of Eq. (A6); it is an empirically measured bias constant, calibrated on a held-out calibration set of edits by comparing the attribution-patching estimator $\widehat{F}_e^{\text{AP}}$ against a more expensive reference Fisher, after removing the off-diagonal contribution. On LLaMA-2-7B the measured linearisation coefficient is approximately $C \approx 0.24$, that is, the *linearisation term* of Eq. (A6) is $\approx 0.24\,\sigma_\theta^2\,\|F_e^{\text{causal}}\|_F$; this coefficient has a fifth-to-ninety-fifth percentile band of 0.18 to 0.31 across edits, and the off-diagonal term is not covered by it. A heuristic surrogate for C is $2L_{\text{Hess}}/\lambda_{\min}^+(F_e^{\text{causal}})$, with H the log-likelihood Hessian at θ_0 and $L_{\text{Hess}} = \|\Pi_{\mathcal{V}_e} H g_e\|/\|g_e\|^2$, evaluated in practice on the spectrum of $\widehat{F}_e^{\text{AP}}$ as the computable proxy for $\lambda_{\min}^+(F_e^{\text{causal}})$; it is consistent with this measured value. The corresponding measured constants on Mistral-7B and GPT-2-XL are $C \approx 0.27$ and $C \approx 0.31$ respectively.

Appendix A.8: Scale and Notation Conventions

Rank counting. Every rank quoted in this paper, including the fixed $r = 1$, $r = 8$, and $r = 32$ of the baselines, is a per-layer rank under the layer set of its backbone, so the comparison is at matched

$|\mathcal{S}|$. Ranks are therefore never aggregated across layers: a per-layer rank r deployed over $|\mathcal{S}|$ layers is reported as rank r, not as $r\,|\mathcal{S}|$, for FISHER-LORA and for every baseline alike.

The linearisation-error constant σ_θ. The slack $\varphi(\varepsilon, \delta)$ of Section 4.3 contains the constant σ_θ, defined as follows: writing the first-order log-likelihood remainder as $r_\theta(\Delta\theta) = \log p_{\theta+\Delta\theta}(y \mid x) - \log p_\theta(y \mid x) - \langle \nabla_\theta \log p_\theta(y \mid x),\, \Delta\theta \rangle$, it is the smallest scalar for which

$$|r_\theta(\Delta\theta)| \;\le\; \sigma_\theta^2 \, \|\Delta\theta\|_F^2 \tag{A9}$$

holds for every update $\Delta\theta$ in the Frobenius ball on $\mathcal{V}_e$; that is, σ_θ^2 upper-bounds the second-order (linearisation) remainder of the log-likelihood on the causal subspace. The constant is a property of the frozen model and of the causal subspace $\mathcal{V}_e$, not of the editor, which is why the same σ_θ enters both the slack φ and the estimator-bias statement of Appendix A.6.

Evaluating the slack. The two terms of Eq. (7) come from the two steps of Appendix A.4. Step 1 gives $\mathrm{KL} \gtrsim \frac{1}{2}\log(1/\varepsilon)$, whose square root is the first term. Step 2 caps the displacement at $\|\Pi_{\mathcal{V}_e}\Delta\theta\|^2 \le 2\delta/\lambda^+_{\min}$, so the norm it admits has scale $\sqrt{2\delta/\lambda^+_{\min}}$, and σ_θ multiplies that scale to the first power because Eq. (A9) bounds the second-order remainder by σ_θ^2 inside the squared norm. At $\varepsilon = 0.05$ and $\delta = 0.1$ the reliability term of Eq. (7) is $\frac{1}{\sqrt{2}}\sqrt{\log(1/0.05)} = \frac{1}{\sqrt{2}}\sqrt{2.996} = 1.224$ dimensions, and the locality factor is $\sqrt{2\delta/\lambda^+_{\min}} = \sqrt{0.2} = 0.447$ at $\lambda^+_{\min} = 1$, so

$$\varphi(\varepsilon, \delta) \;=\; 1.224 \;+\; 0.447\,\sigma_\theta. \tag{A10}$$

On LLaMA-2-7B, $\sigma_\theta = 0.1$ gives $\varphi = 1.224 + 0.045 = 1.269$, the 1.27 quoted in Section 5. The first term is backbone-independent and carries 96% of the LLaMA-2-7B slack; $\lambda^+_{\min}$ and σ_θ are taken at common values on the three backbones, so φ takes the same value on each.

APPENDIX A.9: DATASETS

FISHER-LORA and the baselines are compared on COUNTERFACT (Meng et al., 2022; 2023), the standard knowledge-editing benchmark. COUNTERFACT contains paraphrase prompts and neighbourhood prompts for each fact, supporting the evaluation of reliability and locality. Table A2 summarises the splits and provenance.

Roles of the splits. Three split names recur in this appendix, and they name three roles, not three disjoint sets of edits. The *held-out test split* is the evaluation pool: the 100 random COUNTERFACT edits, the 1,500-edit COUNTERFACT-HARD slice, and the 100 random zsRE edits. The *development split* is a separate 3,500-edit draw from COUNTERFACT, held disjoint from the 100-edit evaluation set and from the 5,000-edit difficulty pool by the same excluded-identifier list, so it contains no edit of the COUNTERFACT-HARD slice and no edit whose $r^\star$ enters any decile; the learning-rate grid, the locality-regulariser weight β, the attribution cutoff τ_{AP} and the optimiser-step budget T are selected there. The *held-out calibration set* is that same 3,500-edit development split in its measurement role, not its selection role: the estimator-bias constant C is measured on it, held out from the headline evaluation sets in the sense that no edit it contains enters Table 1 or Table A6. The COUNTERFACT-HARD slice is the D8 to D10 subset of the same pool, reported as a stratified evaluation; nothing is selected on it or on the pool it is filtered from.

COUNTERFACT. The benchmark consists of 21,919 counterfactual associations. Each is paired with the original fact, the counterfactual target, a set of paraphrase prompts that ought to elicit the same target after editing, a set of neighbourhood prompts that ought to remain unchanged, and a generation-style prompt set used for downstream fluency evaluation. We sample 100 random edits for the main-body Tables 1 and 4 (seed 42), the same 100 across LLaMA-2-7B, Mistral-7B, and GPT-2-XL so that per-model differences are not confounded with edit-set composition. The 5,000-edit pool used for the difficulty-correlation analysis in Section 5.3 is sampled with a different seed (43) and is disjoint from the 100-edit evaluation set, the disjointness enforced by an excluded-identifier list and the draw taken one edit record at a time. Its role in the difficulty analysis is to estimate the Spearman $\rho = 0.74$ between the per-edit complexity $\widehat{r}^\star$ and the joint failure rate of

Table A2: Dataset splits used in the main-body experiments. The COUNTERFACT counts refer to the random subset of facts drawn from the full benchmark; the COUNTERFACT-HARD slice is the union of the top three difficulty deciles of the correlation pool, i.e. the edits with ground-truth $r^\star(e) \geq 16$. Every COUNTERFACT and zsRE split is evaluated on the paraphrase and neighbourhood prompts shipped with its benchmark; WikiText-2 enters as a language-modelling perplexity probe.

Benchmark	Split	Edits	Prompts per edit	License	Source
COUNTERFACT	random	100	~ 30	MIT	Meng et al. (2022)
COUNTERFACT	correlation pool	5,000	~ 30	MIT	Meng et al. (2022)
COUNTERFACT	development	3,500	~ 30	MIT	Meng et al. (2022)
COUNTERFACT-HARD	filtered	1,500	~ 30	MIT	this paper
zsRE	random	100	~ 10	MIT	Levy et al. (2017)
WikiText-2	locality probe	$\sim$ 36k tokens	n/a	CC BY-SA	Merity et al. (2017)

Table A3: Ground-truth $r^\star$ decile boundaries and counts on the 5,000-edit COUNTERFACT pool (LLaMA-2-7B). The median column is the one reported in Table A7. COUNTERFACT-HARD is the union D8 to D10, i.e. the 1,500 edits with $r^\star \geq 16$.

Decile	$r^\star$ range	Median $r^\star$	Edits
D1	1 to 2	1	500
D2	3 to 4	3	500
D3	5 to 6	5	500
D4	7 to 8	7	500
D5	9 to 10	9	500
D6	11 to 13	12	500
D7	14 to 15	15	500
D8	16 to 21	19	500
D9	22 to 27	24	500
D10	28 to 65	30	500
COUNTERFACT-HARD (D8 to D10)	≥ 16	24	1,500

ROME, MEMIT, and ELDER. The COUNTERFACT-HARD slice is filtered from the 5,000-edit pool by retaining edits whose ground-truth $r^\star(e) \geq 16$. That condition selects exactly the union of the top three difficulty deciles D8 to D10 of Table A3, so the slice contains 1,500 edits. Ground-truth $r^\star$ is measured by search over integer ranks $r \in \{1, \ldots, 64\}$, an edit that no rank in that range satisfies being recorded with the sentinel value 65, for the smallest rank achieving $\geq 95\%$ reliability and $\geq 90\%$ locality across the strongest of the LoRA-based baselines (Yu et al., 2024; Li et al., 2025).

zsRE. zsRE is the zero-shot relation-extraction dataset of Levy et al. (2017), adapted into an editing benchmark by De Cao et al. (2021) and Mitchell et al. (2022a), whose edits are single-relation rewrites; we sample 100 random edits (seed 42) for the cross-dataset generalisation check of Table A6, using the paraphrase and neighbourhood prompts shipped with the benchmark for reliability and locality. The average rank allocated on zsRE (9.7 on LLaMA-2-7B) is lower than on COUNTERFACT (11.4).

WikiText-2. A subset of WikiText-2 (Merity et al., 2017) is held out as a language-modelling perplexity probe to verify that no editor catastrophically degrades the unedited model on out-of-edit text. The probe is computed after each edit on a fixed 36,000-token slice of the `wikitext-2-raw-v1` test split and averaged over the edits; we report the perplexity delta versus the unedited base model in Appendix A.17.

Ethical use. The datasets we use permit research use and contain no personal data. The COUNTERFACT counterfactual targets are factually false by construction (the benchmark's purpose is to install counterfactuals), so we do not deploy any edited model.

APPENDIX A.10: BACKBONES AND EDIT SITES

Every editor compared in this paper writes into the same block of a given backbone, so that no reported difference can be attributed to a different edit site. The block is the feed-forward stack identified by causal tracing (Meng et al., 2022): layers 4 to 8 on LLaMA-2-7B (Touvron et al., 2023) and on Mistral-7B (Jiang et al., 2023), and layers 20 to 25 on GPT-2-XL (Radford et al., 2019). We denote that set of layers by $\mathcal{S}$ and keep it fixed for every method, every dataset and every run reported here; the alternative layer sets we tried, and what they cost, are in Appendix A.20.

What one site is. FISHER-LORA attaches one LoRA pair (B_ℓ, A_ℓ) to the down-projection of each layer $\ell \in \mathcal{S}$, so an edit of per-layer rank r occupies r directions in each of the $|\mathcal{S}|$ layers. Every rank quoted anywhere in this paper is a per-layer rank at matched $|\mathcal{S}|$, never a sum across layers; the conventions that fix this accounting are given in Appendix A.8.

Backbone scale. The three backbones span a factor of about five in parameter count (1.5B for GPT-2-XL against 7B for the other two) and two tokenizer families, which is what lets the comparison of Table 1 be stated across backbones rather than on one model.

APPENDIX A.11: TRAINING SCHEDULE AND OPTIMISATION

We record the full per-edit optimisation schedule for FISHER-LORA and the LoRA-based baselines. The schedule was held fixed across all three model families to remove model-specific tuning as a confound in Tables 1 and 2, and was used unchanged on COUNTERFACT-HARD so the comparison there reflects rank-allocation differences, not optimisation-budget differences.

Phase 1 (Causal-Fisher pre-edit). Phase 1 produces the per-edit rank $\widehat{r}^\star$ before any optimiser step is taken. The contrastive gradient $g_e = \nabla_\theta \log[p_{\theta_0}(y_{\text{new}} \mid x)/p_{\theta_0}(y_{\text{old}} \mid x)]$ is formed from one backward pass through the unedited model. The causal subspace $\mathcal{V}_e$ is constructed via Eq. (4) with relative cutoff $\tau_{\text{AP}} = 0.01$ (relative to $\max_{c'} a_{c'}(e)$ as in Eq. (4)); on LLaMA-2-7B this yields $|\mathcal{V}_e|$ in the range $[20, 90]$ across the 100 evaluation edits, with median 43. The estimator $\widehat{F}_e^{\text{AP}}$ is formed via Eq. (8); because it is diagonal in the component basis $\{u_c\}$, its spectrum is read off directly as the squared component scores and no matrix factorisation is performed. The estimate $\widehat{r}^\star$ is then the count of squared scores above the relative cutoff $\tau = 10^{-2}$ (Eq. (6)), obtained by sorting the at most ninety retained scalars. The total Phase 1 wall-clock is approximately 0.7 s per edit, essentially all of it the backward pass; the sort is not measurable against it.

Phase 2 (Fisher-eigenvector LoRA initialisation). Once $\widehat{r}^\star$ and the top-$\widehat{r}^\star$ eigenvectors $\{v_i\}_{i=1}^{\widehat{r}^\star}$ are in hand, we initialise the LoRA factors per layer $\ell \in \mathcal{S}$ as $A_\ell \leftarrow [v_1, \ldots, v_r]^\top|_\ell$ with $r = \min(\widehat{r}^\star, r^{\text{budget}})$ (the restriction of the top eigenvectors to layer ℓ's input coordinates) and $B_\ell \leftarrow \mathbf{0}$. The zero initialisation of B guarantees that the initial output of $\theta_0 + \Delta\theta$ equals the unedited model, so the first optimiser step has access to the correct loss surface. The A initialisation aligns the trainable subspace with the estimated rank $\widehat{r}^\star$.

Phase 3 (per-edit optimisation). We optimise the LoRA factors $\{(A_\ell, B_\ell)\}_{\ell \in \mathcal{S}}$ by Adam on the standard editing loss of Eq. (10), with learning rate 1×10^{-3}, batch size 1, and $T = 25$ optimiser steps. The learning rate is set on the hyperparameter grid of Appendix A.14 (smallest learning rate that meets the $\geq 95\%$ reliability and $\geq 90\%$ locality joint criterion on the development split). The locality regulariser weight is $\beta = 0.1$, selected on the same grid; smaller values inflate locality damage on neighbouring facts, larger values delay reliability convergence beyond the 25-step budget. We use a constant learning rate, no warm-up or decay, and no gradient clipping.

Layer set $\mathcal{S}$. The layer set $\mathcal{S}$ over which the LoRA pair is applied is fixed to the contiguous block of transformer MLP modules identified by the causal-tracing analysis of Meng et al. (2022): layers $\{4, \ldots, 8\}$ on LLaMA-2-7B (32 layers total), layers $\{4, \ldots, 8\}$ on Mistral-7B (32 layers total), and layers $\{20, \ldots, 25\}$ on GPT-2-XL (48 layers total). This range was held fixed across all editors in Tables 1, 2, and 4 so that the rank-selection comparison is not confounded with layer-selection differences; ablating the layer set is reported in Appendix A.20. Within each such layer the edit

targets the MLP down-projection, `mlp.down_proj` on the two LLaMA-family backbones and `mlp.c_proj` under the GPT-2 naming, and the component basis $\{u_c\}$ is the neuron basis of that module, one u_c per output neuron, supported on that neuron's row of input weights. Because $|\mathcal{S}|$ differs across backbones, the matched-$|\mathcal{S}|$ comparison is within each model, not across models: on a given model the same range is applied to every editor, so no method is advantaged by layer selection.

Numerical precision. The Adam moment buffers, the LoRA factors, and the component-score arithmetic are held in single precision to preserve the curvature signal that the empirical rank floor depends on. Reliability is stable to within 0.4 percentage points when the component scores are accumulated in double precision, indicating that single precision suffices for the Phase 1 spectral analysis. The contrastive gradient g_e is accumulated in single precision from the layer-wise gradients to avoid round-off in the eigenvector projection.

Evaluation setup. Post-edit evaluation truncates at a maximum sequence length of 512 tokens and batches eight prompts at a time; the optimiser itself still sees one edit per step.

Convergence diagnostics. The post-edit reliability and locality are tracked after every five optimiser steps; the large majority of converging edits reach within 1 percentage point of their final reliability by step 20 and within 0.3 percentage points by step 25. The 25-step budget therefore clears the convergence horizon; the T sweep is in Appendix A.20.

Software versions. The complete software stack is PyTorch 2.2.0, Transformers 4.40.1, PEFT 0.10.0, Datasets 2.18.0, NumPy 1.26.0, and scikit-learn 1.4.0. The causal-tracing primitives reuse the public reference implementation released with Meng et al. (2022); the attribution-patching primitive reuses the contrastive variant of Syed et al. (2024); the AdaLoRA and GaLore baselines use the official reference implementations of Zhang et al. (2023b) and Zhao et al. (2024). Phase 1 uses no linear-algebra routine beyond the sort of the component scores.

APPENDIX A.12: THE EDITING PROCEDURE

Algorithm 1 states the procedure the three components compose. Lines 1 to 6 are Phase 1: one backward pass gives the contrastive gradient, the attribution scores select the causal subspace, and the estimator's spectrum is read off those same scores and thresholded into the per-edit rank $\widehat{r}^{\star}$, so $\widehat{F}_e^{\mathrm{AP}}$ is never assembled and no eigendecomposition is run. Lines 7 to 14 are Phase 2. The feasibility screen evaluates the slack under the scale convention of Appendix A.8, and is the only branch that ends the procedure without an edit. When it passes, the applied rank is $r = \min(\widehat{r}^{\star}, r^{\mathrm{budget}})$ and each A_ℓ starts at the leading r eigenvectors restricted to layer ℓ with B_ℓ at zero, so the initial output equals the unedited model. Lines 15 to 18 are Phase 3, which fits the factors for T Adam steps on the editing loss. The screen and the initialisation add no pass of their own: both read the spectrum of line 4.

APPENDIX A.13: HARDWARE AND SOFTWARE

Data. We use COUNTERFACT (Meng et al., 2022) and zsRE (Levy et al., 2017; De Cao et al., 2021) with the paraphrase and neighbourhood prompts shipped with each benchmark, and WikiText-2 (Merity et al., 2017) as the language-modelling locality probe.

All experiments run on a single NVIDIA A100 80GB GPU with CUDA 12.1. Per-edit wall-clock on this GPU ranges from 9.4 s (MELO) to 22.5 s (AdaLoRA). The implementation uses PyTorch 2.2.0, the Hugging Face Transformers 4.40 library, and PEFT 0.10. Each edit is processed on its own model instance holding LLaMA-2-7B, Mistral-7B, or GPT-2-XL alongside the per-edit LoRA factors and the locality-regulariser activations, and the $5{,}000$-edit difficulty-correlation study runs as independent per-edit jobs that share no state. Random number generation for the LoRA initialisation and the edit shuffle is drawn from a single fixed seed and reset identically across the three model families so that the edit ordering is reproducible.

Algorithm 1 FISHER-LORA: reading a per-edit LoRA rank from one backward pass, then editing at the resulting rank.

Require: Frozen M_{θ_0}, edit $e = (x, y_{\text{old}}, y_{\text{new}})$, locality set $\mathcal{L}_e$, layer set $\mathcal{S}$ with orthonormal component basis $\{u_c\}$, cutoffs τ_{AP}, τ, tolerances (ε, δ), constant σ_θ, rank budget r^{budget}, loss weight β, steps T.

Ensure: LoRA factors $\{(A_\ell, B_\ell)\}_{\ell \in \mathcal{S}}$, or an empirical infeasibility flag.

Phase 1: read the per-edit rank

$g_e \leftarrow \nabla_\theta \log[\, p_\theta(y_{\text{new}} \mid x) / p_\theta(y_{\text{old}} \mid x)\,]$ at θ_0 {the rank-reading pass}

$a_c(e) \leftarrow |g_e^\top u_c|$ for every component c, by Eq. (3).

$\mathcal{V}_e \leftarrow \text{span}\{u_c : a_c(e) \geq \tau_{\text{AP}} \max_{c'} a_{c'}(e)\}$, by Eq. (4).

$(\lambda_c, v_c) \leftarrow (a_c(e)^2,\, u_c)$ for each retained component c, by Eq. (9) {$\widehat{F}_e^{\text{AP}}$ is never formed}

Sort and relabel the pairs (λ_c, v_c) so that $\lambda_1 \geq \cdots \geq \lambda_{|\mathcal{V}_e|}$.

$\widehat{r}^\star \leftarrow |\{\, i : \lambda_i \geq \tau\, \lambda_1 \,\}|$, by Eq. (6).

Phase 2: screen the budget, then initialise at the eigenvectors

$\varphi(\varepsilon, \delta) \leftarrow \frac{1}{\sqrt{2}} \sqrt{\log(1/\varepsilon)} + \sigma_\theta \sqrt{2\delta / \lambda^+_{\min}}$, by Eq. (7) at the scale convention of Appendix A.8.

if $r^{\text{budget}} < \widehat{r}^\star - \varphi(\varepsilon, \delta)$ **then**

  return INFEASIBLE

end if

$r \leftarrow \min(\widehat{r}^\star,\, r^{\text{budget}})$.

for each $\ell \in \mathcal{S}$ **do**

  $A_\ell \leftarrow [\, v_1, \ldots, v_r \,]^\top|_\ell$ and $B_\ell \leftarrow \mathbf{0}$ {initial output equals M_{θ_0}}

end for

Phase 3: fit the factors

for $t = 1, \ldots, T$ **do**

  Update $\{(A_\ell, B_\ell)\}_{\ell \in \mathcal{S}}$ by one Adam step on $\mathcal{L}_{\text{edit}}(e)$ of Eq. (10).

end for

return $\{(A_\ell, B_\ell)\}_{\ell \in \mathcal{S}}$.

APPENDIX A.14: TRAINING HYPERPARAMETERS

Hyperparameters are set on the hyperparameter grid described in this appendix. The selected configuration is shared across LLaMA-2-7B, Mistral-7B, and GPT-2-XL; per-model overrides were screened on the development split but not required. Random seed 42 is fixed across every training run. The per-(method, benchmark) configuration is listed below.

Per-(method, benchmark) configuration. Every pair evaluated in the main body is configured as follows. Batch size is one edit throughout, and the locality regulariser weight of Eq. (10) is $\beta = 0.1$ for every LoRA-based method, selected on the same grid as the learning rates. Adam runs with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and zero weight decay. The LoRA scaling α equals r and no dropout is applied. The per-layer rank is clipped to the budget of 40 from above and to 1 from below.

- FISHER-LORA, on COUNTERFACT and on COUNTERFACT-HARD: Adam, 25 steps, learning rate searched over $\{3, 5\}\times10^{-4}$ and $\{1, 3\}\times10^{-3}$, selected 1×10^{-3}.
- MELO (Yu et al., 2024), on COUNTERFACT: Adam, 25 steps, the same search grid, selected 1×10^{-3}.
- ELDER, on COUNTERFACT: Adam, 25 steps, searched over $\{1, 3, 5\}\times10^{-4}$ and $\{1, 3\}\times10^{-3}$, selected 5×10^{-4}.
- AdaLoRA (Zhang et al., 2023b) and GaLore (Zhao et al., 2024), on COUNTERFACT: AdamW, 25 steps, searched over $\{3, 5\}\times10^{-4}$ and $\{1, 3\}\times10^{-3}$, selected 1×10^{-3} for both.
- ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), on COUNTERFACT: Adam, 20 steps, searched over $\{1, 5\}\times10^{-1}$ and 1×10^{0}, selected 5×10^{-1} for both. These two optimise an internal target activation, not LoRA factors, so their effective learning rate sits on a different scale from the rank-structured methods.

- On zsRE, FISHER-LORA, ROME, MEMIT, MELO and ELDER reuse the configuration selected on COUNTERFACT for the same method. zsRE is a cross-dataset generalisation check, so no separate search was run there.
- GRACE (Hartvigsen et al., 2023) and WISE (Wang et al., 2024b) run on COUNTERFACT at their released defaults. They enter only as the out-of-family check of Appendix A.15, so no quantity in their setting was selected here.

APPENDIX A.15: BASELINE REPRODUCTION PROTOCOL

We compare against six comparison baselines spanning three tiers, and additionally run GRACE and WISE as out-of-family probes, whose protocol is given at the end of this subsection. Tier 1 (LoRA-based editors) consists of MELO (Yu et al., 2024) at fixed rank $r = 8$ and ELDER at fixed rank $r = 32$; both are drawn from their official public code repositories, with only the data loader adapted to consume the COUNTERFACT splits and locality evaluation prompts used here. Tier 2 (edit-specific rank-1 methods) consists of ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), both taken from the official MEMIT repository. Tier 3 (adaptive low-rank methods proposed for fine-tuning, not editing) consists of AdaLoRA (Zhang et al., 2023b), which schedules the rank of a LoRA product, and GaLore (Zhao et al., 2024), which projects gradients onto a low-rank subspace while training the full parameter set. Their official reference implementations are adapted to operate on the per-edit loss in Eq. (10) with ranks matched to the average rank used by FISHER-LORA.

Out-of-family probes. Beyond the six comparison baselines above, the memory-based editors GRACE (Hartvigsen et al., 2023) and WISE (Wang et al., 2024b) enter only as an out-of-family check on the difficulty predictor of Section 5.3: the joint-failure set is extended from the three rank-based editors to include them, and the resulting correlation is reported there. They are run from their official public repositories at the released default configuration, with no per-edit or per-benchmark search. For that reason no searched value is listed for them above, and they contribute no entry to Tables 1 and 2. They are memory editors, not weight editors, so a matched-rank comparison is undefined for them.

APPENDIX A.16: METRIC DEFINITIONS AND AGGREGATION

We define the five metrics reported in the main body. All metrics aggregate across the 100 random COUNTERFACT edits unless noted otherwise.

Reliability. Reliability is built from a per-(edit, paraphrase) *paraphrase accuracy* taking continuous values in $[0, 1]$: for edit $e = (x, y_{\text{old}}, y_{\text{new}})$ and paraphrase prompt x' of e, it is the fraction of the target tokens of y_{new} that the edited model ranks first at each position given x' and the target prefix, so a partially installed target scores strictly between 0 and 1. Writing $P(e)$ for the paraphrase prompts of e and following the COUNTERFACT protocol of Meng et al. (2022),

$$\text{Rel}(e) \;=\; \frac{1}{|P(e)|} \sum_{x' \in P(e)} \frac{1}{|y_{\text{new}}|} \sum_{t=1}^{|y_{\text{new}}|} \mathbf{1}\left[\arg\max_{y} p_{\theta+\Delta\theta}\big(y \mid x', y_{\text{new}}^{<t}\big) = y_{\text{new}}^{t} \right]. \tag{A11}$$

The number reported in Table 1 is the further average across the 100 edits. The "failure rate" column of Table 2 is a distinct hard-thresholded statistic over the COUNTERFACT-HARD slice: the fraction of edits whose Reliability falls below 0.5. This teacher-forced token-level accuracy is the measurable operational surrogate for the probability-mass constraint $\Pr \geq 1 - \varepsilon$ of Eq. (2): the two thresholds share the symbol ε, with the $\geq 95\%$ accuracy form standing in for the probability-mass form throughout the evaluation.

Locality. Locality on edit e is the indicator $\mathbf{1}[\arg\max_y p_{\theta+\Delta\theta}(y \mid x') = \arg\max_y p_{\theta_0}(y \mid x')]$ averaged across the neighbourhood prompts of e, then averaged across the 100 edits. With $N(e)$ for those neighbourhood prompts, and in the form used by the editing benchmarks (Meng et al., 2022; Zhang et al., 2024),

$$\text{Loc}(e) \;=\; \frac{1}{|N(e)|} \sum_{x' \in N(e)} \mathbf{1}\left[\arg\max_{y} p_{\theta+\Delta\theta}(y \mid x') = \arg\max_{y} p_{\theta_0}(y \mid x') \right]. \tag{A12}$$

These neighbourhood probes are what we call the unrelated prompts throughout; the paraphrase split is used for Reliability alone and never enters Locality.

Rank-floor ratio. Rank-floor ratio is the per-edit ratio $r_{\text{used}}/r^\star$, where r_{used} is the rank FISHER-LORA allocates to that edit (read from Phase 1) and $r^\star$ is the ground-truth complexity from the cross-method rank search over $r \in \{1, \dots, 64\}$ of Appendix A.20. The search runs over the fixed-rank LoRA editors (Yu et al., 2024; Li et al., 2025), which gives

$$\begin{aligned} \mathrm{RFR}(e) \;&=\; \frac{r_{\text{used}}(e)}{r^\star(e)}, \\ r^\star(e) \;&=\; \min\{\, r \in \{1, \dots, 64\} \;:\; \mathrm{Rel}_r(e) \geq 1-\varepsilon \;\text{ and }\; \mathrm{Loc}_r(e) \geq 1-\delta \,\}. \end{aligned} \tag{A13}$$

A ratio of 1.0 is optimal; above it the editor over-allocates, below it under-allocates. The reported summary statistics (mean in Section 5.3, percentiles in Appendix A.19) are computed across the 100 random edits.

Difficulty Correlation. Difficulty Correlation is the Spearman rank correlation ρ between two per-edit quantities: $\widehat{r}^\star$, computed from the Phase 1 single-pass estimator on the unedited model, and the joint failure indicator for ROME, MEMIT, and ELDER. That indicator uses the operational surrogate of Eq. (2) defined above, teacher-forced token accuracy and argmax survival, and is 1 exactly when all three miss the (ε, δ)-correctness criterion of Eq. (2) on that edit (per-edit reliability $< 95\%$ or per-edit locality $< 90\%$) and 0 otherwise. Over the editors of Meng et al. (2022), Meng et al. (2023) and Li et al. (2025),

$$\begin{aligned} F(e) \;&=\; \prod_{m \in \{\text{ROME, MEMIT, ELDER}\}} \mathbf{1}[\mathrm{Rel}^m(e) < 1-\varepsilon \;\text{ or }\; \mathrm{Loc}^m(e) < 1-\delta], \\ \rho \;&=\; \rho_{\mathrm{S}}(\widehat{r}^\star, F), \end{aligned} \tag{A14}$$

with ρ_{S} the Spearman rank correlation; the per-decile rates of the three individual events are in Table A4. The reported $\rho = 0.74$ is computed across the $5{,}000$-edit difficulty-correlation pool described in Appendix A.9; the corresponding Pearson coefficient is 0.70 and the decile-stratified joint failure rate climbs monotonically from 4.0% to 71.6% between deciles 1 and 10.

Edit Cost. Edit Cost is the per-edit wall-clock time in seconds on a single NVIDIA A100 80GB GPU. For FISHER-LORA this includes the Phase 1 backward pass and score sort but not the one-time setup overhead. We split it into the single backward pass that yields the rank and the optimisation that follows, measured on the same wall-clock basis as the lifelong-editing baselines (Hartvigsen et al., 2023):

$$\mathrm{Cost}(e) \;=\; t_{\text{read}}(e) \;+\; t_{\text{opt}}(e). \tag{A15}$$

The fixed optimisation budget ($T = 25$ steps) dominates t_{opt}, so per-edit time varies little. For FISHER-LORA on LLaMA-2-7B, $t_{\text{read}} \approx 0.7$ s (Phase 1, Appendix A.11) and $t_{\text{opt}} \approx 9.6$ s of the 10.3 s in Table 1, so reading the rank takes about 6.8% of the edit. Edit Cost is reported in Table 1 only. The Edit Cost entries of Table 1 are on a single wall-clock scale and are directly comparable across methods; rank is one of several factors entering the per-step cost, and the column is not monotone in rank across the table.

APPENDIX A.17: PER-DECILE RELIABILITY DECOMPOSITION

Table 3 of the main text stratifies edits by ground-truth complexity decile to resolve Reliability into per-difficulty cells. The deciles are formed from the $5{,}000$-edit difficulty pool (Section 5.3), sorted by ground-truth $r^\star(e)$ (search over integer ranks $r \in \{1, \dots, 64\}$) and grouped into ten equal-size difficulty bins; reported numbers are within-bin means. The deciles are equal in size by construction, whereas the headline of Table 1 averages over the 100-edit sample, whose difficulty profile follows the natural COUNTERFACT distribution and is tilted, mildly, toward the easy low-$r^\star$ edits. Figure 2 (left) plots the per-decile median of that distribution against the ranks the fixed-rank editors work at. The headline is recovered from the decile cells by weighting each decile by the number of the 100

sampled edits that fall in it. Assigning each of the 100 edits to its LLaMA-2-7B pool decile using the boundaries of Table A3 gives the occupancy vector

$$(n_1, \ldots, n_{10}) = (11, 13, 10, 12, 9, 10, 9, 8, 10, 8), \tag{A16}$$

whose entries sum to the 100 sampled edits. The two edit sets are disjoint (seed 42 and seed 43, Appendix A.9), so the reweighted decile average estimates the headline. Applying these weights to a method's ten decile cells reproduces its headline entry in Table 1 to within 1.6 points. For FISHER-LORA on LLaMA-2-7B, for instance,

$$\begin{aligned} \tfrac{1}{100}\big(&11{\cdot}99.2 + 13{\cdot}98.7 + 10{\cdot}98.6 + 12{\cdot}98.2 \\ &+ 9{\cdot}96.8 + 10{\cdot}96.0 + 9{\cdot}93.7 + 8{\cdot}93.8 \\ &+ 10{\cdot}91.8 + 8{\cdot}88.1\big) = 95.86, \end{aligned} \tag{A17}$$

against the 96.4 headline, a residual of -0.54 points. The same exercise on the other LLaMA-2-7B rows of Table 3 leaves residuals of $+1.16$ (ROME), $+0.92$ (MEMIT), -1.37 (MELO), -0.68 (ELDER), $+0.69$ (AdaLoRA) and -1.47 (GaLore). On the Mistral-7B rows, evaluated with the Mistral-7B occupancy vector $(12, 14, 11, 9, 11, 9, 8, 10, 8, 8)$, the residuals are $+1.56$, $+1.32$, -0.89, -0.57 and $+0.46$ for ROME, MEMIT, MELO, ELDER and FISHER-LORA. The residuals run up to about a point and a half in either direction and carry no systematic sign, which is the sampling difference one expects between two disjoint draws from the same benchmark.

Rank-one editors (ROME, MEMIT) decline steadily once the decile median $r^\star$ exceeds their fixed rank, and ROME falls below half its decile-1 Reliability by decile 9 on LLaMA-2-7B and by decile 8 on Mistral-7B. Rank-eight MELO holds Reliability above 71% through decile 8 on both models and loses a further 13 to 16 points over the last two deciles. ELDER at rank 32 stays above 79% Reliability on every LLaMA-2-7B decile and above 76% on every Mistral-7B decile, validating that an over-rank choice is broadly recoverable for Reliability at the cost of the Locality damage reported in Table 1. FISHER-LORA preserves Reliability above 85% across all deciles on both models.

Per-decile correctness failure and the joint-failure row. The Reliability cells above are within-decile *means* of a continuous per-edit score and therefore do not by themselves determine how many edits an editor gets *right*. The joint-failure row of Table A7 is built from a different, binary per-edit event. An editor *fails* on edit e when it does not meet the (ε, δ)-correctness criterion of Eq. (2) on that edit, in the same operational form used for the ground-truth $r^\star$ search of Appendix A.9: per-edit reliability $\geq 95\%$ *and* per-edit locality $\geq 90\%$. This is stricter than the 0.5-threshold "Failure" column of Table 2, which counts edits whose Reliability falls below 0.5. An editor can average 99.8% reliability on a decile and still miss the $95\%/90\%$ criterion on nearly half of its edits, because the criterion also has a locality half. The two columns count different events. Table A4 reports the per-decile criterion-failure rate of the three baselines. The joint-failure row of Table A7 is the per-edit intersection of these three events, so each of its cells is at most the column-wise minimum of Table A4, and its unweighted mean over the ten equal-size deciles, 34.82%, is the joint-failure base rate quoted as 34.8% in Section 2. ELDER's criterion-failure rate is U-shaped, not monotone, because the two halves of the criterion bind at opposite ends of the difficulty range. On the easy deciles it is the locality half that binds (its rank-32 update leaves decile-1 Locality at 61.4%, so at most $61.4/90 = 68.2\%$ of those edits can clear the 90% locality requirement and at least 31.8% must fail the criterion however high their reliability), while on the hard deciles it is the reliability half. The two rank-one editors are the binding constraint on the easy end instead, and it is the intersection with them that keeps the joint row low there.

Interval estimates on the per-decile rates. Each cell of Table A4 is a proportion over the 500 edits of its decile. Table A5 gives the exact Clopper-Pearson 95% interval for every cell, computed from that denominator, together with the joint-failure row of Table A7 on the same basis. On the 5,000-edit pool the 34.8% joint-failure base rate carries the interval $[33.5, 36.2]$.

The hard slice under the (ε, δ) criterion. COUNTERFACT-HARD is the union of D8, D9 and D10 in equal 500-edit parts (Appendix A.9). On the slice the criterion-failure rate is 85.7% for ROME, 80.0% for MEMIT and 71.1% for ELDER, with the three failing jointly on 64.6% of the slice. These rates are on the $\geq 95\%$ reliability and $\geq 90\%$ locality criterion of Eq. (2). In counts,

Table A4: Per-decile (ε, δ)-correctness failure rate (%) of the three baselines on the 5,000-edit pool (LLaMA-2-7B). An edit counts as a failure when per-edit reliability $< 95\%$ or per-edit locality $< 90\%$. The "Joint fail" row of Table A7 is the per-edit intersection of the three events and is therefore bounded above by the column-wise minimum of this table, cell by cell.

Method	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
ROME	10.2	16.4	24.6	34.0	46.2	57.4	67.8	77.6	86.4	93.0
MEMIT	8.6	13.0	19.8	28.4	39.6	50.2	60.4	71.0	80.6	88.4
ELDER	46.0	44.2	42.6	42.0	43.8	47.4	53.0	63.4	71.8	78.2
Column-wise min	8.6	13.0	19.8	28.4	39.6	47.4	53.0	63.4	71.8	78.2
Joint fail (Table A7)	4.0	7.8	12.6	19.0	27.4	36.8	46.8	57.2	65.0	71.6

Table A5: Exact binomial 95% confidence intervals (Clopper-Pearson, $n = 500$ edits per decile) for the per-decile (ε, δ)-correctness failure rates of Table A4 and for the joint-failure row of Table A7. The point estimates repeat those tables.

Decile	ROME	MEMIT	ELDER	Joint fail
D1	10.2 [7.7, 13.2]	8.6 [6.3, 11.4]	46.0 [41.6, 50.5]	4.0 [2.5, 6.1]
D2	16.4 [13.3, 19.9]	13.0 [10.2, 16.3]	44.2 [39.8, 48.7]	7.8 [5.6, 10.5]
D3	24.6 [20.9, 28.6]	19.8 [16.4, 23.6]	42.6 [38.2, 47.1]	12.6 [9.8, 15.8]
D4	34.0 [29.9, 38.3]	28.4 [24.5, 32.6]	42.0 [37.6, 46.5]	19.0 [15.7, 22.7]
D5	46.2 [41.8, 50.7]	39.6 [35.3, 44.0]	43.8 [39.4, 48.3]	27.4 [23.5, 31.5]
D6	57.4 [52.9, 61.8]	50.2 [45.7, 54.7]	47.4 [42.9, 51.9]	36.8 [32.6, 41.2]
D7	67.8 [63.5, 71.9]	60.4 [56.0, 64.7]	53.0 [48.5, 57.4]	46.8 [42.4, 51.3]
D8	77.6 [73.7, 81.2]	71.0 [66.8, 74.9]	63.4 [59.0, 67.6]	57.2 [52.7, 61.6]
D9	86.4 [83.1, 89.3]	80.6 [76.9, 84.0]	71.8 [67.6, 75.7]	65.0 [60.6, 69.2]
D10	93.0 [90.4, 95.1]	88.4 [85.3, 91.1]	78.2 [74.3, 81.7]	71.6 [67.4, 75.5]

the slice holds 1,285, 1,200, 1,067 and 969 of these failures, with exact Clopper-Pearson 95% intervals $[83.8, 87.4]$, $[77.9, 82.0]$, $[68.8, 73.4]$ and $[62.1, 67.0]$; on the same basis the Failure column of Table 2 gives $[14.0, 17.7]$ for ELDER (237 edits) and $[6.3, 9.1]$ for FISHER-LORA (114 edits).

Failure mass held by the slice. At 500 edits per decile the three hardest deciles carry 286, 325 and 358 joint failures. COUNTERFACT-HARD holds 969 of the pool's 1,741 joint failures, 55.7% of the failure mass in 30% of the edits, an enrichment of $1.86\times$ on a pool base rate of 34.8%. The remaining 772 joint failures, 44.3% of the total, lie below $r^\star = 16$ and outside the slice.

Locality decomposition. Locality stratifies the opposite way. ELDER's Locality on decile 1 ($r^\star \leq 2$) drops to 61.4% on LLaMA-2-7B, because its rank-32 update has 30+ free directions to leak into the locality set. FISHER-LORA's decile-1 Locality is 97.5% because the small rank it allocates on this decile leaves few free directions to leak into. On decile 10 this reverses: ELDER's Locality is 88.7% (fewer excess directions relative to the larger update) while FISHER-LORA's is 81.2%. The per-decile mean Locality for FISHER-LORA on LLaMA-2-7B is $97.5, 96.9, 95.6, 94.5, 92.9, 91.5, 88.7, 87.1, 85.2, 81.2$ across the ten uniform difficulty deciles of the pool; weighting these by the occupancy vector of Eq. (A16) gives 91.70%, against the 91.3% Locality headline of Table 1, on the same weights used for the Reliability column.

WikiText-2 perplexity probe. The post-edit WikiText-2 perplexity delta versus the unedited base model is +0.04 for FISHER-LORA on LLaMA-2-7B (negligible), +0.18 for ELDER, +0.06 for MELO, +0.02 for ROME, and +0.03 for MEMIT.

APPENDIX A.18: FULL ZSRE RESULTS AND DIFFICULTY STRATIFICATION

Table A6 gives the zsRE comparison on all three backbones, of which Table 2 is the LLaMA-2-7B subset. The baseline Locality column reorders between the two benchmarks; only ELDER's position at the bottom of that column is common to both. Table A7 gives the decile stratification that

Table A6: Cross-dataset generalisation on zsRE (100 random edits): Reliability (%, ↑) and Locality (%, ↑), against the COUNTERFACT pattern of Table 1. Best in **bold**.

		LLaMA-2-7B		Mistral-7B	GPT-2-XL
Method	**Rank**	**Reliability**	**Locality**	**Reliability**	**Reliability**
ROME	1	76.2	83.1	74.3	71.2
MEMIT	1	81.5	86.1	79.1	78.0
MELO	8	88.6	86.5	86.4	86.2
ELDER	32	93.8	81.0	92.6	91.0
FISHER-LORA	$\widehat{r}^\star$	**96.9**	**92.4**	**95.6**	**95.9**

Table A7: Difficulty predictor validity: joint failure rate of ROME, MEMIT, and ELDER by $r^\star$ decile on the 5,000-edit pool (LLaMA-2-7B), 500 edits per decile, against the median ground-truth rank of each decile.

Decile	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Median $r^\star$	1	3	5	7	9	12	15	19	24	30
Joint fail (%)	4.0	7.8	12.6	19.0	27.4	36.8	46.8	57.2	65.0	71.6

Figure 2 (right) plots and from which the 34.8% joint-failure base rate quoted in the main text is the unweighted row mean.

APPENDIX A.19: ESTIMATOR COMPARISON AND RANK-FLOOR AGREEMENT

Estimator alternatives. We test two alternatives to the attribution-patching estimator of Eq. (8) on LLaMA-2-7B / 100 edits: the full causal Fisher of Eq. (5) as an oracle, and random projection at matched rank. The oracle reaches 96.7% Reliability and 91.5% Locality at 620× the per-edit compute cost, since it requires sampling $y \sim p_\theta(\cdot \mid x)$ and accumulating per-token gradient outer products instead of reading one backward pass. The attribution-patched estimator reaches 96.4% at the cost of that single pass, recovering 99.7% of the oracle's Reliability. Random projection is the matched-rank rung of Table 4 and reduces Reliability by 9.7 points. The decisive design choice is the causal-Fisher approximation, not low-rank structure.

Rank-floor agreement. Across the 100 random edits on LLaMA-2-7B the empirical ratio $r_{\text{used}}/r^\star$ has mean 1.02, fifth percentile 0.96, and ninety-fifth percentile 1.10, and the difference $r_{\text{used}} - r^\star$ is at most one dimension on 89% of the edits and at most two on 97%, against the 1.27-dimension one-sided slack that Eq. (7) gives at $\varepsilon = 0.05$, $\delta = 0.1$, $\sigma_\theta = 0.1$ and $\lambda^+_{\min}(F_e^{\text{causal}}) = 1$. The ground-truth $r^\star$ against which the ratio is computed is the cross-method rank search of Appendix A.20.

Out-of-family extension of the difficulty correlation. The Spearman correlation of $\rho = 0.74$ reported in the main text is between the per-edit $\widehat{r}^\star$ and the joint (ε, δ)-failure indicator of ROME, MEMIT, and ELDER on the 5,000-edit pool; the corresponding Pearson coefficient is 0.70. Extending the joint-failure set to the memory-based editors GRACE and WISE, which allocate no rank at all, leaves the correlation at $\rho = 0.71$.

APPENDIX A.20: SENSITIVITY ANALYSES

The sweeps below are reported on the 100 random COUNTERFACT edits / LLaMA-2-7B. The four constants are: the eigenvalue threshold τ that defines $\widehat{r}^\star$, the attribution-patching cutoff τ_{AP} that defines $\mathcal{V}_e$, the locality regulariser weight β, and the optimiser-step budget T. The last three are selected on the development split; τ is a fixed constant of the method and is selected on no split, so its sweep reports sensitivity rather than a choice. Table 5 reports the eigenvalue-threshold sweep already discussed in Section 5.3; here we expand the remaining three sweeps.

Attribution-patching cutoff τ_{AP}. Sweeping the relative cutoff $\tau_{\mathrm{AP}} \in \{0.005, 0.01, 0.02, 0.05\}$ of Eq. (4) changes the dimension of $\mathcal{V}_e$ from a median of 86 at $\tau_{\mathrm{AP}} = 0.005$ to a median of 19 at $\tau_{\mathrm{AP}} = 0.05$. Across this range the cutoff is not binding on the rank: every component that clears the eigenvalue threshold of Eq. (6) is retained at all four settings, so the four give the same edit and the same readings, Reliability 96.4% and Locality 91.3%. The estimator is insensitive to the cutoff across the range of $|\mathcal{V}_e|$ this sweep spans. The deployed value $\tau_{\mathrm{AP}} = 0.01$ was selected on the development split (Appendix A.9).

Locality regulariser weight β. Sweeping $\beta \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$ produces the trade-off $96.7\%, 96.6\%, 96.4\%, 92.0\%, 81.5\%$ Reliability and $80.4\%, 87.6\%, 91.3\%, 95.0\%, 96.8\%$ Locality. The deployed value $\beta = 0.1$, selected on the development split (Appendix A.9), sits on the upper-right shoulder of the trade-off curve.

Optimiser-step budget T. Sweeping $T \in \{10, 15, 20, 25, 50\}$ produces the saturating curve 87.3%, 92.5%, 95.6%, 96.4%, 96.6% Reliability with Locality stable at 90.4% to 91.5%. The deployed value $T = 25$, selected on the development split (Appendix A.9), sits at the saturation knee; doubling to $T = 50$ buys 0.2 percentage points of Reliability at $2\times$ the optimisation compute.

Layer-set $\mathcal{S}$ ablation. We tested four alternative layer sets to the canonical $\{4, \ldots, 8\}$ of Appendix A.11 on LLaMA-2-7B: $\{0, \ldots, 31\}$ (all layers), $\{12, \ldots, 16\}$ (mid-network), $\{20, \ldots, 24\}$ (late mid-network), $\{4, \ldots, 14\}$ (extended early-mid block). Reliability is 94.0% (all layers), 96.4% (canonical), 93.7% (mid-network), 90.4% (late mid-network), 96.1% (extended). The canonical block matches the causal-tracing localisation of Meng et al. (2022) and matches the extended block (96.4% versus 96.1%) at lower compute.

Ground-truth $r^\star$ measurement protocol. The ground-truth $r^\star(e)$ used in the per-decile stratification of Appendix A.17 is computed by search over integer ranks $r \in \{1, \ldots, 64\}$ for the smallest rank achieving $\geq 95\%$ Reliability and $\geq 90\%$ Locality across the strongest of the LoRA-based baselines, evaluated independently of FISHER-LoRA. The cross-method ground truth differs from the same search restricted to FISHER-LoRA by at most one rank dimension on 84% of the edits and by at most two rank dimensions on 95%, consistent with the observed rank floor being a property of the (model, edit) pair, not of any particular editor. That comparison is between two ground-truth protocols; the agreement between the Phase 1 estimate and the ground truth is the separate reading of Appendix A.19.

APPENDIX A.21: SCALE INVARIANCE OF $\widehat{r}^\star$ AND THE CLASS OF NORM-BASED PREDICTORS

Two quantities can be read off the same backward pass that produces the contrastive gradient g_e: the energy it carries on the causal subspace, and the count $\widehat{r}^\star$ of Eq. (6). They are different invariants of that gradient, and this subsection states in what sense.

An energy identity. Under the estimator of Eq. (8) the spectrum is diagonal in the component basis, with $\lambda_c = (g_e^\top u_c)^2$ by Eq. (9). Summing over the components retained by Eq. (4) gives

$$\left\|\Pi_{\mathcal{V}_e} g_e\right\|^2 \;=\; \sum_{c:\, u_c \in \mathcal{V}_e} \lambda_c, \tag{A18}$$

so the squared gradient norm on the causal subspace is the total spectral mass, while $\widehat{r}^\star$ of Eq. (6) is the number of eigenvalues within a factor $\tau = 10^{-2}$ of the largest. One is a sum of the spectrum, the other a count of its entries above a relative cutoff.

Homogeneity. Let $g_e \mapsto c\, g_e$ for any $c > 0$. The attribution scores of Eq. (3) become $c\, a_c(e)$, and since the cutoff of Eq. (4) is relative to $\max_{c'} a_{c'}(e)$, the retained set $\mathcal{V}_e$ is unchanged. Every λ_c becomes $c^2\lambda_c$, hence $\lambda_{\max}$ becomes $c^2\lambda_{\max}$, and the criterion $\lambda_i \geq \tau\lambda_{\max}$ of Eq. (6) holds for exactly the same indices. Therefore

$$\widehat{r}^\star(c\, g_e) \;=\; \widehat{r}^\star(g_e) \quad \text{for all } c > 0, \qquad \left\|\Pi_{\mathcal{V}_e}(c\, g_e)\right\| \;=\; c\left\|\Pi_{\mathcal{V}_e} g_e\right\|. \tag{A19}$$

$\widehat{r}^\star$ is invariant under the positive scaling group; the gradient norm is homogeneous of degree one under it. Any non-constant monotone predictor built from $\|\Pi_{\mathcal{V}_e} g_e\|$ alone varies under that scaling and cannot have the first, so the two belong to different invariant classes and no reparametrisation of the norm recovers $\widehat{r}^\star$.

The equal-energy surface. Invariance alone leaves open whether the two quantities are related once the scale is fixed. They are not. Fix the energy at $\|\Pi_{\mathcal{V}_e} g_e\|^2 = E$ and let $m = |\mathcal{V}_e|$ be the number of retained components, empirically in $[20, 90]$ with median 43 on LLaMA-2-7B. For any integer k with $1 \le k \le m$, the score profile that places all but a fraction of order $m\tau_{\mathrm{AP}}^2$ of E in equal parts on k components, and spreads the remainder over the rest above the attribution cutoff of Eq. (4) but below $\tau\lambda_{\max}$, satisfies Eq. (A18) with the same E, has $\lambda_{\max}$ equal to E/k up to that fraction, and has exactly k eigenvalues equal to $\lambda_{\max}$, so Eq. (6) returns $\widehat{r}^\star = k$. At $k = 1$ the energy concentrates on a single direction and $\widehat{r}^\star = 1$; at $k = m$ the energy is uniform across the subspace and $\widehat{r}^\star = m$. The same E is therefore consistent with every value of $\widehat{r}^\star$ the estimator can return. In the reverse direction, holding a spectral profile fixed and rescaling by c moves the norm over the whole positive line at constant $\widehat{r}^\star$, by Eq. (A19). The pair $(\|\Pi_{\mathcal{V}_e} g_e\|, \widehat{r}^\star)$ thus takes values across the full product of its two ranges.

Consequence for difficulty prediction. The gradient norm measures the size of the edit's causal gradient; $\widehat{r}^\star$ measures how many directions that movement is spread over, a peaked spectrum giving a small count and a flat one a large count. The rank floor of Section 4.3 and the correlation of Spearman $\rho = 0.74$ in Section 5.3 are statements about the second property, and a construction of the kind above shows that the first does not determine it at fixed scale. The same argument applies to any quantity that is a monotone function of the total causal energy on $\mathcal{V}_e$, since such a function is determined by E and is therefore constant along the surface on which $\widehat{r}^\star$ ranges from 1 to $|\mathcal{V}_e|$.

Appendix A.22: The Ablation Ladder and the Point Projection

Ablation ladder rows. In Table 4 of the main text, the rank-only row uses the per-edit rank $\widehat{r}^\star$ with an unconstrained LoRA subspace and random initialisation; the random-subspace row additionally confines the update to a *random* $\widehat{r}^\star$-dimensional subspace, isolating the effect of constraining the subspace from the effect of choosing the causal one; the causal-subspace row replaces that random subspace with the attribution-patched one at the identical rank, still under random initialisation; the full row adds the Fisher-eigenvector initialisation. The fixed-rank reference row is the ELDER configuration of Table 1.

Point projection. Random projection at matched rank confines the update to a random $\widehat{r}^\star$-dimensional subspace and carries no spectrum from which to initialise; it reaches 86.7 Reliability and 88.0 Locality, a drop of 9.7 Reliability points against the full method.

Appendix A.23: Future Work

Three directions extend the Causal-Fisher rank framework. The first is the slack term. Sharpening the Cramér-Rao argument so that the bound conditions on the spectral gap between $\lambda_{\max}$ and $\lambda_{\min}^+$ instead of on $\lambda_{\min}^+$ alone would improve failure prediction on the 14 near-null edits identified in Section 5.3. The second is scope. The attribution-patching estimator $\widehat{F}_e^{\mathrm{AP}}$ is a sum of per-component outer products of a contrastive gradient, and is therefore defined for single-fact edits. Extending it to multi-fact and sequential editing in the style of MEMIT and WISE requires replacing $g_e g_e^\top$ with a low-rank covariance over an interleaved edit batch, and lifting the energy argument of Appendix A.4 to the joint rank of the batch's Causal FIM. That would convert the present per-edit floor into a per-trajectory rank-budget law evaluable on 10^4-edit zsRE streams. The third is reach beyond additive LoRA. The Cramér-Rao argument only requires the editor to live in a low-rank-structured submanifold of the parameter space. Deriving matching rank floors and Fisher-eigenvector achievability constructions for prefix tuning, IA^3, and adapter modules would test whether the Causal Fisher rank is the correct invariant across the full parameter-efficient family or specific to additive LoRA blocks.