UAR-LoRA: Uncertainty-Aware Rank-Adaptive Low-Rank Adaptation for Lifelong Knowledge Editing

Anonymous authors
Paper under double-blind review

Abstract

Lifelong knowledge editing keeps a deployed large language model factually current one edit at a time, so thousands of corrections must accumulate without eroding earlier edits or unrelated knowledge. Existing low-rank adapter editors do not control, for each arriving edit, how much capacity its expert takes, which queries reach it and where it sits: a rank fixed before the stream starves hard corrections, a router without a calibrated threshold serves queries the pool should leave to the frozen base, and randomly initialised experts overlap earlier ones. We propose UAR-LoRA, an uncertainty-aware, rank-adaptive framework that treats capacity, admissibility and geometry as three conditions of one heuristic capacity-coherence criterion and meets them as each expert enters the pool. An Entropy Probe and Adaptive Rank Selector sizes each expert from the frozen base's predictive entropy, a Confidence-Gated Router returns queries below a calibrated threshold to the frozen base, and a Rank-Aware Orthogonal Initialiser places each new expert in directions no earlier expert occupies while free ones remain, with a truncated-SVD fallback once they run out. Averaged over three backbones on CounterFact at $T{=}1000$, UAR-LoRA reaches mean Reliability 92.9 against 90.4 for ELDER, and on LLaMA-2-7B it lowers Expected Calibration Error from 0.19 to 0.11 at a $0.44\times$ trainable-parameter ratio relative to fixed rank-32 LoRA. No single-module variant exceeds 91.0 Reliability against 93.0 with all three, and the probe matches a fixed rank of 32 at $0.44\times$ rather than $1.09\times$. Forgetting also accumulates more slowly: over 5000 ZsRE edits UAR-LoRA retains 94.1 Reliability where ELDER falls to 86.2, losing 2.4 points from the first edit against ELDER's 7.8.

1 Introduction

Lifelong knowledge editing keeps a deployed large language model (LLM) factually current by correcting its errors one edit at a time, without retraining (Wang et al., 2024b; Hartvigsen et al., 2023; Li et al., 2025). Such models are served with their weights frozen (Touvron et al., 2023; Grattafiori et al., 2024), yet the facts they encode keep changing: entities acquire new attributes and earlier assertions need correction. The edits therefore arrive as an unbounded stream, and the difficulty is not whether one edit applies but whether thousands accumulate without eroding what came before. A lifelong editor is judged on three counts after every edit: reliability on the edited fact, generality under paraphrase and entailment, and locality on the knowledge it should leave unchanged.

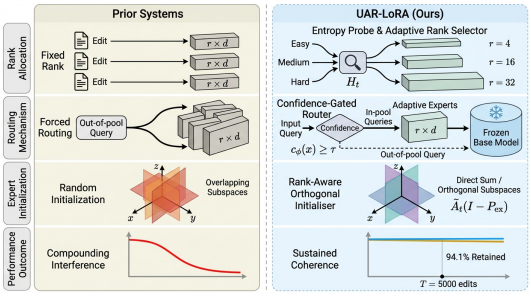

Figure 1: **Left:** prior editors. **Right:** UAR-LoRA: sized rank, calibrated gate, orthogonal init.

Editors differ in where they store a correction: direct weight editors rewrite the projections that hold a fact (Meng et al., 2022; 2023), memory editors consult a growing external store at inference (Hartvigsen et al., 2023), and parameter-efficient editors install each correction as a low-rank adapter beside the frozen backbone (Hu et al., 2022; Yu et al., 2024). The adapter line has advanced along

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.

three axes that matter for a stream: rank allocation spends adapter capacity by importance rather than uniformly (Zhang et al., 2023); routing over a growing adapter pool keeps early edits from being overwritten, with ELDER learning the routing jointly and deferring to the base (Li et al., 2025) and SoLA routing by semantic scope (Luo et al., 2026); and orthogonal subspaces keep successive updates apart, in continual learning (Wang et al., 2023) and in editing (Xu et al., 2025). Together, these advances make a pool of low-rank adapters a practical basis for lifelong editing.

However, no existing editor controls all three axes for each arriving edit, and Figure 1 (left) shows how the resulting failures compound in a long stream. A rank fixed before the stream over-provisions a one-token attribute swap yet starves a correction that must reach entailed queries; a routing score without a calibrated threshold serves queries the pool should leave to the frozen base; and experts drawn at random write into directions that earlier edits already occupy. The three decide one quantity, the interference each new expert exerts on the edits and knowledge in place, through how much room it takes, which queries reach it, and where it sits. Capacity, admissibility and geometry are thus not three knobs to tune apart but three conditions of one heuristic capacity-coherence criterion.

To address these failures, we propose UAR-LoRA, an uncertainty-aware, rank-adaptive low-rank adaptation framework that meets the three conditions edit by edit. Figure 1 (right) shows its three components: an Entropy Probe and Adaptive Rank Selector that sizes each expert from the frozen base's predictive entropy; a Confidence-Gated Router that returns queries below a calibrated threshold to the frozen base; and a Rank-Aware Orthogonal Initialiser that places each new expert in directions no earlier expert occupies while free ones remain. The key insight is that interference in an adapter pool is set when each expert enters it: how much room the expert takes, which queries can reach it, and which directions it occupies together determine how far it can disturb the edits already stored. Our main contributions are as follows:

- We identify three failures of lifelong adapter pools, capacity misallocation, uncalibrated routing and interference accumulation, and cast them as three conditions of a single heuristic capacity-coherence criterion.
- We propose UAR-LoRA: an Entropy Probe and Adaptive Rank Selector sizes each expert by entropy, a Confidence-Gated Router abstains below a calibrated threshold, and a Rank-Aware Orthogonal Initialiser places each expert in unoccupied directions while any remain.
- Across three backbones on CounterFact, UAR-LoRA reaches 92.9 mean Reliability against 90.4 for ELDER, at a per-edit 0.44× parameter ratio on LLaMA-2-7B relative to fixed rank-32 LoRA; over 5000 ZsRE edits it retains 94.1 where ELDER falls to 86.2.

2 MOTIVATION

In lifelong knowledge editing, we identify three challenges that limit the adapter-routing recipe over a long edit stream.

Challenge #1: One rank cannot fit heterogeneous edits. Edits are not equally hard. Overwriting a single-token factual swap (one entity's country of citizenship) demands almost no capacity, whereas a diffuse multi-hop correction that must propagate through several entailed queries demands far more. A pipeline that commits to one bottleneck rank for the entire stream over-provisions the easy edits and under-provisions the hard ones. Figure 2 (left) makes the cost visible: a narrow rank pool $R = \{4, 8\}$ reaches 90.2 Reliability, whereas allowing the rank to adapt across $\{4, 8, 16, 32\}$ lifts Reliability to 93.0 at a higher but still sub-unit parameter cost. The failure is structural: any fixed rank pays either an accuracy tax on hard edits or a parameter tax on easy ones. *From what signal, available before the edit is trained, should the capacity of each edit be decided?*

Challenge #2: The routing score is uncalibrated. A closed-set softmax router distributes unit probability mass across the current expert pool. Every query therefore receives a within-pool score, including queries that merely resemble an edited fact and queries that fall just outside the edited region. Nothing in the score decides whether the pool should touch them at all. Figure 2 (middle) shows the trade-off the gate governs: at the loosest gate we sweep, the abstention decision trails in-pool routing by 24.6 points, and closing the gate recovers that decision up to $\tau = 0.8$ while eroding in-pool accuracy throughout. *How can a route back to the frozen base be governed by a calibrated threshold, so that a query the pool should not touch is left untouched?*

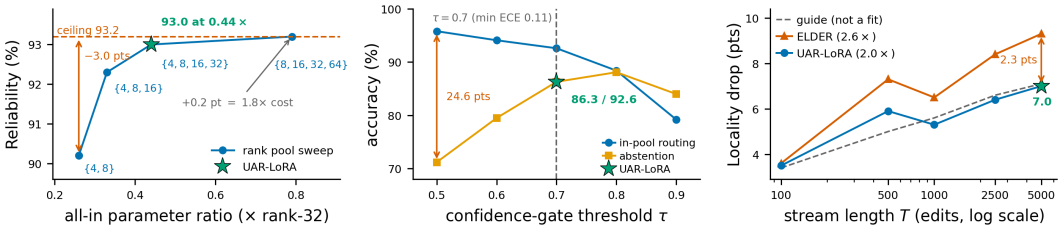

Figure 2: Three diagnostics motivating UAR-LoRA, one per challenge. *Left:* capacity misallocation. *Middle:* uncalibrated routing. *Right:* interference accumulation.

Challenge #3: Random initialisation makes interference structurally unavoidable. When each new expert is initialised at random, its row space overlaps arbitrarily with subspaces already occupied by earlier experts. Successive edits collide in parameter space, and collisions perturb previously edited facts as the stream lengthens. Figure 2 (right) quantifies the drift: ELDER's cumulative Locality drop reaches 9.3 at $T = 5000$, whereas UAR-LoRA holds the drop to 7.0 over the same stream. *Where in representation space should a new expert be placed so that installing it does not disturb the edits already stored?*

The failures act through one quantity, the disturbance a new expert causes to stored edits: its rank sets how much room it takes, the router which queries reach it, and its initialisation which directions it shares. Capacity, admissibility and geometry are thus conditions of one capacity-coherence criterion.

3 Problem Formulation

To address these challenges mentioned in Section 2, we formulate lifelong editing as a constrained optimisation problem: a growing family of low-rank experts maximises the editing criteria,

$$\max_{\{(A_t,B_t,r_t,c_\phi)\}} \frac{1}{T}\sum_{t=1}^{T}\Big[\mathrm{Rel}_t + \mathrm{Gen}_t + \mathrm{Loc}_t\Big], \tag{1}$$

subject to an average adapter-parameter ceiling,

$$\sum_{t=1}^{T} r_t\,(d_{\text{in}} + d_{\text{out}}) \;\le\; r_{\text{ref}}\,(d_{\text{in}} + d_{\text{out}})\,T. \tag{2}$$

Here $e_t = (x_t, y_t)$ is the t-th of T edits arriving one at a time to a frozen base f_{θ_0}; $\Delta W_t = B_t A_t$ is the rank-r_t expert that edit t installs into the pool E_t (Hu et al., 2022), and c_ϕ decides whether a query is served by the pool; Rel_t, Gen_t and Loc_t are the reliability, generality and locality of the pool after t edits; d_{in} and d_{out} are the input and output widths of an adapted projection; and r_{ref} is the reference rank, so an edit may spend no more on average than one fixed-rank adapter.

To address Challenge #1, the rank r_t is drawn per edit from the frozen base's uncertainty on x_t rather than fixed in advance. To address Challenge #2, the gate may withhold the pool entirely and rarely admits a query unrelated to any stored edit. To address Challenge #3, each new expert is placed in directions no earlier expert occupies, so the occupied directions accumulate as a direct sum at placement. These are the three conditions of the capacity-coherence criterion.

4 Method

4.1 Overall Framework

In order to solve the problem formulated in Section 3, we propose UAR-LoRA, whose per-edit pipeline is shown in Figure 3. UAR-LoRA comprises three modules: the Entropy Probe and Adaptive Rank Selector sizes each new expert from the frozen base's uncertainty; the Confidence-Gated Router decides whether a query is served by the pool or returned to the frozen base; and the Rank-Aware Orthogonal Initialiser places each new expert in unoccupied directions while any remain.

The three modules form one system rather than three dials. All three act on the same expert as it enters the pool: the probe fixes how much room the expert occupies, the gate decides which queries can reach the expert at all, and the initialiser places that room in directions no earlier expert holds. Each module thus controls one way a new expert can disturb stored edits.

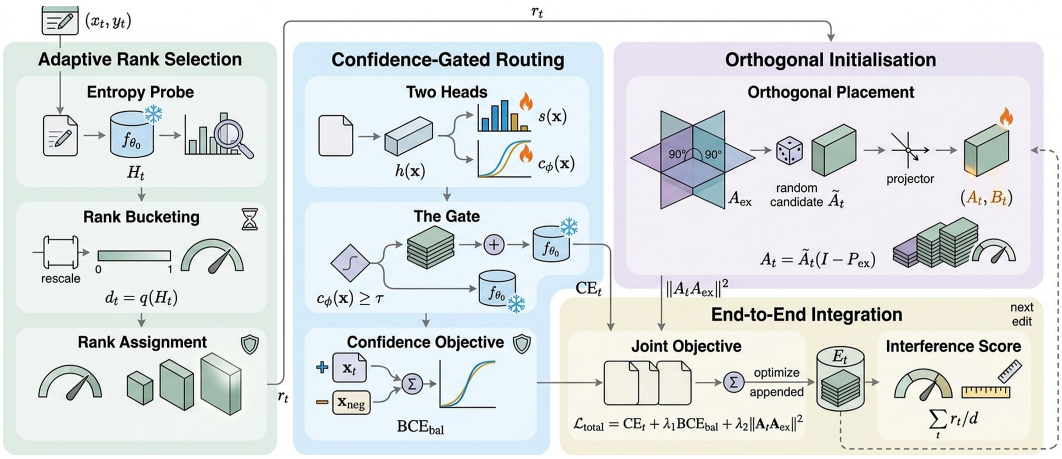

Figure 3: UAR-LoRA for one edit. **Left:** frozen-base entropy sets the rank. **Middle:** the gate returns queries below τ to the frozen base. **Right:** orthogonal placement. **Bottom:** one joint objective.

4.2 Adaptive Rank Selection

To address Challenge #1, we propose the Entropy Probe and Adaptive Rank Selector, which sizes each new expert by the frozen base's uncertainty. Figure 3 (left) shows its two steps: a probe reads the entropy of the edit query, and a rule maps it to a rank.

Entropy probe. We read the predictive entropy of the frozen base on the edit query,

$$H_t \;=\; -\sum_{y \in V} p_{\theta_0}(y \mid x_t)\, \log p_{\theta_0}(y \mid x_t), \tag{3}$$

where $p_{\theta_0}(\cdot \mid x_t)$ is the frozen base's next-token distribution on the query x_t and V is the vocabulary. A base that is confident about the fact yields a low H_t, and an uncertain one a high H_t. The probe uses this signal because it is available before the edit is trained and costs a single forward pass of the frozen base.

Rank bucketing. Raw entropy is not comparable across prompts, so we normalise it by its own recent history: $d_t = q(H_t) \in [0, 1]$ rescales H_t by the range of entropies over a trailing window of W edits, with a cold-start value $d_t = d_0$ for $t \leq t_0$ before enough history has accumulated. Normalising by recent history makes the rank depend on how hard an edit is relative to the stream rather than on the entropy scale of its prompt, so the whole rank pool stays in use as the stream drifts. Details about the normaliser, its clipping, and its behaviour before the window fills in Appendix A.4.

Rank assignment. The normalised difficulty is mapped to a discrete rank bucket,

$$r_t \;=\; R\big[\min\big(\lfloor d_t\,|R|\rfloor,\; |R|-1\big)\big]\,, \tag{4}$$

where R is a fixed increasing rank pool and $|R|$ its size. The rule splits $[0, 1]$ into $|R|$ equal bins, so low-difficulty edits receive $\min R$ and the hardest receive $\max R$. Discrete buckets keep the experts to a small set of shapes, so the parameter cost of every edit is known before training and bounded by $\max R$. Prompts whose entropy estimate is unreliable fall back to a clamped rank r_{fb}. More details are provided in Appendix A.6.

Through this module we obtain a per-edit rank r_t read from the entropy proxy, so low-entropy edits cost the smallest rank in the pool and high-entropy edits the largest. That rank fixes how much room the orthogonal initialiser must find.

4.3 Confidence-Gated Routing

To address Challenge #2, we propose the Confidence-Gated Router, which gives the pool a calibrated route back to the frozen base. Figure 3 (middle) shows its two heads: one chooses which expert serves a query, and one decides whether any expert serves it.

Two heads. Both heads read $h(x)$, the frozen base's hidden representation of a query x. An expert-weight head produces a distribution over the $|E|$ stored experts,

$$s(x) \;=\; \mathrm{softmax}(W_s\, h(x))\,, \tag{5}$$

where W_s holds one row per stored expert, while a separate scalar confidence head

$$c_\phi(x) \;=\; \sigma\big(w_c^\top h(x)\big) \in (0,1), \tag{6}$$

with weight vector w_c and logistic function σ, estimates whether x belongs to the edited region at all. The separate head is what makes abstention possible: the softmax of Eq. (5) always distributes unit mass over the pool, so a second, scalar output expresses that no expert should serve x.

The gate. The two heads are combined by a hard-threshold gate. On an adapted projection with frozen weight W_0 and input u, the gate adds the expert mixture only when the confidence clears the threshold τ:

$$g_x(u) = \begin{cases} W_0 u + \displaystyle\sum_i s_i(x)\, B_i A_i\, u, & c_\phi(x) \ge \tau, \\ W_0 u, & c_\phi(x) < \tau, \end{cases} \tag{7}$$

so an abstained query is served by the frozen base $f_{\theta_0}(x)$ exactly. A hard threshold gives an abstained query no adapter contribution at all, which is how the gate protects the knowledge the stream should leave unchanged.

Confidence objective. The confidence head is trained with a balanced binary cross-entropy on two constructed anchors, the edit's own query x_t as the positive and a query x_{neg} drawn from the edits not yet installed as the negative,

$$\mathrm{BCE}_{\mathrm{bal}} \;=\; \mathrm{BCE}\big(c_\phi(x_t), 1\big) + \mathrm{BCE}\big(c_\phi(x_{\mathrm{neg}}), 0\big). \tag{8}$$

The head is then Platt-scaled (Platt et al., 1999) so that c_ϕ reads as a probability, and τ is chosen on a held-out calibration set. A query unrelated to any stored edit, written $x \notin E_T$, should be admitted with probability at most δ,

$$\Pr(c_\phi(x) \ge \tau \mid x \notin E_T) \;\le\; \delta, \tag{9}$$

and δ is the admissibility term that the capacity-coherence criterion carries alongside the geometric one. More details are provided in Appendices A.5 and A.11.7.

Through this module we obtain a router that may decline, returning the query to the frozen base. What remains is where the admitted experts are placed.

4.4 Orthogonal Initialisation

To address Challenge #3, we propose the Rank-Aware Orthogonal Initialiser, which places each new expert in directions no earlier expert holds while free ones remain. Figure 3 (right) shows how a random candidate is projected onto the complement of the occupied span.

Orthogonal placement. The placement is carried out on a single adapted projection with input width d. Since each existing $A_s \in \mathbb{R}^{r_s \times d}$, we collect the occupied directions by stacking the transposes: $A_{\mathrm{ex}} = [A_1^\top, \ldots, A_{t-1}^\top]$ has shape $d \times \sum_{s<t} r_s$, and its column span is the space earlier experts read. The orthogonal projector onto that span is

$$P_{\mathrm{ex}} \;=\; A_{\mathrm{ex}}\big(A_{\mathrm{ex}}^\top A_{\mathrm{ex}}\big)^{-1} A_{\mathrm{ex}}^\top, \tag{10}$$

written for the full-column-rank case and defined with the Moore-Penrose pseudoinverse in general. We draw a random candidate $\tilde{A}_t \sim \mathcal{N}(0, I)$ in $\mathbb{R}^{r_t \times d}$ and project it onto the orthogonal complement of the occupied span,

$$A_t \;\leftarrow\; \tilde{A}_t\big(I - P_{\mathrm{ex}}\big), \tag{11}$$

which yields $A_t A_{\mathrm{ex}} = 0$ regardless of the heterogeneous ranks $\{r_s\}$. At placement, the span of the union $\bigcup_t A_t$ is therefore a direct sum of dimension $\sum_t r_t$, for as long as this sum stays below d. The new expert therefore maps every input direction that earlier experts read to zero. More details about the general form of the projector are provided in Appendix A.5.

When the occupied span approaches saturation the projection switches to a truncated placement that keeps the projector of Eq. (10) well-conditioned. A soft regulariser $\lambda_2 \|A_t A_{\text{ex}}\|_F^2$ penalises the overlap that later gradient updates reintroduce. Details about the guard and its triggers in Appendix A.6.

Through this module we obtain a pool whose occupied directions accumulate as a direct sum at placement for as long as unused ones remain, which is what makes the geometric term additive. The three modules are then fitted together.

4.5 End-to-End Integration

Figure 3 (bottom) shows the three modules fitted in one pass per edit. For an arriving edit, the frozen forward pass yields its entropy and rank, the new expert is drawn in the orthogonal complement of the occupied span, and the expert and the two heads are optimised jointly under

$$\mathcal{L}_{\text{total}} \;=\; \text{CE}_t + \lambda_1 \,\text{BCE}_{\text{bal}} + \lambda_2 \left\|A_t A_{\text{ex}}\right\|_F^2, \tag{12}$$

where CE_t is the token-wise cross-entropy of the target y_t under the edited model, and λ_1 and λ_2 weight the confidence loss and the orthogonality regulariser. Fitting the modules in one pass keeps them consistent: the rank fixed by the probe sets the shape of the candidate that the initialiser projects, and the confidence head is trained on the same edit the new expert stores. The fitted expert joins the pool, and at inference every query passes the gate, as detailed in Appendix A.4.

Framework Summary. The three modules act on one analysis quantity, the interference score of the heuristic capacity-coherence criterion: while the occupied rank stays below d, $\text{Interf}(T) \leq \delta + \sum_{t \leq T} r_t/d$, where δ bounds the false-admission rate of Eq. (9) and the second term bounds the expected squared overlap of an isotropic query direction with the occupied span. The rank selector sets each r_t in that term, the orthogonal initialiser keeps the term additive at placement, and the gate trained with BCE_{bal} targets δ, so every part of the score has an owner. Details in Appendix A.5.

5 Experiments

5.1 Implementation Details

Datasets. We use four editing benchmarks: CounterFact (Meng et al., 2022) with 21,919 edits, ZsRE (De Cao et al., 2021; Levy et al., 2017) with 19,085, WikiBio (Lebret et al., 2016) with 5,946 long-form biographical edits, and a multi-hop MQuAKE-style set (Zhong et al., 2023) with 3,000 edits, each paired with one-, two- and three-hop questions. Every edit carries five paraphrase probes for Generality and ten locality probes, neighbouring facts sharing its subject or relation, for Locality. Edits are reported at stream lengths $T \in \{1, 100, 500, 1000, 2500, 5000\}$, and every report scores one fixed pool of 1,000 edits drawn in advance, so Reliability, Generality and Locality are proportions over 1,000, 5,000 and 10,000 decisions. Details in Appendix A.7.

Baselines. We compare against Fine-Tune, the locate-then-edit editors ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), the memory editors GRACE (Hartvigsen et al., 2023) and MELO (Yu et al., 2024), and the routing LoRA editors SoLA (Luo et al., 2026) and ELDER (Li et al., 2025), the last being the strongest baseline. Each uses its published configuration. Every result is reproduced under our protocol rather than carried over from the cited work. Details in Appendix A.8.

Metrics. We report Reliability (Rel ↑), Generality (Gen ↑), Locality (Loc ↑), Expected Calibration Error (ECE ↓), and the trainable-parameter ratio (Param ↓), the last taken against the fixed rank-32 LoRA reference footprint. ECE uses $M = 10$ equal-width confidence bins and is computed on the evaluation predictions, except in the gate-threshold sweep, where it is computed on the calibration set and scores the gate's admit-or-abstain decision. The reference footprint is 46.4M parameters, and the Param numerator of UAR-LoRA is its adapters averaged over the stream together with the router weights and the confidence head at their full size. More details are provided in Appendix A.9.

Hyperparameters. Adapters are trained with AdamW ($\eta = 1\text{e}{-3}$, $S = 50$ steps) over the rank pool $R = \{4, 8, 16, 32\}$ with $r_{\text{ref}} = 32$, $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$. The probe normalises over $W = 1000$ edits, the first $t_0 = 100$ edits take $d_0 = 0.5$, and unreliable estimates fall back to $r_{\text{fb}} = 16$. The threshold $\tau = 0.7$ minimises ECE over $\{0.5, \ldots, 0.9\}$ on 500 paraphrases disjoint from all evaluation probes, with settings shared across datasets and backbones. Details in Appendix A.4.

Table 1: Main comparison, CounterFact, LLaMA-2-7B, T=1000; bold: best.

Method	Rel ↑	Gen ↑	Loc ↑	ECE ↓	Param ↓
Fine-Tune	68.5	61.2	58.4	0.42	145×
ROME	82.1	71.4	79.8	0.34	0.97×
MEMIT	85.7	76.9	85.2	0.37	4.86×
GRACE	87.3	74.6	**94.1**	0.28	1.00×
MELO	88.9	80.3	90.5	0.22	0.86×
SoLA	89.6	81.1	91.0	0.26	0.79×
ELDER	90.5	82.2	91.6	0.19	1.00×
UAR-LoRA	**93.0**	**86.4**	92.0	**0.11**	0.44×

Table 2: Backbone transfer on CounterFact, T=1000: UAR-LoRA against ELDER on three backbones; bold: the better of the two.

Backbone	Method	Rel ↑	Gen ↑	Loc ↑
GPT-2-XL	ELDER	88.4	79.8	91.2
	UAR-LoRA	**90.2**	**83.5**	90.6
LLaMA-2-7B	ELDER	90.5	82.2	91.6
	UAR-LoRA	**93.0**	**86.4**	92.0
LLaMA-3-8B	ELDER	92.2	84.0	92.7
	UAR-LoRA	**95.6**	**88.1**	93.8

Backbones. Backbones are GPT-2-XL (Radford et al., 2019), LLaMA-2-7B (Touvron et al., 2023), and LLaMA-3-8B (Grattafiori et al., 2024). The backbone stays frozen and only the experts and the two heads are learned. The adapters act on the feed-forward sublayers, and the residual width is $d = 4096$ on both LLaMA backbones. More details are provided in Appendix A.9.2.

Protocol. Edits are streamed in a fixed order, and the model is queried after the last edit of a stream without resets. Each value is a single run (seed 0, greedy decoding, one A100 80GB GPU); unless stated otherwise, results use LLaMA-2-7B at $T = 1000$. Details in Appendices A.7 and A.10.

5.2 Main Results

Main comparison. This subsection tests whether meeting the three conditions at every edit improves accuracy and calibration at a lower parameter cost, and whether that lead holds across backbones, stream lengths, long-form targets and multi-hop composition. Table 1 compares UAR-LoRA with seven editors on CounterFact after one thousand edits. UAR-LoRA reaches 93.0 Reliability and 86.4 Generality against 90.5 and 82.2 for ELDER, the strongest baseline, at a parameter ratio of 0.44× against ELDER's 1.00×. The Generality margin (4.2 points) is larger than the Reliability margin (2.5), and the one quality column UAR-LoRA does not lead is Locality, where GRACE holds 94.1 against our 92.0 while its Generality (74.6) trails ours by nearly twelve points. GRACE isolates each edit inside a discrete codebook entry, so neighbours are protected but each paraphrase must fall within a stored key's deferral radius, whereas a continuous adapter lets the paraphrases of one edit share a subspace. The capacity sharing that costs UAR-LoRA two Locality points against a codebook is what buys its lead on paraphrases, and the entropy-sized ranks deliver that lead at less than half of ELDER's parameter ratio. More details about the setting, the parameter accounting and the error bounds of this comparison are provided in Appendices A.11.1 and A.11.8.

Calibration and accuracy. Table 1 also scores how far each editor's confidence can be trusted. UAR-LoRA has the lowest ECE, 0.11 against 0.19 for ELDER, a reduction of 42%. Calibration does not simply track editing accuracy: MEMIT is the worse calibrated of the ROME–MEMIT pair (0.37 against 0.34) despite the higher Reliability of the two, and the same reversal recurs between MELO and SoLA (0.22 against 0.26). A softmax router always distributes unit mass over the pool, so its score cannot express that no expert should serve a query; UAR-LoRA adds a separate scalar confidence head, Platt-scaled into a probability, and abstains below a threshold set on held-out data. Calibration has to be modelled by a head of its own rather than inherited from editing accuracy.

Backbone transfer. Table 2 repeats the comparison with ELDER on GPT-2-XL, LLaMA-2-7B and LLaMA-3-8B under the same CounterFact protocol. UAR-LoRA reaches a mean Reliability of 92.9 against 90.4 for ELDER (a 2.6-point difference before rounding) and a mean Generality of 86.0 against 82.0. The Reliability lead holds on every backbone (90.2 vs 88.4, 93.0 vs 90.5, 95.6 vs 92.2) and widens from 1.8 points on GPT-2-XL to 3.4 on LLaMA-3-8B, while the three Locality differences (-0.6, $+0.4$ and $+1.1$ points) change sign, so the methods tie on that axis. The configuration is held fixed across backbones and still transfers, because each module reads only the frozen base's own signals: the probe reads its predictive entropy, and both router heads read its hidden representation. Nothing is therefore tuned to a particular backbone. Details in Appendix A.11.2.

Table 3: Stream-length scaling on ZsRE, LLaMA-2-7B.

Method	T=1	T=100	T=1000	T=5000
Rel ↑				
MEMIT	92.1	87.3	81.5	76.9
GRACE	93.2	90.1	87.4	85.0
ELDER	94.0	91.4	88.0	86.2
UAR-LoRA	**96.5**	**95.4**	**94.6**	**94.1**
Loc ↑				
GRACE	**97.1**	**95.2**	**94.2**	**92.1**
ELDER	96.7	93.1	90.2	87.4
UAR-LoRA	96.9	93.4	91.6	89.9

Table 4: Component ablation on CounterFact, LLaMA-2-7B, T=1000: w/o rows remove one module, only rows keep one; bold: best.

Variant	Rel ↑	Gen ↑	Loc ↑	Param ↓
UAR-LoRA (full)	**93.0**	**86.4**	92.0	0.44
w/o Probe (r=32)	92.1	85.0	91.7	1.09
w/o Probe (r=8)	89.7	81.2	91.9	0.34
w/o Gate (soft-MoE)	91.4	84.2	88.9	0.44
w/o Ortho Init (random)	90.8	83.1	90.4	0.44
only Probe	90.9	82.7	91.0	0.44
only Gate	90.3	82.4	89.8	1.09
only Ortho Init	91.0	82.9	90.8	1.09

Stream-length scaling. Table 3 tracks ZsRE Reliability and Locality from one edit to five thousand, and it carries the lifelong claim. At $T = 5000$ UAR-LoRA retains 94.1 Reliability against 86.2 for ELDER: it loses 2.4 points from $T = 1$ against ELDER's 7.8, roughly a third. GRACE degrades least in Locality (97.1 to 92.1) but starts and ends below us in Reliability, and our own Locality (96.9 to 89.9) stays within about two points of GRACE's codebook. Up to the nominal orthogonal-capacity point $T^\star = \lfloor d/\mathbb{E}[r] \rfloor = 362$ of LLaMA-2-7B, with the measured mean rank $\mathbb{E}[r] = 11.3$, each new expert is placed orthogonally to the occupied span. The long streams lie past $T^\star$, where placement falls back to a truncated SVD and orthogonality holds only against the retained directions, so the flat curve beyond $T^\star$ is an empirical property of the complete system, not of orthogonal placement alone, and each module's separate contribution is measured only at $T = 1000$. Details in Appendices A.6 and A.11.3.

Long-form editing. Table 5 moves to WikiBio, whose targets are generated-text spans rather than single tokens. UAR-LoRA reaches 89.1 Reliability and 82.6 Generality against 85.9 and 77.8 for ELDER. The Generality margin (4.8 points) is again the larger, and GRACE again edges Locality (90.7 vs 89.7) at a Generality cost (72.9). As on single-token edits, a continuous adapter lets the paraphrases of one edit share a subspace, whereas a codebook entry serves only queries within its deferral radius. The CounterFact ordering carries over to long-form targets. Details in Appendix A.11.4.

Table 5: Long-form editing on WikiBio, LLaMA-2-7B, T=1000.

Method	Rel ↑	Gen ↑	Loc ↑
MEMIT	79.4	70.1	82.3
GRACE	83.6	72.9	**90.7**
MELO	84.2	76.0	87.1
ELDER	85.9	77.8	88.4
UAR-LoRA	**89.1**	**82.6**	89.7

Multi-hop composition. Table 6 scores edits whose answers must propagate through one, two and three reasoning hops. UAR-LoRA leads ELDER at every depth, from 88.4 against 85.1 at one hop to 48.6 against 41.2 at three. All methods decay with depth, but the margin over ELDER widens from 3.3 points at one hop to 6.6 at two and 7.4 at three. A correction that must reach entailed queries is the diffuse edit a single fixed rank under-provisions, so the widening margin is consistent with sizing each expert by the frozen base's uncertainty. Composition therefore widens the lead rather than eroding it. Details in Appendix A.11.5.

Table 6: Multi-hop composition, MQuAKE-style, LLaMA-2-7B, T=1000.

Method	1-hop ↑	2-hop ↑	3-hop ↑
MEMIT	80.2	52.4	33.1
GRACE	82.7	55.9	35.8
ELDER	85.1	61.3	41.2
UAR-LoRA	**88.4**	**67.9**	**48.6**

5.3 Ablation Study and Analysis

Component ablation. This subsection tests whether each module is needed for its condition and how the modules interact. Table 4 removes each module in turn from the full system, which reaches 93.0/86.4/92.0 at 0.44×. Pinning a fixed rank of 32 in place of the entropy probe (Challenge #1) gives 92.1 Reliability at 1.09×; the 0.9-point gap lies below its 1.17-point difference bound, so the probe matches it at 0.44×. A fixed rank of 8 cuts the budget to 0.34× but costs 3.3 Reliability points (89.7). Replacing the confidence gate with soft mixture-of-experts routing (Challenge #2) costs 3.1 Locality points (88.9), which ties Locality to hard abstention: a soft router always spreads unit mass over the pool, whereas an abstained query receives no adapter contribution. Replacing the orthogonal initialisation with a random one (Challenge #3) costs 2.2 Reliability points (90.8), consistent with orthogonal placement limiting how far a new edit overwrites earlier ones. Each module thus guards

its own condition: the probe holds accuracy at the lower budget, the gate holds Locality, and the initialiser holds the Reliability of stored edits. Details in Appendix A.11.6.

Interaction of the modules. Table 4 also keeps one module in each of three only-rows, none of which exceeds 91.0 Reliability, against 93.0 with all three. The losses do not add: removing the gate costs 1.6 Reliability points and the initialiser 2.2, yet removing both (only Probe) costs 2.1 rather than their sum of 3.8, so 1.7 points overlap. The initialiser's Locality effect changes sign across pairs (-1.6 against the full method, $+2.1$ between only Probe and w/o Gate), so we attribute none to it, and the gate alone holds the lowest single-module Locality (89.8 against 91.0 for the probe), so its Locality gain appears only with the other two. The rows separate capacity from routing: pinning the rank moves Reliability by 0.9 and 3.3 points but Locality by at most 0.3, so capacity drives Reliability, while the largest Locality shift follows the gate. The gate and the initialiser both limit how far a new expert disturbs stored edits, through which queries reach it and where it is written, and their overlapping losses are consistent with shared protection. Details in Appendix A.12.

Sensitivity to the gate threshold. Table 7 sweeps the confidence-gate threshold τ on the held-out calibration set. Calibration is tightest at $\tau = 0.7$ (ECE 0.11), where abstention decisions reach 86.3 and in-pool routing 92.6, whereas the abstention-maximising $\tau = 0.8$ reaches 88.1 on abstention but only 88.4 on routing. In-pool routing erodes steadily as the gate closes, from 95.8 at $\tau = 0.5$ to 79.2 at $\tau = 0.9$, and beyond $\tau = 0.8$ over-abstention on in-pool queries pulls abstention accuracy back to 84.0, so raising the threshold further erodes both. The threshold choice is a locality-reliability trade-off: a false accept admits an out-of-pool query and costs Locality, whereas a false abstain refuses an in-pool query and costs Reliability. Moving from 0.7 to 0.8 gains 1.8 abstention points for 4.2 routing points. Details in Appendix A.11.7.

Table 7: Confidence-gate threshold sweep on the held-out calibration set.

τ	0.5	0.6	0.7	0.8	0.9
ECE ↓	0.18	0.14	**0.11**	0.13	0.17
Abstain acc. ↑	71.2	79.5	86.3	**88.1**	84.0
Route acc. ↑	**95.8**	94.1	92.6	88.4	79.2

6 Related Work

Direct weight editing. Locate-then-edit methods find a fact in feed-forward projections (Dai et al., 2022) and rewrite it with a learned editor (Mitchell et al., 2022a) or in closed form, at one layer (Meng et al., 2022) or across layers (Meng et al., 2023). Writes land in shared weights, so a long stream perturbs earlier edits; AlphaEdit projects each update onto the null space of preserved knowledge (Fang et al., 2025). UAR-LoRA keeps the backbone frozen and gives each edit an adapter.

Adapter and LoRA editing. Low-rank editors store each correction beside the frozen backbone (Hu et al., 2022; Yu et al., 2024), and lifelong variants route among a growing pool, from ELDER's joint routing with deferral (Li et al., 2025) to SoLA's semantic scope (Luo et al., 2026) and WISE's sharded side memories (Wang et al., 2024a). Rank allocation (Zhang et al., 2023), adapter mixtures (Dou et al., 2024) and orthogonal subspaces (Wang et al., 2023; Xu et al., 2025) each supply one ingredient; UAR-LoRA sets all three for every expert.

Memory and codebook editors. A third family keeps corrections outside the parameters, in an episodic memory (Das et al., 2024), the prompt (Zheng et al., 2023), a codebook of edited activations (Hartvigsen et al., 2023), a counterfactual model behind a scope classifier (Mitchell et al., 2022b) or one patch per correction (Huang et al., 2023). These avoid parameter interference, but their store and scope decision grow with the stream; UAR-LoRA keeps corrections parametric in an orthogonally placed expert pool whose confidence gate abstains to the frozen base below a calibrated threshold.

7 Conclusion

UAR-LoRA treats capacity, admissibility and geometry as conditions of one heuristic capacity-coherence criterion, met as each expert enters the pool: the Entropy Probe and Adaptive Rank Selector sizes it, the Confidence-Gated Router abstains below a calibrated threshold, and the Rank-Aware Orthogonal Initialiser places it in unoccupied directions while any remain. No single-module variant exceeds 91.0 Reliability versus 93.0 with all three, and over 5000 ZsRE edits UAR-LoRA loses 2.4 points against ELDER's 7.8. This suggests fixing each expert's size, reach and placement together.

REFERENCES

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*, 2022.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurelie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiri Navratil, Soham Dan, and Pin-Yu Chen. Larimar: Large language models with episodic memory control. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10109–10126. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24a.html.

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*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-Style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.acl-long.106.

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 *International Conference on Learning Representations*, 2025.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

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*, 2022.

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*, 2023.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

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*, 2017.

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.

Haihua Luo, Xuming Ran, Tommi Kärkkäinen, Zhonghua Chen, Jiangrong Shen, Qi Xu, and Fengyu Cong. Reversible lifelong model editing via semantic routing-based LoRA. *arXiv preprint arXiv:2603.11239*, 2026.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, 2022b.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

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

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*, 2024a.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 2024b.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Haoyu Xu, Pengxiang Lan, Enneng Yang, Guibing Guo, Jianzhe Zhao, Linying Jiang, and Xingwei Wang. Knowledge decoupling via orthogonal projection for lifelong editing of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. doi: 10.18653/v1/2025.acl-long.646.

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*, 2024.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 2019.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.296.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

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, including the per-edit algorithm, the method constants and hyperparameters, the datasets and probe construction, the baseline configurations, the metric definitions and parameter accounting, the evaluation protocol and the compute environment, are disclosed in this appendix.

APPENDIX

Table S1 lists the contents of each part of the appendix.

Table S1: What each part of the appendix holds.

Section	Contents
Appendix A.0	Limitations
Appendix A.1	Extended related work
Appendix A.2	Motivation figure in detail
Appendix A.3	Editing criteria and parameter budget
Appendix A.4	Per-edit algorithm and method constants
Appendix A.5	Derivation of the capacity-coherence criterion
Appendix A.6	Failure cases and boundary conditions
Appendix A.7	Datasets and probe construction
Appendix A.8	Baseline configurations
Appendix A.9	Metric definitions and parameter accounting
Appendix A.10	Evaluation protocol, compute and randomness
Appendix A.11	Experiments in detail
Appendix A.12	Discussion

APPENDIX A.0: LIMITATIONS

Inference cost of the gated pool. Every query that reaches a deployed editor passes through its confidence gate and, when admitted, through a mixture over a pool of experts that grows with the stream, so the time an editor adds to each query matters as much as the parameters it trains. The evaluation does not include this inference latency against other lifelong editors, nor the way the edits of a stream spread over the ranks of the pool, which is what decides where the parameter saving comes from.

Sensitivity to the rank pool. The rank pool fixes the smallest and the largest capacity an edit can receive: a pool that is too narrow starves hard edits, and one that is too wide spends parameters on easy ones. The evaluation relates the default pool to Reliability only and does not include how Generality, which asks whether an edit carries over to paraphrases of its query, responds to narrower and wider pools, although a pool that suits exact recall may not be the best trade-off for transfer.

Editing beyond text. Deployed assistants increasingly answer questions about images as well as text, and a fact that changes in the world must then change in both, while the knowledge to be edited may sit partly in the visual encoder rather than in the language model. This work edits text-only language models and does not examine whether per-edit capacity, calibrated abstention and orthogonal placement carry over to multimodal models, whose edits may need a different notion of difficulty and of which queries an edit should reach.

APPENDIX A.1: EXTENDED RELATED WORK

Prior work on editing falls into three families, direct weight editing, adapter and LoRA editing, and memory and codebook editors; this appendix covers each in full, together with the continual-learning, calibration and benchmark work UAR-LoRA builds on.

A.1.1 DIRECT WEIGHT EDITING

A foundational line treats a fact as content stored in specific feed-forward projections (Dai et al., 2022) and edits it directly, either through a learned editor network (Mitchell et al., 2022a) or in closed form in one step. ROME locates the causal layer for a subject-relation pair and applies a rank-one update to overwrite the stored association (Meng et al., 2022), and MEMIT generalises this to batched updates across a span of layers (Meng et al., 2023). Both suit a small, static batch, but their in-place edits are difficult to isolate as a stream accumulates: each write perturbs the weights the previous writes touched. AlphaEdit answers this within the same family by projecting each perturbation onto the null space of preserved knowledge (Fang et al., 2025). UAR-LoRA leaves the frozen backbone untouched.

One further branch is the meta-learned editor, which trains a hypernetwork or auxiliary model to emit the weight change and so decouples the edit from a hand-chosen layer (De Cao et al., 2021). It sits with the locate-then-edit family for our purposes: the edit is still written into a shared parameter tensor, so a long sequence accumulates the same cross-edit interference, and the choice of where to write is automated rather than removed. SERAC (Mitchell et al., 2022b) is sometimes grouped with these meta-learned methods because it also trains an auxiliary model, but it belongs to the memory family: it leaves the base parameters untouched and routes in-scope queries to a separate counterfactual model, so it carries the scope-classifier cost of that family rather than the cross-edit interference of this one.

A.1.2 ADAPTER AND LORA EDITING

A second family expresses each correction as a small trainable module while freezing the base model. Adapters insert bottleneck layers (Houlsby et al., 2019) and low-rank adaptation reparameterises the update as a product of two thin matrices (Hu et al., 2022), with later work allocating rank by importance (Zhang et al., 2023) or routing among several adapters (Dou et al., 2024). MELO attaches low-rank blocks keyed to edited facts (Yu et al., 2024), and recent lifelong editors route among a growing pool of adapters so that early edits survive later ones, with ELDER learning the routing jointly (Li et al., 2025) and SoLA routing by semantic scope (Luo et al., 2026). WISE routes each query between the main memory and a side memory whose edits are sharded by batch into distinct parameter subspaces and then merged (Wang et al., 2024a), and KDE projects the gradient of each edit onto the space orthogonal to cached bases of past edits (Xu et al., 2025). UAR-LoRA belongs to this family but differs in three specifics: rank adapts to edit difficulty, initialisation is orthogonal rather than random, in the spirit of the orthogonal subspaces used for continual learning (Wang et al., 2023), and the router abstains to the base at a threshold set on held-out calibration data.

The low-rank and adapter machinery we build on is standard (Hu et al., 2022; Houlsby et al., 2019; Dettmers et al., 2023), the confidence-gated router reuses the soft mixture-of-experts for-

mulation (Puigcerver et al., 2024), the score calibration follows Platt scaling (Platt et al., 1999), and the forgetting template of Appendix A.5 follows McAllester (McAllester, 1999). The contribution is the coupling of the three conditions.

A parallel line in continual learning limits forgetting by constraining the directions in which updates may move the parameters. Gradient-projection and subspace methods restrict each task's update to directions orthogonal to the span occupied by earlier tasks, so that new learning cannot disturb the representations already committed: OGD projects each gradient away from the directions of previously stored gradients (Farajtabar et al., 2020), OWM confines updates to the orthogonal complement of the input space seen so far (Zeng et al., 2019), GPM builds a memory of bases spanning the activations of earlier tasks and projects out their component (Saha et al., 2021), and Adam-NSCL steers updates into the null space of the feature covariance of previous tasks (Wang et al., 2021). Separately from that group, O-LoRA carries the orthogonality on low-rank adapters rather than on raw gradients, learning each task in a low-rank vector subspace kept orthogonal to those of the earlier tasks (Wang et al., 2023). It holds that orthogonality as a penalty throughout training and gives every task the same rank, whereas our initialiser pays the projection once, at placement, and lets the rank follow the difficulty of the single edit at hand. Our orthogonal initialiser shares the premise that subspace separation limits interference, but realises it through the placement of a low-rank adapter (Hu et al., 2022) or bottleneck module (Houlsby et al., 2019) rather than through a gradient constraint. The distinction is where orthogonality is enforced: projection methods let a full update form and then remove its component in the occupied span at every step, whereas UAR-LoRA initialises each new expert directly into the complement of that span and leaves fitting under only a soft penalty thereafter. Those gradient-projection methods apply a projection at every update step, with the subspace bases themselves recomputed once per task, while all tasks continue to share one parameter tensor; UAR-LoRA pays the full projection once at placement and places each edit in its own block, which is what turns the separation into the additive direct-sum accounting the capacity-coherence criterion relies on.

A.1.3 Memory and Codebook Editors

A third family keeps corrections in an external store consulted at inference, whether an episodic memory (Das et al., 2024) or the prompt itself (Zheng et al., 2023). GRACE maintains a discrete codebook of edited activations growing with the stream (Hartvigsen et al., 2023), SERAC routes queries to a counterfactual model through a learned scope classifier (Mitchell et al., 2022b), and continual approaches add one lightweight module per correction (Huang et al., 2023). These sidestep parameter interference, at the cost of a scope classifier whose errors can grow with the store. UAR-LoRA keeps corrections parametric and outside the backbone, replacing that store with an orthogonally structured, abstaining pool.

A.1.4 Uncertainty and Calibration

Our entropy probe and confidence gate build on classical tools for turning model scores into calibrated decisions: Platt scaling maps raw scores to probabilities (Platt et al., 1999), and PAC-Bayes analysis bounds the generalisation of stochastic predictors (McAllester, 1999), whose functional form we borrow as a heuristic template (Appendix A.5). We use uncertainty both to size each edit and to license abstention, connecting calibration to the capacity-coherence criterion.

A.1.5 Sequential and Lifelong Editing Benchmarks

Editing benchmarks now include sequential and compositional tests. ZsRE frames editing as zero-shot relation extraction (Levy et al., 2017), the biographies of WikiBio probe locality on generated text (Lebret et al., 2016), and MQuAKE measures whether multi-hop entailments update after an edit (Zhong et al., 2023); a recent survey catalogues the metric landscape and its trade-offs (Wang et al., 2024b). We evaluate at streams up to $T = 5000$, the regime this paper is concerned with.

A.1.6 Positioning Summary

Memory and routing methods manage accumulation by moving state outside the backbone but leave adapter capacity and routing confidence as separate engineering choices; calibration methods supply

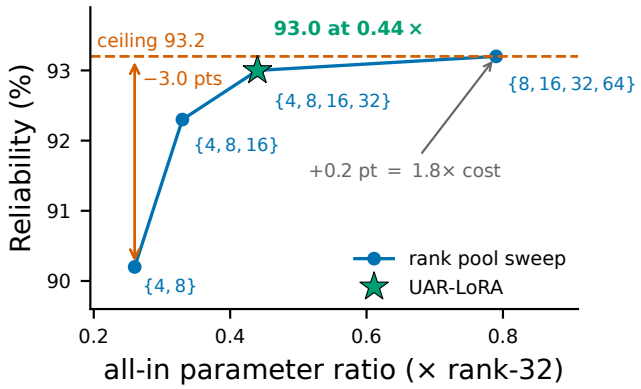

Figure S1: The diagnostic for Challenge #1 in the motivation figure: reliability against parameter cost as the rank pool widens.

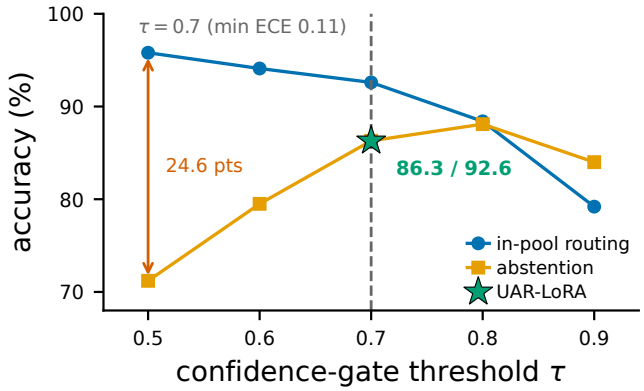

Figure S2: The diagnostic for Challenge #2 in the motivation figure: the two accuracies the confidence gate governs, swept over the threshold.

the confidence signal but are not tied to how capacity is allocated. What distinguishes UAR-LoRA is that it addresses the three failure modes together: a per-edit rank read from the frozen base's predictive entropy and an abstention decision licensed by a learned confidence head, two uncertainty signals of different origin that serve two conditions of one criterion, together with an orthogonal placement that keeps that capacity from colliding with earlier edits at placement.

APPENDIX A.2: MOTIVATION FIGURE IN DETAIL

The motivation figure compresses three diagnostics into one row, one per challenge. Each panel is reproduced below with its axes and the provenance of every annotated value.

Challenge #1, capacity misallocation. The horizontal axis is the all-in trainable-parameter ratio against a fixed rank-32 LoRA reference, counting adapters together with the router, confidence head, entropy probe and rank selector; the vertical axis is mean Reliability in percent. Both coordinates are measured on CounterFact with LLaMA-2-7B at T=1000. The four blue markers are the four rank pools swept, annotated with the pool itself: $\{4, 8\}$ at $0.26\times$ and 90.2, $\{4, 8, 16\}$ at $0.33\times$ and 92.3, $\{4, 8, 16, 32\}$ at $0.44\times$ and 93.0, and $\{8, 16, 32, 64\}$ at $0.79\times$ and 93.2. The dashed horizontal line marks the widest pool actually swept, 93.2, which is a ceiling within this sweep. The vermilion double arrow is the 3.0-point deficit a fixed small pool carries against that ceiling. The grey annotation reads the last two markers together: the final 0.2 point of Reliability costs $0.79/0.44 = 1.8\times$ the parameters. The green star is the pool UAR-LoRA adopts, which reaches 93.0 at $0.44\times$ against 93.2 at $0.79\times$.

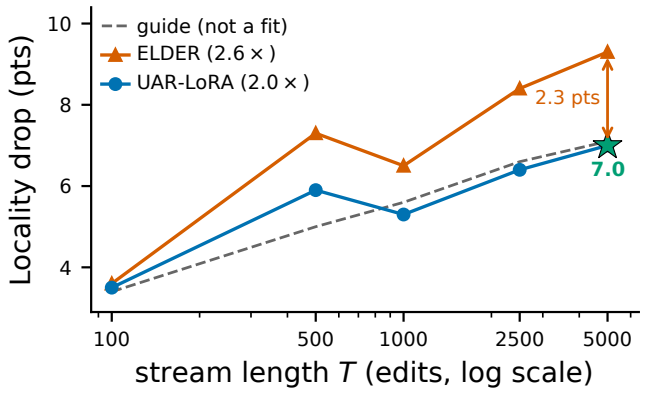

Figure S3: The diagnostic for Challenge #3 in the motivation figure: cumulative locality damage against stream length.

Challenge #2, uncalibrated routing. The horizontal axis is the confidence-gate threshold τ, swept over $\{0.5, 0.6, 0.7, 0.8, 0.9\}$; the vertical axis is accuracy in percent. Both curves and the calibration-set ECE quoted below come from the confidence-gate threshold sweep of Appendix A.11.7. The blue curve is in-pool routing accuracy, which falls monotonically from 95.8 at τ=0.5 to 79.2 at τ=0.9 as the gate closes. The orange curve is the abstention decision accuracy over the two classes defined in Appendix A.11.7, which rises from 71.2 to a maximum of 88.1 at τ=0.8 and falls back to 84.0 at τ=0.9. The vermilion double arrow at the left edge is the 24.6-point spread between the two curves at τ=0.5, the loosest gate in the sweep. The dashed vertical line is the operating point τ=0.7, chosen because calibration-set ECE is smallest there (0.11); the green star reports the resulting pair, 86.3 abstention and 92.6 in-pool. The threshold is selected on held-out calibration data alone.

Challenge #3, interference accumulation. The horizontal axis is stream length T on a logarithmic scale at the five measured points $\{100, 500, 1000, 2500, 5000\}$; the vertical axis is cumulative Locality drop in points, so lower is better. Both series are measured on ZsRE with LLaMA-2-7B, and each Locality reading has denominator $N_{\text{Loc}} = 10{,}000$. ELDER rises from 3.6 to 9.3, a factor of 2.6×; UAR-LoRA rises from 3.5 to 7.0, a factor of 2.0×. Neither curve is monotonic, and both dip at T=1000. The vermilion double arrow at the right edge is the 2.3-point separation at T=5000. The dashed grey line is a guide sequence of slow, sub-linear shape drawn on the same axes. The contrast is between two complete systems, so the panel shows the two systems' total separation rather than isolating orthogonal placement; the isolated effect is measured in Appendix A.11.6.

APPENDIX A.3: EDITING CRITERIA AND PARAMETER BUDGET

An edit enters as a query-target pair, and the entropy probe reads its difficulty and selects a rank from the pool. A low-rank adapter is then placed in an unused subspace and fitted. At inference the confidence gate decides whether the query is close enough to an applied edit to be routed at all.

A.3.1 THE THREE EDITING CRITERIA

The lifelong objective is stated over three canonical editing criteria. Reliability is the edit accuracy on each edit query x_t, that is, whether the requested change is recalled when the edit query itself is presented. Generality is the accuracy on paraphrases of x_t, so that a change survives rewordings rather than being memorised at a single surface form. Locality is the edited model's agreement with f_{θ_0} on neighbouring inputs, that is, the preservation of behaviour on material the edit was never meant to touch. The measurement protocol by which each criterion is scored, including probe construction and split sizes, is given in Appendices A.7 and A.9.

A.3.2 THE PARAMETER BUDGET IN FULL

The parameter budget constrains the ranks of the stream by

$$\sum_{t=1}^{T} r_t \left(d_{\text{in}} + d_{\text{out}}\right) \;\le\; r_{\text{ref}} \left(d_{\text{in}} + d_{\text{out}}\right) T, \tag{S1}$$

where d_{in} and d_{out} are the input and output widths of an adapted projection and $r_{\text{ref}} = 32$ is the reference rank. Equation (S1) is an admissibility ceiling rather than a discriminative criterion: it is equivalent to $\frac{1}{T}\sum_{t\le T} r_t \le 32$, so the fixed rank-32 baseline meets it with equality and every policy drawn from the default pool $R = \{4, 8, 16, 32\}$ meets it, with equality only where that policy selects the largest rank in the pool at every edit. It therefore says which allocations are admissible, not which is preferable; the choice among admissible allocations is made on the accuracy and parameter-ratio columns of the experimental tables, where spending fewer parameters on easy edits so that hard edits can be served at full rank is what the measured ratios reward. The two parameter accountings in this paper are stated at different scopes: the direct-sum bookkeeping of the method is per adapted projection and in the input width d of that projection, whereas the reference footprint against which the Param ratios are normalised is the whole-model placement of Appendix A.9.2. Eq. (S1) is unaffected by the difference, since it reduces to a constraint on the mean selected rank and the Param column is a ratio taken against that same whole-model footprint on both sides. Every configuration reported in this paper satisfies it.

APPENDIX A.4: PER-EDIT ALGORITHM AND METHOD CONSTANTS

A.4.1 PER-EDIT ALGORITHM

Each incoming edit is handled in four steps.

1. **Probe.** The entropy probe reads the backbone's predictive entropy on the edit query and maps it, through the rolling normalisation window, to a normalised difficulty score. This score selects one rank r from the pool R: low-entropy easy edits receive small ranks, high-entropy hard edits receive larger ones.
2. **Place.** A LoRA adapter of the selected rank r is initialised orthogonally to the span already occupied by previous adapters, so that the new edit is written into an unused subspace wherever one remains.
3. **Fit.** The adapter is trained with AdamW (steps, learning rate, and the confidence and orthogonality loss weights λ_1, λ_2 as listed in Table S2), minimising the edit objective together with a confidence loss and an orthogonality regulariser that discourages overlap with existing experts.
4. **Gate.** At inference, the confidence gate compares the routing confidence against the calibrated threshold τ listed in Table S2. Queries below threshold are abstained (routed to the unedited backbone), which is what protects locality when a query falls outside the pool of applied edits.

The four steps map one-to-one onto the three modules and their fitting. In the Probe step the raw predictive entropy is not used directly but passed through the rolling normalisation window, so that difficulty is judged relative to the recent stream rather than on an absolute entropy scale that would drift across datasets. In the Gate step the confidence head w_c is trained with the adapters inside the Fit step, on the edit stream, whereas the Platt scaler and the threshold τ are fit afterwards on the calibration set alone.

Algorithm 1 gives a schematic view of UAR-LoRA as one procedure: the per-edit installation of an expert, the calibration of the gate, and the decision the gate makes on every query. Stage 1 sizes the edit, reading the frozen-base predictive entropy, re-estimating it over sub-prompt windows on short prompts, and normalising it against the trailing window before bucketing it into one rank. Stage 2 spends that rank where no earlier expert has written, projecting a random candidate onto the orthogonal complement of the occupied span by Eq. (S11) and falling back to a truncated projector once the occupied span nears saturation, at which point the placement is orthogonal only to the

Algorithm 1 UAR-LoRA: per-edit installation, gate calibration, and gated inference

Require: frozen base f_{θ_0} with hidden state $h(\cdot)$; edit stream $\{e_t = (x_t, y_t)\}_{t=1}^{T}$; rank pool R; the fixed configuration of Table S2
Ensure: expert pool E_T, routing heads (W_s, w_c), calibrated confidence c_ϕ, threshold τ
$E_0 \leftarrow \varnothing$; initialise W_s and w_c; θ_0 stays frozen
for $t = 1$ to T **do**
  Stage 1: Entropy Probe and Adaptive Rank Selector
  Run f_{θ_0} once on x_t; read $p_{\theta_0}(\cdot \mid x_t)$ and $h(x_t)$ ▷ frozen forward
  $H_t \leftarrow -\sum_{y \in V} p_{\theta_0}(y \mid x_t) \log p_{\theta_0}(y \mid x_t)$ ▷ predictive entropy
  if x_t has fewer effective tokens than the short-prompt threshold **then**
    $H_t \leftarrow$ mean of this entropy over the 5 sub-prompt windows at uniform stride
  end if
  if $t \le t_0$ **or** $\max \mathcal{H}_t - \min \mathcal{H}_t = 0$ **then**
    $d_t \leftarrow d_0$ ▷ cold start; vanishing range
  else
    $d_t \leftarrow q(H_t)$ on $\mathcal{H}_t$, the whole prefix while $t < W$ ▷ Eq. (S2)
  end if
  $r_t \leftarrow R\big[\min(\lfloor d_t\,|R| \rfloor, |R| - 1)\big]$ ▷ rank bucketing
  if H_t was re-estimated and is still flagged unreliable **then**
    $r_t \leftarrow r_{\mathrm{fb}}$ ▷ clamped fallback rank
  end if
  Stage 2: Rank-Aware Orthogonal Initialiser
  $\tilde{A}_t \sim \mathcal{N}(0, I)$ in $\mathbb{R}^{r_t \times d}$
  $A_{\mathrm{ex}} \leftarrow [A_1^\top, \ldots, A_{t-1}^\top]$
  if the condition number of $A_{\mathrm{ex}}^\top A_{\mathrm{ex}}$ exceeds the threshold of Table S2 **or** A_{ex} has at least $0.95\,d$ columns **then**
    $P_{\mathrm{ex}} \leftarrow$ projector onto the truncated SVD of A_{ex} at $\min\big(\mathrm{rank}(A_{\mathrm{ex}}), \lfloor 0.95\,d \rfloor\big)$ ▷ near-saturation guard, Appendix A.6
  else
    $P_{\mathrm{ex}} \leftarrow A_{\mathrm{ex}} A_{\mathrm{ex}}^{+}$ ▷ Eq. (S12); Proposition 1 gives this general form
  end if
  $A_t \leftarrow \tilde{A}_t\big(I - P_{\mathrm{ex}}\big)$; $B_t \leftarrow 0$ ▷ Eq. (S11); exactly orthogonal unless the guard fired
  Stage 3: Confidence-Gated Router, fitted with the expert
  Take x_t as the positive anchor and draw x_{neg} from the edits not yet installed ▷ anchors of $\mathrm{BCE}_{\mathrm{bal}}$
  for S AdamW steps **do**
    $\mathcal{L}_{\mathrm{total}} \leftarrow \mathrm{CE}_t + \lambda_1\,\mathrm{BCE}_{\mathrm{bal}} + \lambda_2 \|A_t A_{\mathrm{ex}}\|_F^2$ ▷ joint objective
    Update $\{A_t, B_t, W_s, w_c\}$ at rate η; earlier experts stay fixed ▷ the router and the confidence head supply $s(\cdot)$ and $c_\phi(\cdot)$
  end for
  $E_t \leftarrow E_{t-1} \cup \{(x_t, y_t, A_t, B_t, r_t)\}$
end for
Stage 4: Gate calibration, on held-out data alone
Fit the Platt scaler of c_ϕ on the held-out paraphrases of Appendix A.11.7
$\tau \leftarrow$ the grid value of Table S2 minimising calibration-set ECE; freeze the scaler and τ
Stage 5: Gated inference, once per query
for each query x **do**
  if $c_\phi(x) \ge \tau$ **then**
    serve x with $W_0 u + \sum_i s_i(x)\, B_i A_i\, u$ on every adapted projection ▷ Eq. (S7), admit
  else
    serve x with the frozen base $f_{\theta_0}(x)$ ▷ Eq. (S7), abstain
  end if
end for
return E_T, (W_s, w_c), the calibrated c_ϕ, and τ

retained directions. Stage 3 fits the adapter and the two heads together against the joint objective, whose three terms own the edit fit, the confidence calibration, and the residual overlap. Stages 1 to 3 run once per edit and are never rolled back. Stage 4 runs once afterwards, fitting the Platt scaler and the threshold on held-out data alone. Stage 5 is consulted on every query and is a hard threshold: the expert mixture of Eq. (S7) when the calibrated confidence clears it, the untouched base otherwise. The constants the listing names in words are those of Table S2.

Table S2: Full hyperparameter configuration.

Hyperparameter	Value
Rank pool R	$\{4, 8, 16, 32\}$
Confidence-gate threshold τ	calibration-set ECE minimiser on the sweep grid
Loss weight λ_1 (confidence)	0.5
Loss weight λ_2 (orthogonality)	0.1
Optimiser	AdamW
Learning rate	$1e{-3}$
Steps per edit	50
Batch size (per edit)	1
Rolling normalisation window	1000 (whole prefix before it fills)
Cold-start length	100
Cold-start fallback rank	$r = 16$ (from $d_t = 0.5$)
Platt calibration set size	500 held-out paraphrases
Confidence-threshold sweep grid	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$
Short-prompt threshold	8 effective tokens
Entropy fallback (short prompts)	mean over 5 sub-prompt windows, uniform stride
Entropy fallback clamp rank	$r = 16$
Ill-conditioning threshold	10^8 on the condition number of $A_{\text{ex}}^\top A_{\text{ex}}$
Near-saturation guard	truncated SVD clamped at $\lfloor 0.95\, d \rfloor$
Projector ridge	10^{-6}
Negatives per edit	1, drawn from edits not yet installed
Loss-explosion guard	raise at $5\times$ the median of the trailing 50 losses
ECE bins M	10

A.4.2 HYPERPARAMETERS

Table S2 lists the full configuration. All numbers are held fixed across datasets and backbones unless a table explicitly sweeps one of them.

Two degenerate cases are handled explicitly by the normaliser. Before the window fills, that is for $t_0 < t < W$, it reads the whole available prefix rather than a fixed-width window. If the window range $\max \mathcal{H}_t - \min \mathcal{H}_t$ vanishes, the difficulty falls back to the cold-start value $d_t = d_0$.

The difficulty normaliser. Explicitly, the normaliser rescales the current entropy against the range of the entropies that preceded it,

$$q(H_t) = \min\Big(1, \max\Big(0, \frac{H_t - \min \mathcal{H}_t}{\max \mathcal{H}_t - \min \mathcal{H}_t}\Big)\Big), \tag{S2}$$

where the window it reads is $\mathcal{H}_t = \{H_{t'}\}_{\max(1,\, t-W) \le t' \le t-1}$, that is, the entropies of the W edits immediately before t. The outer clipping in Eq. (S2) is what keeps $d_t \in [0, 1]$: the window excludes H_t itself, so a new extreme would otherwise place the raw ratio outside the unit interval. The behaviour of the window before it fills and when its range vanishes is the pair of degenerate cases stated above.

The edit loss in full. The term CE_t of the joint objective in Algorithm 1 is the standard cross-entropy edit loss on the target, that is, the negative log-likelihood the edited model assigns to y_t given x_t, written token-wise over the target positions j,

$$\begin{aligned} \mathrm{CE}_t &= -\log p_{\theta_t}\big(y_t \mid x_t\big) \\ &= -\sum_j \log p_{\theta_t}\big(y_{t,j} \mid x_t,\, y_{t,<j}\big), \end{aligned} \tag{S3}$$

where p_{θ_t} is the next-token distribution of the edited model after edit t, as opposed to the frozen-base distribution p_{θ_0} read by the entropy probe.

APPENDIX A.5: DERIVATION OF THE CAPACITY-COHERENCE CRITERION

The capacity-coherence criterion below is heuristic, as is the forgetting template given below.

The capacity-coherence criterion. The relation below is stated under five assumptions. (a) Entropy-normalised ranks $r_t = R[\min(\lfloor q(H_t)\,|R|\rfloor, |R|-1)]$. (b) Calibrated abstention $\Pr(c_\phi(x) \geq \tau \mid x \notin E_T) \leq \delta$. (c) Exact pairwise orthogonality $A_t A_s^\top = 0$ for $s < t$; we treat this as an idealisation that the orthogonal initialiser and the soft penalty encourage but do not guarantee; its empirical counterpart is the accuracy cost of replacing the orthogonal initialiser with a random one (Appendix A.11.6). (d) The induced rank sequence approximately stationary with a well-defined mean $\mathbb{E}[r]$; this is a modelling assumption, and $\mathbb{E}[r]$ is read from the observed rank distribution, whose mean is 11.3 on LLaMA-2-7B (Appendix A.6). (e) Each A_t of full row rank r_t (almost surely under the Gaussian initialiser), with the realized cumulative rank at most d over the horizon; the expected budget makes this plausible but does not by itself guarantee it. Writing $T^\star = \lfloor d/\mathbb{E}[r] \rfloor$, then under (a)–(e) we have, for all $T \leq T^\star$,

$$\mathbb{E}\left[\sum_{t\leq T} r_t\right] \;\leq\; \mathbb{E}[r]\,T \;\leq\; d, \tag{S4}$$

so the pool stays within the ambient width in expectation over the horizon. Under the same assumptions the second statement of the criterion reads

$$\mathrm{Interf}(T) \;\leq\; \delta + \frac{\sum_{t\leq T} r_t}{d}, \tag{S5}$$

where the interference score $\mathrm{Interf}(T)$ is the sum of the out-of-pool calibration miss-probability and the expected squared geometric overlap of an isotropic in-domain query with the occupied span. This is a heuristic modelling relation, not a proven bound. Its inequalities are stated to hold approximately under the assumptions above, and their empirical counterparts are those noted in (c) and (d): the ablations and the measured rank distribution.

A.5.1 Derivation of the capacity-coherence criterion

We derive the two claims separately: first the expected budget inequality $\mathbb{E}[\sum_{t\leq T} r_t] \leq \mathbb{E}[r]\,T \leq d$, then the interference-score inequality $\mathrm{Interf}(T) \leq \delta + \sum_{t\leq T} r_t/d$, where $\mathrm{Interf}(T)$ is the calibration miss-probability plus the expected geometric overlap, not a single event probability.

Budget inequality. By condition (a) each rank is $r_t = R[\min(\lfloor q(H_t)\,|R|\rfloor, |R|-1)]$ with $R = \{4, 8, 16, 32\}$. The rolling range-normaliser $q(\cdot)$ maps the current entropy, against the trailing window of entropies, to the unit interval (Eq. (S2)); it is the subsequent bucketing map of condition (a) that turns that normalised difficulty into one of the $|R|$ ranks. We model the induced rank sequence as approximately stationary with a well-defined mean $\mathbb{E}[r]$ (a modelling assumption, consistent with the observed rank distribution rather than a proven property of the non-stationary stream); under this assumption, by linearity of expectation applied to the partial sum,

$$\mathbb{E}\left[\sum_{t\leq T} r_t\right] = \sum_{t\leq T} \mathbb{E}[r_t] = \mathbb{E}[r]\,T. \tag{S6}$$

For $T \leq T^\star = \lfloor d/\mathbb{E}[r] \rfloor$ this gives $\mathbb{E}[r]\,T \leq \mathbb{E}[r]\,\lfloor d/\mathbb{E}[r] \rfloor \leq d$, so the expected cumulative rank $\mathbb{E}[\sum_{t\leq T} r_t] = \mathbb{E}[r]\,T$ is at most d. We therefore state the budget in expectation: $\mathbb{E}[\sum_{t\leq T} r_t] \leq \mathbb{E}[r]\,T \leq d$. The bound is on the expectation only: the realised sum $\sum_{t\leq T} r_t$ need not satisfy $\sum_{t\leq T} r_t \leq d$ deterministically. The orthogonal placement below is therefore stated under the assumption that a free orthogonal complement is available, that is, while the realised occupied rank $\sum_{s\leq t} r_s$ stays below d. That assumption is a condition on the realised occupied rank rather than on the edit count, and the cold-start schedule of Appendix A.4.2, which places the first t_0 edits at the cold-start rank, reaches the ceiling earlier than the nominal $T^\star$ suggests; past that point the near-saturation guard of Appendix A.6.1 switches the placement projector to a truncated singular value decomposition so that it stays well-conditioned.

Geometric interference term. By condition (c), $A_t A_s^\top = 0$ for all $s < t$, so the row spaces of the experts are pairwise orthogonal and their union is a direct sum. By Proposition 2 the dimensions then add without overlap over the T experts. Its total dimension is therefore $\dim\big(\bigoplus_{t\leq T} \mathrm{row}(A_t)\big) =$

$\sum_{t\le T}\dim\big(\mathrm{row}(A_t)\big)=\sum_{t\le T} r_t$ (each A_t has full row rank r_t almost surely under the Gaussian initialiser; Proposition 2); while this realized sum stays below d, the occupied subspace $\mathcal{S}_T$ is a genuine proper subspace of $\mathbb{R}^d$ whenever $\sum_{t\le T} r_t < d$. Consider an in-domain query whose normalised input direction u at the adapted projection is isotropic on the unit sphere. Edit interference, meaning that an admitted query is captured by the occupied subspace and served by the experts rather than left to the frozen base, occurs to the extent that u carries mass in $\mathcal{S}_T$. The geometric content we use is the expected squared projection of u onto $\mathcal{S}_T$, not a discrete hitting probability: an isotropic continuous direction lies in a proper subspace with probability zero, so the relevant quantity is the average overlap. For an isotropic direction the expected squared projection onto a subspace of dimension k is exactly k/d, since $\mathbb{E}\big[\|\Pi_{\mathcal{S}_T}u\|^2\big]=\mathrm{tr}(\Pi_{\mathcal{S}_T})/d=\dim(\mathcal{S}_T)/d$, where $\Pi_{\mathcal{S}_T}$ is the orthogonal projector onto $\mathcal{S}_T$ (Proposition 3). We therefore quantify the geometric interference term by this expected squared overlap, $\mathbb{E}\big[\|\Pi_{\mathcal{S}_T}u\|^2\big]=\dim(\mathcal{S}_T)/d=\sum_{t\le T} r_t/d$. The dimension count alone holds for any placement as an upper bound, since $\dim(\mathcal{S}_T)\le\sum_{t\le T} r_t$ whether or not the row spaces overlap; condition (c) turns it into an equality. What (c) adds beyond the count is the zero-map property: $A_tA_s^\top=0$ gives $A_tv=0$ for every $v\in\mathrm{row}(A_s)$ with $s<t$, so at placement a new expert produces no update on the input directions that earlier experts read.

Calibration term. The second failure mode is a miss-route on an out-of-pool query, that is, a query $x\notin E_T$ that the gate wrongly admits. The gate is a hard threshold: on an adapted projection with frozen weight W_0 and input u it serves

$$g_x(u)=\begin{cases} W_0u+\displaystyle\sum_i s_i(x)\,B_iA_i\,u, & c_\phi(x)\ge\tau,\\ W_0u, & c_\phi(x)<\tau,\end{cases}\tag{S7}$$

where $s(x)$ is the router distribution over the stored experts. By this gate such a query is served by an expert only when $c_\phi(x)\ge\tau$. Condition (b) states precisely that $\Pr(c_\phi(x)\ge\tau\mid x\notin E_T)\le\delta$, so the probability of an out-of-pool miss-route is at most δ.

Combining the two terms. The two failure modes are distinct: the geometric term is the expected overlap of in-domain queries with the occupied span, and the calibration term is the probability that an out-of-pool query is wrongly admitted. We define the interference score as the sum of these two contributions. The calibration contribution is a genuine probability, at most δ by condition (b); the geometric contribution is the expected squared overlap $\sum_{t\le T} r_t/d$ established above. Adding them,

$$\text{interference score}\ \le\ \delta+\frac{\sum_{t\le T} r_t}{d},\tag{S8}$$

which is the interference-score relation stated at the head of this appendix.

A forgetting template. Using the capacity-coherence criterion, we compare the average locality drop Forget_T, for a prior P and posterior Q_T over the expert pool and any $\zeta\in(0,1)$, against the McAllester PAC-Bayes template

$$\mathbb{E}_{Q_T}[\ell]\ \lesssim\ \widehat{\mathbb{E}}_{Q_T}[\ell]+\sqrt{\frac{\mathrm{KL}(Q_T\|P)+\log\big(2\sqrt{T}/\zeta\big)}{2(T-1)}}.\tag{S9}$$

We use this display only as a descriptive template, not a proven bound; the PAC-Bayes hypotheses fail for Forget_T (detailed in the analysis below), so the template is read as a qualitative rationale for why excess forgetting should decrease with stream length rather than as a rate.

A.5.2 Heuristic analysis of the forgetting template

We apply a McAllester PAC-Bayes inequality (McAllester, 1999) to the per-edit forgetting indicator.

Forgetting indicator and PAC-Bayes. For each earlier edit $s<T$ define the forgetting indicator $\ell_s=\mathbf{1}[\,\text{the edited model after edit } T \text{ disagrees with } f_{\theta_0} \text{ on the locality neighbourhood of } x_s\,]$, a bounded loss in $[0,1]$. The average locality drop is $\mathrm{Forget}_T=\frac{1}{T-1}\sum_{s<T}\mathbb{E}_{Q_T}[\ell_s]$, the posterior-averaged empirical forgetting over the $T-1$ prior edits. The McAllester bound states that for any

prior P independent of the sample, any posterior Q_T, and any $\zeta \in (0,1)$, with probability at least $1-\zeta$,

$$\mathbb{E}_{Q_T}[\ell] - \widehat{\mathbb{E}}_{Q_T}[\ell] \;\le\; \sqrt{\frac{\mathrm{KL}(Q_T\|P) + \log\big(2\sqrt{T}/\zeta\big)}{2(T-1)}}. \tag{S10}$$

We caution that a rigorous PAC-Bayes application requires a fixed sampling distribution from which the forgetting losses are drawn i.i.d. and a prior independent of that sample, neither of which is cleanly defined for Forget_T: the quantity is the empirical forgetting over the $T-1$ prior edits of a single non-stationary stream, not a population loss under a fixed distribution. We read it as a descriptive template suggesting that an "excess" forgetting quantity beyond the empirical level should be governed by a complexity term that shrinks with stream length. The measured forgetting itself does not decrease: its cumulative level rises overall with T, and what we treat as the empirical observation is the slow, sub-linear shape of that rise, as shown in Figure S3.

Why the complexity term should stay controlled (heuristic). The intuition behind the template is that the posterior Q_T over the expert pool departs from the prior P only on the directions actually occupied by installed experts. By the orthogonal initialisation of Eq. (S11) (Propositions 1 and 5), each expert is placed in a fresh block of the direct sum, and the installation step by itself adds no new term to the mixture, since the freshly installed expert contributes the zero map at that instant (Proposition 4). While the realised occupied rank stays below d, the departure is therefore confined to a subspace of bounded dimension. We stress that confinement to $\le d$ coordinates does not yield a numerical bound on $\mathrm{KL}(Q_T\|P)$: a KL divergence depends on the displacement and variance of the posterior on the occupied coordinates, which support alone does not constrain. We therefore make no claim of the form $\mathrm{KL}(Q_T\|P) = O(d)$ or of a per-coordinate KL bounded by r_t/d. The upshot is qualitative. To the extent the posterior's movement stays confined to a bounded occupied span, one expects the excess-forgetting template to decrease with stream length. The slow, sub-linear growth of the measured forgetting we report (its cumulative level rises, by a factor 2.0 over a fifty-fold longer stream) is an empirical observation, shown in Figure S3, for which this heuristic offers a rationale rather than a proof.

A.5.3 Supporting propositions

The propositions below are standard. Throughout, $A_{\mathrm{ex}} \in \mathbb{R}^{d\times m_{\mathrm{ex}}}$ with $m_{\mathrm{ex}} = \sum_{s<t} r_s$ is the column concatenation $[A_1^\top, \dots, A_{t-1}^\top]$ of the existing factors, and A^+ denotes the Moore-Penrose pseudoinverse of a matrix A. A new expert is placed by projecting a Gaussian candidate $\tilde{A}_t \in \mathbb{R}^{r_t\times d}$ onto the orthogonal complement of the occupied span,

$$A_t \;\leftarrow\; \tilde{A}_t\big(I - P_{\mathrm{ex}}\big), \tag{S11}$$

with the projector written for the full-column-rank case as

$$P_{\mathrm{ex}} \;=\; A_{\mathrm{ex}}\big(A_{\mathrm{ex}}^\top A_{\mathrm{ex}}\big)^{-1} A_{\mathrm{ex}}^\top. \tag{S12}$$

The implementation uses the form of Eq. (S12) while the near-saturation guard of Appendix A.6.1 has not fired and switches to the truncated singular value decomposition of that guard once the span approaches saturation; both are the AA^+ projector of the corresponding matrix, the second of the truncated factor rather than of A_{ex} itself.

Proposition 1 (The existing-span projector is an orthogonal projector)**.** *Let $P_{\mathrm{ex}} = A_{\mathrm{ex}}A_{\mathrm{ex}}^+$. Then $P_{\mathrm{ex}}^2 = P_{\mathrm{ex}}$, $P_{\mathrm{ex}}^\top = P_{\mathrm{ex}}$, and $\mathrm{range}(P_{\mathrm{ex}}) = \mathrm{col}(A_{\mathrm{ex}})$; that is, P_{ex} is the orthogonal projector onto the occupied span. When A_{ex} has full column rank, $A_{\mathrm{ex}}^+ = (A_{\mathrm{ex}}^\top A_{\mathrm{ex}})^{-1}A_{\mathrm{ex}}^\top$ and the expression coincides with Eq.* (S12)*.*

Proof. Write $A = A_{\mathrm{ex}}$. The Moore-Penrose conditions give $AA^+A = A$ and $(AA^+)^\top = AA^+$, which is symmetry. Idempotence follows from the first condition: $(AA^+)(AA^+) = (AA^+A)A^+ = AA^+$. For the range, $P_{\mathrm{ex}}x = A(A^+x) \in \mathrm{col}(A)$ for every x, and conversely any Ay satisfies $P_{\mathrm{ex}}(Ay) = (AA^+A)y = Ay$, so $\mathrm{col}(A) \subseteq \mathrm{range}(P_{\mathrm{ex}})$. A symmetric idempotent matrix is the orthogonal projector onto its range. The full-column-rank identity for A^+ is the standard one, and substituting it recovers Eq. (S12). Note that the pseudoinverse form requires no rank condition, so the three properties hold for every A_{ex}, including the rank-deficient case in which $(A_{\mathrm{ex}}^\top A_{\mathrm{ex}})^{-1}$ does not exist. □

Proposition 2 (Direct-sum dimension count)**.** *If the row spaces* $\mathrm{row}(A_t)$*,* $t \le T$*, are pairwise orthogonal and each* A_t *has full row rank* r_t*, then their sum is direct and* $\dim\big(\bigoplus_{t\le T}\mathrm{row}(A_t)\big) = \sum_{t\le T} r_t$*.*

Proof. For subspaces U, V one has $\dim(U+V) = \dim U + \dim V - \dim(U\cap V)$. Pairwise orthogonality forces $U\cap V = \{0\}$, so the correction vanishes and $\dim(U\oplus V) = \dim U + \dim V$. Iterating over $t \le T$, each new row space is orthogonal to the accumulated span and hence meets it trivially, so the dimensions add. Full row rank gives $\dim(\mathrm{row}(A_t)) = r_t$. □

Proposition 3 (Isotropic overlap identity)**.** *Let* $\mathcal{S} \subseteq \mathbb{R}^d$ *be a subspace of dimension* k *with orthogonal projector* $\Pi_{\mathcal{S}}$*, and let* u *be uniform on the unit sphere. Then* $\mathbb{E}\big[\|\Pi_{\mathcal{S}}u\|^2\big] = \mathrm{tr}(\Pi_{\mathcal{S}})/d = k/d$*. The quantity* k/d *is an average overlap, not a collision probability.*

Proof. $\|\Pi_{\mathcal{S}}u\|^2 = u^\top\Pi_{\mathcal{S}}u$, so $\mathbb{E}[u^\top\Pi_{\mathcal{S}}u] = \mathrm{tr}\big(\Pi_{\mathcal{S}}\,\mathbb{E}[uu^\top]\big)$. Isotropy gives $\mathbb{E}[uu^\top] = I/d$, whence the expectation equals $\mathrm{tr}(\Pi_{\mathcal{S}})/d$, and a rank-$k$ orthogonal projector has trace k. □

Proposition 4 (A freshly installed expert contributes the zero map at installation)**.** *With the up-projection factor initialised as* $B_t \leftarrow 0$*, the weight update of the new expert is* $\Delta W_t = B_tA_t = 0$*, so the newly installed expert contributes the zero map to the mixture of Eq.* (S7) *at the instant of installation, whatever* A_t *and whatever router weight it receives.*

Proof. $B_t = 0$ gives $B_tA_t = 0$ entrywise, so the summand $s_t(x)\,B_tA_t\,u$ of Eq. (S7) vanishes identically. Hence the installation step by itself introduces no new term: any change in the prediction at the instant of installation comes from the router renormalising over the enlarged pool, not from the map of the new expert, whose own contribution to the locality neighbourhood is produced by the subsequent fitting of B_t. □

Proposition 5 (Orthogonality at initialisation)**.** *At initialisation, that is, before any gradient step is taken on* A_t*, and when the near-saturation guard has not fired, so that the projector in use is the full* $P_{\mathrm{ex}} = A_{\mathrm{ex}}A_{\mathrm{ex}}^{+}$ *rather than a truncated one, the placement rule of Eq.* (S11) *gives* $A_tA_{\mathrm{ex}} = 0$ *for every candidate* $\tilde{A}_t$ *and every configuration of the heterogeneous ranks* $\{r_s\}_{s<t}$*.*

Proof. $A_tA_{\mathrm{ex}} = \tilde{A}_t(I - P_{\mathrm{ex}})A_{\mathrm{ex}} = \tilde{A}_t\big(A_{\mathrm{ex}} - P_{\mathrm{ex}}A_{\mathrm{ex}}\big)$, and $P_{\mathrm{ex}}A_{\mathrm{ex}} = A_{\mathrm{ex}}A_{\mathrm{ex}}^{+}A_{\mathrm{ex}} = A_{\mathrm{ex}}$ by the first Moore-Penrose condition, which holds for the pseudoinverse of A_{ex} itself whatever its rank; the premise that the near-saturation guard has not fired is what guarantees that the projector actually applied is this P_{ex} and not the truncated one of the near-saturation fallback, whose projector, built from only the retained singular factor of A_{ex}, does not reproduce A_{ex} in general. The bracket vanishes. The statement is confined to initialisation: assumption (c) above concerns pairwise orthogonality after fitting, which the soft regulariser encourages but does not guarantee. □

APPENDIX A.6: FAILURE CASES AND BOUNDARY CONDITIONS

This appendix sets out the regimes in which UAR-LoRA degrades and the fallbacks that handle them, so the long-stream results are read with their operating envelope attached.

A.6.1 CAPACITY SATURATION BEYOND THE NOMINAL HORIZON

The orthogonal initialiser can only place a new expert in a fresh direction while the occupied span has room. Once the realised cumulative rank approaches the residual width d, no free orthogonal complement remains, and the direct-sum accounting behind the appendix analysis (Appendix A.5) no longer holds. This defines a nominal orthogonal-capacity reference point $T^\star = \lfloor d/\mathbb{E}[r]\rfloor$ at which the occupied span approaches the residual width: 362 for the d=4096 backbones and 141 for GPT-2-XL, both at the $\mathbb{E}[r] = 11.3$ measured on LLaMA-2-7B. Both are read at the residual width, which is the input width of the projections that expand it; the projection that contracts back to it takes the wider input and reaches the same point later in the stream. Our long-stream results lie beyond it. Past $T^\star$ the near-saturation guard takes over. Firing once the condition number of $A_{\mathrm{ex}}^\top A_{\mathrm{ex}}$ degrades or A_{ex} reaches $0.95\,d$ columns, it switches placement to a truncated SVD clamped

at $\min\big(\mathrm{rank}(A_{\mathrm{ex}}),\ \lfloor 0.95\,d \rfloor\big)$ (Appendix A.4.2), so late edits are written into the orthogonal complement of the retained singular directions alone. Because that complement still contains the occupied directions the truncation discarded, orthogonality to the occupied span as a whole no longer holds. Past the nominal horizon the method continues through this fallback, and the retention reported at long streams is an empirical result measured under it; the heuristic analysis characterises only the pre-saturation regime.

A.6.2 Unreliable difficulty estimates

The entropy probe assumes the frozen-base predictive entropy is a usable difficulty signal. On short or degenerate prompts the estimate is unstable, which would missize the adapter. The fallback is triggered by a prompt-length check: prompts with too few effective tokens have their entropy recomputed as the mean over five sub-prompt windows, and prompts still flagged unreliable have their rank clamped to $r = 16$ rather than trusting an unstable extreme. This trades per-edit optimality for stability on the hardest-to-read prompts.

A.6.3 Out-of-pool queries and the abstention edge

The confidence gate protects locality by abstaining to the frozen base when routing confidence falls below τ. Two boundary behaviours follow. A false accept, an out-of-pool query wrongly admitted, is the calibration failure the gate is meant to suppress. The calibration-set ECE of 0.11 (Appendix A.11.7) is the gap between confidence and accuracy, not $\Pr(c_\phi \geq \tau \mid x \notin E_T)$, so it reports how far the gate's confidence can be trusted rather than how often this edge is crossed. A false abstain, an in-pool query wrongly refused, costs Reliability rather than Locality, so the threshold choice is a locality-reliability trade-off and $\tau = 0.7$ is the calibration-set minimiser.

A.6.4 Locality against a dedicated codebook

A dedicated codebook editor (GRACE) retains higher Locality across the stream because it never shares capacity between entries, whereas UAR-LoRA accepts a small locality gap in exchange for the generality that capacity sharing buys.

A.6.5 The two boundary conditions of the long-stream regime

The first boundary is the failure of the free-complement assumption; the second is divergence between the difficulty proxy and the true difficulty. We view widening the effective span (through periodic consolidation) as the natural next step.

APPENDIX A.7: Datasets and Probe Construction

All four editing benchmarks are read into the sequential-stream protocol used throughout the paper. CounterFact supplies clean counterfactual triples with explicit subject–relation–object structure, and its $21{,}919$ edits are the source of the headline comparison; each edit swaps a stored object for a counterfactual one, which makes it a strict test of whether a single targeted change takes effect without a rewrite of the surrounding context. ZsRE frames each of its $19{,}085$ items as a zero-shot relation-extraction question-answer pair, so an edit here is a question whose answer the model must be made to return, and it stresses whether an edit survives a change of surface phrasing. WikiBio contributes $5{,}946$ long-form biographical edits whose targets are spans of generated text rather than single tokens. The multi-hop set follows a MQuAKE-style construction in which each edited fact is paired with one-, two-, and three-hop compositional questions whose answer depends on the edit propagating through intermediate facts, so it tests downstream use of the edit.

Table S3 records the provenance of the four benchmarks. Edit counts are as used in this paper; the source is the originating benchmark.

The probe structure is uniform across benchmarks. For every edit we hold five paraphrase probes, which rephrase the query while preserving its meaning and are used to score Generality, and ten locality probes, which are drawn from facts that share a subject or relation with the edit but whose

Table S3: Dataset provenance. Edit counts and form are as used here. At every report the three accuracy denominators are $N_{\mathrm{Rel}} = 1{,}000$, $N_{\mathrm{Gen}} = 5{,}000$ and $N_{\mathrm{Loc}} = 10{,}000$ on every one of these benchmarks.

Dataset	Origin	Edits	Form
CounterFact	ROME benchmark	21,919	triple
ZsRE	zero-shot RE	19,085	QA
WikiBio	long-form bios	5,946	long-form
MQuAKE-style	MQuAKE	3,000	multi-hop

answer must remain unchanged and are used to score Locality. The rationale for scoring Locality on neighbouring rather than distant facts is given with the metric definitions in Appendix A.9.1. The sequential stream is formed by sub-sampling edit blocks in a fixed presentation order and reporting at stream lengths T, so that a report at a given length always queries the model after all T edits have been applied in sequence. Within a block, each successively longer stream contains the shorter ones as a prefix, which is what lets the scaling behaviour be read as the growth of a single accumulating pool rather than as unrelated runs.

Scoring is separated from stream length: at every report the three metrics are evaluated on one fixed probe pool of $1{,}000$ edits drawn once before the stream begins and held identical across all lengths and all methods, so that lengths are compared on the same instrument rather than on pools that grow with T. The pool is held out from the calibration set of Appendix A.11.7 and from every hyperparameter choice; it is not held out from the stream, since Reliability asks whether an installed target is recalled and can only be scored on edits the stream has installed. Reliability scores one indicator per pool edit and so has denominator $1{,}000$; Generality scores five paraphrase probes per pool edit and so has denominator $5{,}000$; Locality scores ten locality probes per pool edit and so has denominator $10{,}000$. Streams are drawn from the benchmark's edit set and are arranged so that a report at any length scores the same $1{,}000$ pool edits. At $T \geq 1000$ a single stream installs the whole pool and is carried on to length T with further edits of the same benchmark. At $T < 1000$ the pool is covered by $\lceil 1000/T \rceil$ disjoint streams of T pool edits each: ten streams of one hundred at $T = 100$, and the single-edit setting at $T = 1$. Within a stream the edits are applied in one fixed order and never rolled back, so T controls how many edits are resident when a pool edit is scored and not how many edits are scored. Summing the three, every report rests on $16{,}000$ $\arg\max$ decisions ($1{,}000 + 5{,}000 + 10{,}000$), the same count at every length, which is what makes the reports at different lengths directly comparable. Every Reliability, Generality and Locality entry in the tables is a proportion over one of these three denominators, evaluated once under the single presentation order, with greedy decoding and seed 0.

A.7.1 Probe Construction

Edits are streamed in blocks and evaluated at $T \in \{1, 100, 500, 1000, 2500, 5000\}$. The stream-length scaling of Appendix A.11.3 reports the four lengths $T \in \{1, 100, 1000, 5000\}$; the two intermediate lengths $T = 500$ and $T = 2500$ enter the cumulative Locality drop of Figure S3 under the same protocol.

Each edit in every benchmark carries five paraphrase probes, used to score generality, and ten locality probes drawn from neighbouring facts, used to score locality. The locality probes are neighbouring subject–relation pairs that share an entity or a relation with the edit target but whose correct answer the edit is not meant to change, so that a locality failure registers only when the model's behaviour moves on material the edit never addressed. The same probe counts are used at every reporting length, so the generality and locality columns are comparable across T. The per-benchmark edit counts and the probe counts that define the evaluation material are given above.

APPENDIX A.8: Baseline Configurations

We describe each baseline together with the published configuration it uses.

- **Fine-Tune**: unconstrained gradient updates to the backbone weights, no locality constraint.

- **ROME** (Meng et al., 2022): rank-one update to a single located MLP layer.
- **MEMIT** (Meng et al., 2023): batched rank-one updates spread over a set of critical MLP layers (five critical layers in our LLaMA-2-7B configuration).
- **GRACE** (Hartvigsen et al., 2023): a discrete codebook of edited activations, one entry retrieved per edit key.
- **MELO** (Yu et al., 2024): dynamic low-rank memory that grows adapters as edits arrive.
- **SoLA** (Luo et al., 2026): a reversible lifelong editor that routes each query to a LoRA module by semantic scope.
- **ELDER** (Li et al., 2025): a fixed-rank LoRA lifelong editor with a learned router, the strongest baseline.

APPENDIX A.9: METRIC DEFINITIONS AND PARAMETER ACCOUNTING

A.9.1 METRIC DEFINITIONS

Let f_θ be the edited model and $\mathcal{P}$ the fixed probe pool on which every report is scored (Appendix A.7). For an edit $e \in \mathcal{P}$ let (x_e, y_e) be its query and target, $P(e)$ its paraphrase set, and $N(e)$ its locality neighbours with pre-edit answers y_n.

Reliability. The fraction of edits whose target is recalled:

$$\mathrm{Rel} = \frac{1}{|\mathcal{P}|} \sum_{e \in \mathcal{P}} \mathbb{1}\left[\arg\max_y f_\theta(y \mid x_e) = y_e\right]. \tag{S13}$$

Generality. Generality applies the same criterion to the paraphrase set, so that an edit counts only if it survives rewordings of the query rather than being memorised at a single surface form:

$$\mathrm{Gen} = \frac{1}{|\mathcal{P}|} \sum_{e \in \mathcal{P}} \frac{1}{|P(e)|} \sum_{x \in P(e)} \mathbb{1}\left[\arg\max_y f_\theta(y \mid x) = y_e\right]. \tag{S14}$$

Locality. Locality scores the neighbouring queries whose pre-edit answers the edit was not meant to change:

$$\begin{aligned} \mathrm{Loc} = \frac{1}{|\mathcal{P}|} \sum_{e \in \mathcal{P}} \frac{1}{|N(e)|} \\ \times \sum_{(x_n, y_n) \in N(e)} \mathbb{1}\left[\arg\max_y f_\theta(y \mid x_n) = y_n\right]. \end{aligned} \tag{S15}$$

Expected Calibration Error. The three accuracies above say nothing about whether the confidence driving the gate can be trusted, which is what the calibration error measures. With predictions binned into M equal-width confidence bins B_m,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \tag{S16}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the mean accuracy and mean predicted confidence within the bin. We use $M = 10$. In the confidence-gate threshold sweep of Appendix A.11.7 the scored event is the correctness of the gate's admit-or-abstain decision. All accuracy metrics are reported as percentages and ECE on the $[0, 1]$ scale.

Scope of the cost claims. Compute cost is expressed as the trainable-parameter ratio, whose accounting is given in Appendix A.11.8.

Numerator of the Param column. The numerator is stated per family: for methods that rewrite backbone weights it is the weights they rewrite; for GRACE and MELO, whose stores grow with the stream, it is the parameters they hold at the reported stream length; for UAR-LoRA it is its adapter parameters trained for each edit, averaged over the stream, together with the control modules charged at their full size at the reported stream length, the trainable parameters among those modules being the router weights and the confidence head. The store-based and the per-edit numerators are two time bases, so the column is read within a family rather than as one cost across families.

The four metrics trade against one another and are reported separately. Reliability and Generality score the effect of the edit; Locality scores unchanged behaviour elsewhere. A method can trade one for the other, and a lifelong editor must hold all three simultaneously. The indicator in Reliability and Generality is an exact match on the argmax prediction, so partial credit is never awarded: a paraphrase counts only if it recovers the target, which makes Generality a strict test of whether the edit transfers beyond the exact query surface. Locality is measured on neighbours that share a subject or relation with the edit, which are the facts most likely to change; distant facts are preserved by default and would inflate the score. ECE is separated from accuracy because a model can be accurate yet overconfident, and it is overconfidence on out-of-pool queries that the confidence gate is designed to suppress, so ECE is the metric that most directly reflects the gate's contribution.

Reliability is a proportion of independent indicator outcomes over a known denominator, one per pool edit, so the sampling error attached to a single Reliability cell is $\sqrt{p(1-p)/N_{\mathrm{Rel}}}$. For the UAR-LoRA cell on CounterFact at $T = 1000$ this gives 0.81 points ($p = 0.930$, $N = 1000$), and the corresponding ELDER cell carries 0.93 points. Generality and Locality are not proportions of independent probes. The five paraphrase probes of an edit are rewordings of one query, and its ten locality probes are neighbours that share its subject or relation, so the outcomes within an edit are dependent and the independent unit is the edit, of which the pool holds $1{,}000$, rather than the probe, of which it holds $5{,}000$ and $10{,}000$. Placing no constraint on the correlation within an edit, that is reading an edit's probes as if they moved together, gives the conservative bound $\sqrt{p(1-p)/1{,}000}$: 1.08 points on Generality ($p = 0.864$) and 0.86 on Locality ($p = 0.920$). The probe-level values 0.48 and 0.27 are the opposite extreme and would hold only if the probes of one edit were independent of one another. Under non-negative within-edit correlation the sampling error lies between the two, and the figures quoted here and below are the conservative end. On the ZsRE stream at $T = 5000$ the figures are 0.75 points on Reliability ($p = 0.941$) and 0.95 on Locality ($p = 0.899$); the WikiBio Reliability cell carries 0.99 points and the three-hop cell 1.58. An exact interval is available where the count is binomial: the Clopper–Pearson 95% interval for 930 of 1000 on Reliability is $[91.2, 94.5]$. These figures bound only the finite-probe component of the uncertainty; they do not cover the variation a different presentation order would produce (Appendix A.10.3).

For a difference between two cells scored on the same probe set, $\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$ is an upper bound on the sampling error of the difference whenever the two methods' per-item outcomes are non-negatively correlated, which is what one expects when both are scored on the same queries; under negative correlation the expression understates it. On that reading the 2.5-point Reliability margin over ELDER at $T = 1000$ carries a difference bound of 1.23 points and stands at about 2.0 times it, the 4.2-point Generality margin carries a bound of 1.62 points and stands at about 2.6 times it, and the 0.9-point Reliability gap to the fixed-rank-32 variant of Appendix A.11.6 carries a bound of 1.17 points and so falls below one, which is the arithmetic behind reading that row as unsettled.

A.9.2 Backbone Models

The evaluation uses the three backbones in Table S4, each loaded from the HuggingFace checkpoint named there. The method is written for a single adapted projection and applied identically at every projection of every adapted feed-forward sublayer, so both the budget of Eq. (S1) and the reported parameter ratios are summed over all of them.

All edits act on the feed-forward sublayers of the backbone; the backbone weights themselves remain frozen and only the injected adapters, the confidence head, and the router weights are trained. The fixed rank-32 LoRA reference baseline against which the Param ratios are normalised places a rank-32 adapter on the gate, up, and down projections of every one of the backbone's feed-forward sublayers (all 32 layers of LLaMA-2-7B), which comes to $32 \times 3 \times 32 \times (4096 + 11008) = 46.4$M parameters. Every Param entry in the paper is a count of the parameters that method trains or

Table S4: Backbone models and their HuggingFace identifiers.

Backbone	Scale	HuggingFace identifier
GPT-2-XL	1.5B	`gpt2-xl`
LLaMA-2-7B	7B	`meta-llama/Llama-2-7b-hf`
LLaMA-3-8B	8B	`meta-llama/Meta-Llama-3-8B`

rewrites divided by this same 46.4M. For a locate-then-edit method the numerator is the backbone weights it rewrites, namely the full projection matrix of each edited layer. That is why a single-layer rewrite and a five-layer rewrite land near and above $1\times$ respectively, rather than at the size of their rank-one deltas. Concretely, one rewritten down-projection is $11008 \times 4096 = 45.09$M parameters, so a single-layer rewrite gives $45.09\text{M}/46.4\text{M} = 0.97\times$, and the five critical layers of the batched variant give $5 \times 45.09\text{M}/46.4\text{M} = 4.86\times$. For unconstrained fine-tuning the numerator is the whole backbone, 6.74B parameters, giving $6.74\text{B}/46.4\text{M} = 145\times$. For UAR-LoRA it is the rank-weighted adapters together with the control modules of the confidence-gated routing path, giving the all-in $0.44\times$. Itemised at $T = 1000$, that numerator is the rank-weighted adapters, $11.3/32 \times 46{,}399{,}488 = 16{,}384{,}819$; the router weights W_s at $|E| \times d = 1000 \times 4096 = 4{,}096{,}000$; and the confidence head w_c at $d = 4{,}096$. The entropy probe and the rank selector contribute none. The three terms sum to $20{,}484{,}915$, and $20{,}484{,}915/46{,}399{,}488 = 0.4415$, which is the $0.44\times$ reported; the two heads alone are $4{,}100{,}096$, or 0.088 of the reference footprint. The $32 \times 3 \times 32 \times (4096 + 11008)$ reference expands to $46{,}399{,}488$ parameters, which is the 46.4M used as the denominator throughout.

APPENDIX A.10: EVALUATION PROTOCOL, COMPUTE AND RANDOMNESS

A.10.1 EVALUATION PROTOCOL

Edits are applied in a fixed presentation order and never rolled back. At each reported stream length T, the model state after all T edits is queried on the edit set, the paraphrase set, and the locality set to produce Reliability, Generality, and Locality; ECE is computed over the same predictions with $M = 10$ bins, so the ECE reported in Appendix A.11.1 is an evaluation-set quantity. The Platt scaler and the threshold τ of the confidence gate are fit only on the held-out calibration set of Appendix A.11.7 and are never exposed to evaluation items. Every reported number is a single value under this fixed protocol.

Reporting at increasing T over one fixed order makes the scaling and forgetting tables comparable across methods. No edit is ever rolled back or replayed, so a report at length T always reflects the cumulative effect of all earlier edits on the current one. All four metrics are read from the same model states at each length, so Reliability, Generality, Locality, and ECE at a given T describe one measurement rather than separately optimised runs. The threshold is frozen after calibration, so the abstention behaviour scored on the locality set is the behaviour the deployed gate would exhibit.

A.10.2 COMPUTE ENVIRONMENT

The configuration is a single NVIDIA A100 80GB GPU with CUDA 12.8, PyTorch 2.11.0 and HuggingFace Transformers 5.14.1, in `bfloat16` precision, as listed in Table S5. The remaining pinned dependencies are PEFT 0.19.1, Datasets 5.0.0, Accelerate 1.14.0, Safetensors 0.8.0, NumPy 2.4.6, SciPy 1.18.0, scikit-learn 1.9.0 and Matplotlib 3.10.9.

A.10.3 RANDOMNESS AND DETERMINISM

Three sources of randomness enter the method: the Gaussian draw of each adapter's down-projection factor before orthogonal projection (Eq. (S11)), the sub-sampling of edit blocks into the fixed presentation stream, and the draw of the negative query used by the confidence objective, as listed in Table S2. All runs use a fixed random seed of 0, applied to Python's `random` module, to NumPy and to PyTorch on every CUDA device, and each reported number is a single run under one fixed presentation order.

Table S5: Compute environment.

Field	Value
GPU	NVIDIA A100 80GB
CUDA	12.8
Framework	PyTorch 2.11.0, Transformers 5.14.1
Precision	`bfloat16`
Random seed	0

A.10.4 Prompt Templates and Decoding

Edit, paraphrase, and locality queries follow the native templates of their source benchmarks (CounterFact, ZsRE, WikiBio, and the MQuAKE-style multi-hop set of Appendix A.7). Structurally, an edit query presents the subject–relation cue and scores the target continuation; a paraphrase query rewrites that cue while holding the target fixed; a locality query presents a neighbouring subject–relation whose pre-edit answer must be preserved. Decoding for every metric is greedy: all of Reliability, Generality, and Locality (Appendix A.9.1) score the $\arg\max$ prediction against the target, with no sampling, temperature, or beam search, so the reported accuracies are deterministic given the model state.

APPENDIX A.11: Experiments in Detail

Each unit below gives the setting of one experiment, how each of its columns and rows is read, and its extended analysis, including the finite-probe error bounds of Appendix A.9.1. Unless a unit states otherwise, every value is a single run under one fixed presentation order, seed 0, with greedy decoding (Appendix A.10).

A.11.1 Main Comparison, Calibration and Accuracy

Setting. CounterFact on LLaMA-2-7B at $T = 1000$, with the seven baselines of Appendix A.8 and UAR-LoRA scored on the same fixed probe pool, so the accuracy denominators are $N_{\mathrm{Rel}} = 1000$, $N_{\mathrm{Gen}} = 5000$ and $N_{\mathrm{Loc}} = 10000$. UAR-LoRA runs at the operating threshold $\tau = 0.7$ of Appendix A.11.7.

Reading the columns. Reliability, Generality and Locality are percentages, higher is better. ECE is on the $[0, 1]$ scale, lower is better, and is computed with $M = 10$ bins on the evaluation predictions of Appendix A.10.1; it is not the calibration-set ECE of the threshold sweep in Appendix A.11.7. Bold marks the best value in each quality column. Param is a ratio to the fixed rank-32 LoRA reference footprint of 46.4M parameters and is not bolded, since its numerator is stated per family (Appendix A.9.1). ELDER ($1.00\times$) and UAR-LoRA ($0.44\times$) are costed on the same per-edit basis, ours all-in with the router weights and the confidence head charged at their full size at $T{=}1000$; GRACE's $1.00\times$ is the store it holds at $T{=}1000$ and is read within the store-based family. The locate-then-edit and fine-tuning entries are derived in Appendix A.9.2.

Error bounds. The UAR-LoRA Reliability cell carries 0.81 points of sampling error and the ELDER cell 0.93. Read with the edit as the independent unit, the 2.5-point Reliability margin over ELDER stands at about 2.0 times its difference bound of 1.23 points, and the 4.2-point Generality margin at about 2.6 times its bound of 1.62 (Appendix A.9.1). ECE is the metric that most directly reflects the gate's contribution, because it is overconfidence on out-of-pool queries that the confidence gate is designed to suppress.

A.11.2 Backbone Transfer

Setting. CounterFact at $T = 1000$ on the three backbones of Appendix A.9.2, comparing ELDER and UAR-LoRA under one configuration held fixed across backbones (Appendix A.4.2). The LLaMA-2-7B rows are the ELDER and UAR-LoRA entries of Appendix A.11.1.

Reading the rows. Each backbone contributes one pair of rows. Bold marks the better of the two methods on that backbone in Reliability and Generality; Locality is not bolded, for the reason given below.

The three margins as a set. Taken across the three backbones the Reliability differences are $+1.8$, $+2.5$, $+3.4$ with a median of $+2.5$, and the Generality differences $+3.7$, $+4.2$, $+4.1$ with a median of $+4.1$; both are one-signed. The Locality differences are -0.6, $+0.4$, $+1.1$ with a median of $+0.4$, and they change sign. Finite-probe error alone already leaves them unseparated: read with the edit as the independent unit, the difference bound of Appendix A.9.1 is about 1.2 points, which none of the three reaches. We therefore mark no Locality win.

A.11.3 Stream-Length Scaling

Setting. ZsRE on LLaMA-2-7B at $T \in \{1, 100, 1000, 5000\}$. Every length is scored on the same fixed pool of $1{,}000$ edits; at $T \geq 1000$ one stream installs the pool and continues to length T, and at $T < 1000$ the pool is covered by disjoint streams of T edits each (Appendix A.7).

Reading the rows. The upper block reports Reliability for MEMIT, GRACE, ELDER and UAR-LoRA, and the lower block Locality for GRACE, ELDER and UAR-LoRA. Bold marks the best value per column within each block.

Margins and error. Read on the accuracy scale of this stream, the Reliability margins over ELDER at $T \in \{1, 100, 1000, 5000\}$ are $+2.5$, $+4.0$, $+6.6$ and $+7.9$ with a median of $+5.3$, and the Locality margins $+0.2$, $+0.3$, $+1.4$ and $+2.5$ with a median of $+0.85$; both are one-signed and both widen with T, while the Reliability sampling error at $T = 5000$ is 0.75 points and the Locality error 0.95 on the edit as the independent unit.

The regime past the saturation point. On LLaMA-2-7B the nominal orthogonal-capacity reference point is $T^\star = 362$ (Appendix A.6.1), so the $T = 1000$ and $T = 5000$ columns lie past it. There the near-saturation guard places each new expert in the complement of a truncated SVD of the occupied span, and the retention in those columns is an empirical result measured under that fallback. The isolated effect of each module is measured at $T = 1000$ in Appendix A.11.6.

A.11.4 Long-Form Editing

Setting. WikiBio on LLaMA-2-7B at $T = 1000$, whose edit targets are spans of generated text rather than single tokens. Denominators are $N_{\mathrm{Rel}} = 1000$, $N_{\mathrm{Gen}} = 5000$ and $N_{\mathrm{Loc}} = 10000$. Bold marks the best value per column.

Margins and error. The three differences over ELDER are $+3.2$, $+4.8$ and $+1.3$ with a median of $+3.2$, all one-signed, and the Reliability cell here carries 0.99 points of sampling error at $N = 1000$. Read with the edit as the independent unit, the Reliability and Generality margins exceed their difference bounds of 1.48 and 1.78 points, while the Locality margin of 1.3 sits just below its bound of 1.40 and is not separated by the probe set.

A.11.5 Multi-Hop Composition

Setting. The MQuAKE-style set on LLaMA-2-7B at $T = 1000$. Each hop column scores one question per edit, so $N = 1000$ per cell, and reports end-to-end accuracy at that depth. Bold marks the best value per column.

Margins and error. UAR-LoRA leads ELDER by 3.3 points at one hop (88.4 vs 85.1), by 6.6 points at two hops (67.9 vs 61.3), and by 7.4 points at three hops (48.6 vs 41.2). The three margins are $+3.3$, $+6.6$ and $+7.4$ with a median of $+6.6$; the three-hop cell sits nearest the mid-range of the scale and so carries the largest sampling error of the three, 1.58 points at $N = 1000$, which the 7.4-point margin still exceeds several times over.

A.11.6 Component Ablation and Interaction of the Modules

Setting. CounterFact on LLaMA-2-7B at $T = 1000$, under the protocol of Appendix A.11.1.

Reading the rows. The first row is the full system. The four remove-one rows each take out one module and name its replacement in parentheses: a fixed rank of 32 or 8 in place of the entropy probe, soft mixture-of-experts routing in place of the confidence gate, and a random draw in place of the orthogonal initialiser. The three only-rows keep the named module alone. Param is all-in, counted as in Appendix A.11.1; the full row is $11.3/32 + 0.088 = 0.44$, and the two fixed-rank rows are $32/32 + 0.088 = 1.09$ and $8/32 + 0.088 = 0.34$, with the control share 0.088 itemised in Appendix A.9.2.

Scope. The ablation covers the three modules and nothing else. Two further pieces of the method carry no row of their own. The unreliable-entropy fallback is a stability guard specified with the rest of the hyperparameters (Appendix A.4.2). The interference score is an analysis quantity (Appendix A.5) and is not varied experimentally.

Reading the rank rows. Neither fixed rank matches the adaptive allocation on accuracy and budget at once. Since the fixed-rank-32 row is within a point on Reliability yet costs about $2.5\times$ the adaptive budget, the case for adaptive rank rests on the accuracy and budget columns read together. Read against the difference bound of Appendix A.9.1, the three Reliability removals separate very differently: the fixed-rank-32 row differs by 0.9 points against a bound of 1.17, so it sits below one multiple and reads as equal accuracy; the random-initialisation row differs by 2.2 points against 1.22, about 1.8 multiples; the fixed-rank-8 row differs by 3.3 points against 1.25, about 2.6. On Locality the w/o Gate row falls 3.1 points below the full system against a bound of 1.31, about 2.4 multiples. These multiples order the rows by how far each exceeds finite-probe error alone and are not significance statements.

Reading the only-rows. The Reliability losses of the remove-one rows are 1.6 for the gate and 2.2 for the initialiser, and removing both leaves the only-Probe row at a loss of 2.1, below their sum of 3.8. The capacity and coherence reading of these rows is given in Appendix A.12.

A.11.7 Sensitivity to the Gate Threshold

Calibration set. The Platt scaler applied to the confidence head and the abstention threshold τ are fit on held-out facts drawn from a calibration set of 500 paraphrases; the confidence head itself is trained on the edit stream with the adapters. These paraphrases are disjoint from both the edit queries and the locality probes: no calibration item shares its surface form with any item used to compute Reliability, Generality, or Locality. The threshold sweep of $\tau \in \{0.5, \ldots, 0.9\}$ is carried out on this held-out set, and the operating point $\tau = 0.7$ is the one that minimises calibration-set ECE. The ECE row of the sweep is therefore a calibration-set quantity, whereas the ECE of Appendix A.11.1 is computed on the evaluation predictions of Appendix A.10.1; the two are reported separately and neither is derived from the other.

The split is at the level of the fact rather than the individual probe, because a naive split by probe would still let a calibration paraphrase and an evaluation paraphrase of the same underlying fact land on opposite sides, leaking surface information across the boundary: once an edit contributes any paraphrase to the calibration set, none of its remaining probes may appear in the evaluation sets, and the reverse holds as well.

Reading the rows. ECE is binned with $M = 10$ and scores the correctness of the gate's admit-or-abstain decision. The abstention decision accuracy is the fraction of correct decisions over out-of-pool queries that should be abstained and in-pool queries that should not. The routing accuracy is the in-pool routing accuracy. Bold marks the best value per row. The sweep carries no Locality column.

The choice of the operating point. A false abstain costs Reliability and a false accept costs Locality, so the threshold choice is a locality-reliability trade-off (Appendix A.6.3).

A.11.8 COST PROFILE

Setting. CounterFact on LLaMA-2-7B at $T = 1000$. The parameter ratio is the all-in per-edit ratio against the fixed rank-32 LoRA reference.

Per-edit cost. Each edit does three things. The entropy probe is a single frozen-base forward pass on the edit query. Orthogonal initialisation forms and applies the projector of Eq. (S12) onto the occupied span; its cost grows with the cumulative occupied rank $\sum_{s<t} r_s$ rather than with the ambient width alone. Fitting runs AdamW on the selected rank-r adapter together with the two heads (schedule in Table S2).

Parameter footprint. The $0.44\times$ ratio is a per-edit ratio against a fixed rank-32 LoRA adapter: its numerator is the mean adapter size over the stream plus the control modules of the confidence-gated routing path (router, confidence head, entropy probe, and rank selector) charged at their full size at $T = 1000$. Two of those four control components carry no trainable parameters of their own: the entropy probe is a frozen-base forward pass and the rank selector is the deterministic rank bucketing, so the control share is carried by the router weights W_s and the confidence head w_c; the two are itemised in Appendix A.9.2 and enter the parameter ratio rather than appearing as a line of their own. The ratio counts all trainable components, not the adapter rank alone.

Per-query cost. At inference each query pays one probe-free forward pass through the frozen base, one confidence-head evaluation, and, when the gate admits, the selected expert mixture. Abstained queries skip the expert mixture entirely and cost only the gate plus the frozen-base pass. The entropy probe runs once per edit and is not on the inference path.

APPENDIX A.12: DISCUSSION

Coherence versus capacity. The component ablation of Appendix A.11.6 exposes an asymmetry between two ways of failing under a long stream. Capacity, the ability of a single adapter to absorb one hard edit, drives raw Reliability: pinning the rank moves Reliability while leaving Locality almost where it was. The second is coherence, whether a new edit can be written without disturbing the ones already present. The largest Locality movement in the ablation follows from removing the gate, which is a routing decision taken at inference rather than part of the writing geometry, so it is evidence about what a route back to the frozen base buys on out-of-pool queries and not about coherence in the sense just given. The initialiser's Locality effect changes sign within that table, so the case for coherence rests on Reliability: at $T = 1000$, replacing the orthogonal initialiser with a random one costs 2.2 Reliability points. The flatter long-stream curve of the ZsRE stream (Appendix A.11.3) is a property of the complete system.

Abstention as locality insurance. The threshold sweep of Appendix A.11.7 shows what the gate trades on the calibration set: as τ rises from 0.5 to 0.7, the abstention decision accuracy climbs from 71.2 to 86.3 while in-pool routing accuracy falls from 95.8 to 92.6. The sweep carries no Locality column, and the Locality readings below are the evaluation-set ones of the ZsRE stream. The gate is conservative: borderline queries are routed to the unedited base rather than to an adapter. That trade is consistent with the Locality behaviour we report as T grows. On that stream Locality holds within about two points of a dedicated codebook even as the stream reaches five thousand edits.

Positioning versus prior editors. Compared with recent lifelong editors, UAR-LoRA differs in the three conditions of the criterion. ELDER (Li et al., 2025) routes softly over a fixed-rank adapter set: it controls neither the capacity condition, since rank is fixed regardless of edit difficulty, nor the admissibility condition, since its deferral tracks no calibrated score. MELO (Yu et al., 2024) grows low-rank memory but likewise fixes how each edit is sized and activates blocks by discrete index, so a perturbed paraphrase can miss the pool. GRACE (Hartvigsen et al., 2023) isolates each edit in a discrete codebook entry, which explains its strong Locality, but sharing reaches only as far as each key's deferral radius, which is the mechanism behind its Generality shortfall. UAR-LoRA approaches the separation of a codebook while widening that reach, by making the placement continuous and orthogonal rather than discrete, and by setting its threshold on held-out calibration data, at the point where calibration is tightest.

Scope of the analysis. The capacity-coherence criterion is a heuristic modelling argument that relies on approximate stationarity of the rank sequence and on a realised-capacity condition. It is stated up to the nominal orthogonal-capacity reference point $T^\star$, the stream length at which the occupied span approaches the ambient dimension d. The entropy probe reads difficulty from the frozen base's own predictive distribution.