IMPROVING ATOMIC-FACT RECALL VIA FOCUSED VIEWS IN UNSTRUCTURED KNOWLEDGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly serve as general-purpose interfaces to factual knowledge, but their parameters do not automatically reflect information that changes after pretraining. Knowledge editing (KE) provides a targeted alternative to costly retraining by modifying selected knowledge while preserving unrelated knowledge and general capabilities. Conventional KE uses structured factual triples, whereas unstructured KE (UKE) uses free-form passages containing multiple facts. Nonetheless, existing UKE editors exhibit a failure mode known as *context reliance*: edited LLMs can often reproduce the editing passage but fail to reliably recall its individual facts without the original passage context. We identify *context-induced difficulty underestimation* under the standard passage-level editing objective: later facts receive increasingly rich ground-truth context and consequently incur lower initial losses, making them appear easier to learn. In response, we propose FOVEATED, a plug-and-play framework that constructs focused views of each sentence by randomly shifting the Rotary Position Embedding (RoPE) positions assigned to the keys of its preceding context. The perturbation is applied during editing and removed afterward, leaving the model's native positional encoding unchanged at inference time. We instantiate FOVEATED for both direct-optimization and locate-then-edit editors. We theoretically analyze how FOVEATED counteracts context-induced difficulty underestimation and empirically demonstrate consistent improvements across five KE editors, two LLM backbones, and three benchmarks.

1 INTRODUCTION

Large language models (LLMs) have become a foundation for modern natural language processing tasks (Vaswani et al., 2017; Brown et al., 2020; Zhou et al., 2023). As model scale continues to grow, these capabilities make LLMs increasingly useful as general-purpose interfaces for accessing and reasoning with knowledge (Bommasani et al., 2021; Roberts et al., 2020; Kojima et al., 2022).

However, an LLM's parametric knowledge remains largely fixed after pretraining and can quickly become outdated as real-world knowledge evolves, leading to factual errors in downstream applications (De Cao et al., 2021; Hartvigsen et al., 2023). Keeping the model current through repeated retraining is prohibitively expensive. As a remedy, *knowledge editing* (KE) addresses this problem by incorporating specific new knowledge through selective model updates, while preserving unrelated knowledge and general capabilities (Wang et al., 2023; Yao et al., 2023; Zhang et al., 2024b).

Standard KE typically represents new knowledge as a separate factual association, often a *(subject, relation, object)* triple. Given such structured facts, some methods directly optimize selected parameters (Wang et al., 2024a;b), while *locate-then-edit* (LTE) methods first identify where the target knowledge is represented and then write a targeted update into the corresponding layers (Meng et al., 2022; 2023; Fang et al., 2025). Recent work relaxes this structured-input assumption through *unstructured knowledge editing* (UKE), where the new knowledge is provided as a free-form passage (Wu et al., 2024a; Yang et al., 2025). Unlike a structured edit, such a passage may express multiple interrelated facts without specifying which ones a user may later query. UKE methods accordingly expand the editable modules and decompose the passage into smaller units and coordinate their updates (Deng et al., 2025; Jiang et al., 2025; Su et al., 2025). Together, these methods expand the scope of KE from individual factual associations to free-form passages.

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.

Nonetheless, existing methods often fall short of making effective use of the edited knowledge beyond the original editing context. In particular, an edited model may recall the complete passage under the original editing query, yet fail to retrieve the new knowledge when answering more targeted questions about individual facts (Zhou et al., 2026; Wang et al., 2026). Recent studies have begun to examine this failure mode, termed *context reliance*, in which the new knowledge may remain accessible only as part of the complete passage, where the edited model can rely on the preceding context as a shortcut during generation. To mitigate this problem, Zhou et al. (2026) align predictions under global and local contexts; Wang et al. (2026) synthesize atomic question–answer supervision with external models; and Liu et al. (2026) construct hybrid self-distillation trajectories from a privileged in-context state. *Yet the learning dynamics underlying context reliance remain underexplored.*

In light of this gap, we examine how context affects the apparent learning difficulty under the standard passage-level objective in this work. Interestingly, we find that the pre-update editing loss decreases for facts appearing later in the passage, even though all edited facts are new to the model (Wu et al., 2024a; Deng et al., 2025). Intuitively, under teacher forcing (Lamb et al., 2016), later content is predicted from longer ground-truth prefixes and thus receives richer contextual support. This growing support acts as a *shortcut*, lowering the initial losses of later facts and making them appear easier to learn. We call this *context-induced difficulty underestimation*; Sec. 2 details our analysis. *This observation is the first contribution of this work.*

In response, we propose FOVEATED to construct *focused views* that make the preceding context less reliable as a shortcut during editing. By varying how a target sentence retrieves its preceding context while preserving the sentence itself, FOVEATED trains the model on each fact across randomized contextual views and directs the editing signal toward the target content. Concretely, FOVEATED constructs such views by randomly shifting the *key-side* Rotary Position Embedding (RoPE; Su et al., 2021) positions assigned to the prompt and preceding sentences, while leaving all query rotations and the target sentence's key rotations unchanged. The perturbation is applied only during editing and removed afterward, leaving the model's native positional encoding unchanged at inference time. We further provide theoretical support for the focused-view design.

FOVEATED is a flexible framework that intervenes in the editing computation without prescribing how the updates are parameterized. It can be applied to a wide range of existing editors, and we instantiate FOVEATED for two broad families. Direct optimization (DO) editors update selected parameters with gradient descent (Hu et al., 2022; Wang et al., 2024a;b; Liu et al., 2025b). For this family, FOVEATED generalizes the standard passage-level objective to a weighted expectation of sentence-level losses under focused views. Setting the perturbation magnitude to zero exactly recovers the standard passage-level objective. For LTE editors (Deng et al., 2025; Su et al., 2025), applying FOVEATED is less direct because their passage objectives optimize intermediate hidden targets rather than the eventual parameter updates. We follow Wang et al. (2026) and adopt a two-stage design by first writing sentence-level targets into shallow layers and then applying the base editor in deeper layers to learn the complete passage. The key distinction of our approach is that FOVEATED obtains its sentence-level targets directly from the editing passage under focused views, without synthesizing atomic question–answer supervision through external models. See Sec. 3 for details. *This general editing framework and our theoretical analysis form the main technical contribution of this work.*

The remainder of this paper is organized as follows. Sec. 2 introduces context-induced difficulty underestimation, Sec. 3 presents FOVEATED and its theoretical analysis, and Sec. 4 evaluates the method. Secs. 5 and 6 review related work and conclude the paper, respectively.

2 Difficulty Underestimation in Unstructured Knowledge Editing

This section reveals context-induced difficulty underestimation in UKE. Background is also provided.

2.1 Preliminaries

For a text sequence $\boldsymbol{x}$ of length $|\boldsymbol{x}|$, an autoregressive LLM $\pi(\cdot)$ factorizes its likelihood as

$$\pi(\boldsymbol{x}) = \prod_{t=1}^{|\boldsymbol{x}|} \pi(x_t \mid x_1, \ldots, x_{t-1}) \triangleq \prod_{t=1}^{|\boldsymbol{x}|} \pi(x_t \mid \boldsymbol{x}_{<t}),$$

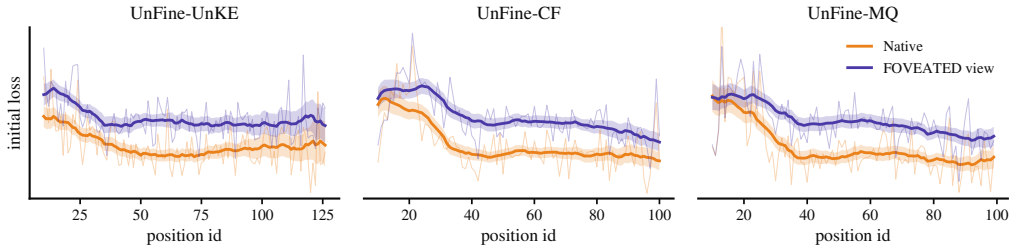


Figure 1: Pre-edit NLL loss in UnFine passages. Native loss gradually decreases toward later passage positions, and FOVEATED views yield consistently higher losses at the same positions.

where $\pi(x_t \mid \boldsymbol{x}_{<t})$ is the predicted probability assigned to token x_t given its prefix $\boldsymbol{x}_{<t}$. To generate text, the LLM sequentially draws x_t from $\pi(\cdot \mid \boldsymbol{x}_{<t})$ and appends it to the prefix, until an end-of-sequence token is produced, or the length limit is reached (Bengio et al., 2000).

Knowledge Editing (KE) modifies selected knowledge in a pre-trained LLM while preserving knowledge outside the scope of the edit (Wang et al., 2023; Zhang et al., 2024b). Structured KE represents the knowledge to be updated as a factual triple *(subject, relation, object)*, with the edit replacing the object with a new value. The edit is typically specified by a prompt-response pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ identifies the subject and relation, and $\boldsymbol{y}$ the updated object. *Unstructured KE (UKE)* instead associates $\boldsymbol{x}$ with a free-form passage $\boldsymbol{c}$ rather than explicitly annotated triples (Wu et al., 2024a; Yang et al., 2025). A passage may express multiple facts, each a separately queryable *atomic fact* (Wang et al., 2026). We focus on atomic-fact recall in UKE.

2.2 Learning Difficulty under Teacher Forcing

UKE aims to edit multiple new facts encoded in a free-form passage. However, as revealed by Zhou et al. (2026); Wang et al. (2026), many UKE methods encounter a *context reliance* failure mode: recall of the edited knowledge depends heavily on the context supplied during editing. When that context is absent, the edited LLMs struggle to recall the new knowledge, especially facts appearing later in the passage. Such failure limits the effectiveness of UKE beyond the original editing context.

Whereas this pattern appears to be a contextual "overfitting" problem of KE (Zhang et al., 2024a; Liu et al., 2025a), we show that it is analogous to "underfitting": the presence of contextual support makes new knowledge appear easy to predict even when the LLM has not sufficiently learned it.

Specifically, although (U)KE methods differ in the variables they optimize to incorporate new knowledge (Fang et al., 2025; Jiang et al., 2025), their supervision often comes from fitting the editing passage through a negative log-likelihood (NLL) loss (Liu et al., 2025a). Let θ denote the editor-specific optimization variables and π_θ the corresponding edited model. Given prompt $\boldsymbol{x}$ and passage $\boldsymbol{c}$, a common full-passage objective is

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{|\boldsymbol{c}|}\sum_{t=1}^{|\boldsymbol{c}|} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}), \tag{1}$$

where the ground-truth $\boldsymbol{c}_{<t}$ is provided to predict token c_t, a procedure called *teacher forcing* (Lamb et al., 2016). Intuitively, the prefix provides increasingly rich context as the passage unfolds. Consequently, even if the LLM does not yet hold the new knowledge, it can still see the fact as a plausible continuation of the surrounding text. This contextual shortcut can make the loss understate the difficulty of acquiring the knowledge. We call this effect *context-induced difficulty underestimation*.

To examine the effect empirically, we evaluate Qwen2.5-7B-Instruct (Qwen Team, 2024) on free-form passages from UnFine (Wang et al., 2026) and measure the initial token-level NLL before any editing updates. Although these passages are designed to contain new knowledge throughout (Deng et al., 2025; Wang et al., 2026), Fig. 1 shows a clear decrease in initial loss toward later positions. Later content thus appears easier to predict despite still introducing new knowledge, consistent with the growing prefix providing contextual support before the model has learned anything from the edit. To test whether intrinsic sentence difficulty explains the decline, we randomly shuffle the sentence order in each passage and observe the same pattern (App. D.1).

To address this issue, we introduce the plug-and-play method FOVEATED in the next section.

3 Proposed Method

We propose FOVEATED (FOcused ViEws of ATomic facts in unstructured knowledge EDiting) and integrate it with multiple UKE editors.

3.1 Breaking Context Shortcuts through Positional Encoding Perturbation

Given the difficulty underestimation induced by contextual shortcuts, we propose to make the context less reliable as a shortcut during the editing phase, so that the fact needs to be learned consistently across randomized views of the surrounding context.

Inspired by how LLMs use positional encoding to understand context (Vaswani et al., 2017; Press et al., 2021), and how manipulating positional encoding during fine-tuning can help them handle contexts longer than those seen during training (Zhu et al., 2024; Peng et al., 2024), FOVEATED randomly perturbs selected positional encodings to break the context shortcut in a fine-grained way. Intuitively, when positional information is perturbed, the shortcut becomes less reliable, encouraging the LLM to learn the new knowledge through a more focused view of the current content.

We focus on Rotary Position Embedding (RoPE; Su et al., 2021), which is widely used in modern LLMs (Dubey et al., 2024; Qwen Team, 2024; Liu et al., 2024). Given an attention head of even dimension d_h, let $\boldsymbol{h}_m$ denote the input hidden state at position m. The attention head maps $\boldsymbol{h}_m$ into d_h-dimensional query and key vectors via $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$, respectively. To incorporate relative-position information into attention score computation, RoPE rotates the query and key vectors based on their positions before computing the attention score. The query and key vectors lie in $\mathbb{C}^{d_h/2}$, with coordinate j rotated by $m\omega_j$ at position m. Let $\boldsymbol{\Theta} = \mathrm{diag}(\omega_1, \ldots, \omega_{d_h/2})$ collect the rotation frequencies. The rotated query and key vectors are given by

$$\boldsymbol{q}_t = e^{\mathrm{i}t\boldsymbol{\Theta}}\boldsymbol{W}_Q\boldsymbol{h}_t, \qquad \boldsymbol{k}_s = e^{\mathrm{i}s\boldsymbol{\Theta}}\boldsymbol{W}_K\boldsymbol{h}_s,$$

where i is the imaginary unit, and $\boldsymbol{q}_t$ and $\boldsymbol{k}_s$ denote the rotated query and key at positions t and s. The attention score is then given by

$$a_{ts} = \mathrm{Re}(\boldsymbol{q}_t^*\boldsymbol{k}_s)/\sqrt{d_h} = \mathrm{Re}\left((\boldsymbol{W}_Q\boldsymbol{h}_t)^* e^{\mathrm{i}(s-t)\boldsymbol{\Theta}}(\boldsymbol{W}_K\boldsymbol{h}_s)\right)/\sqrt{d_h},$$

where * and $\mathrm{Re}(\cdot)$ denote conjugate transpose and the real part, respectively. The term $e^{\mathrm{i}(s-t)\boldsymbol{\Theta}}$ encodes the *relative position* $s-t$, allowing the attention score to capture distance information.

Using this notation, FOVEATED randomly perturbs the key-side RoPE. For each target sentence, let $\mathcal{B}$ contain the prompt and each preceding sentence as separate context blocks. We independently sample one offset for every block $b \in \mathcal{B}$,

$$\epsilon_b \sim \mathcal{N}(0, \sigma^2), \qquad D_b = \mathrm{round}(\epsilon_b),$$

where σ controls the perturbation scale. Every eligible key position s in block b uses $d_s = D_b$, while $d_s = 0$ in the target sentence. We then apply these offsets to the context keys while keeping all query rotations and the target sentence's key rotations unchanged:

$$\tilde{\boldsymbol{q}}_t = \boldsymbol{q}_t, \qquad \tilde{\boldsymbol{k}}_s = e^{\mathrm{i}(s+d_s)\boldsymbol{\Theta}}\boldsymbol{W}_K\boldsymbol{h}_s = e^{\mathrm{i}d_s\boldsymbol{\Theta}}\boldsymbol{k}_s,$$
$$\tilde{a}_{ts} = \mathrm{Re}\left((\boldsymbol{W}_Q\boldsymbol{h}_t)^* e^{\mathrm{i}(s+d_s-t)\boldsymbol{\Theta}}(\boldsymbol{W}_K\boldsymbol{h}_s)\right)/\sqrt{d_h}.$$

Each block offset is shared across attention layers and resampled at every step. As shown in Fig. 1, the resulting focused views generally exhibit higher initial losses than the native view. The perturbation is removed after editing, and the LLM returns to its original RoPE.

The following theorem gives a mean-margin and view-consistency decomposition of FOVEATED.

Theorem 3.1 (Mean-margin and view-consistency decomposition)**.** *Let m_0 and m_D denote the native and perturbed margins of a target token, where D denotes the sampled positional offsets, with $\ell(m) = \log(1 + e^{-m})$. Under the integrability condition $\mathbb{E}_D|m_D| < \infty$, we have*

$$\mathbb{E}_D[\ell(m_D)] = \ell(m_0 - \Delta) + \mathcal{R}, \qquad \mathcal{R} \geq 0,$$

where $\Delta = m_0 - \mathbb{E}_D[m_D]$ *is the change in mean margin and* $\mathcal{R}$ *is an expected Bernoulli KL divergence that vanishes if and only if the correct-token confidence is constant across views. Under additional positive-shortcut assumptions (App. A.4),* $\Delta \geq 0$ *holds and hence* $\mathbb{E}_D[\ell(m_D)] \geq \ell(m_0)$.

In words, FOVEATED aims to learn a fact prediction that is stable across randomized positional views of the context: $\mathcal{R}$ penalizes view-dependent confidence, and, under a positive shortcut, the term Δ additionally reduces the mean contextual support. The complete version and proofs are in App. A.

3.2 Integrating FOVEATED with Existing Editors

We integrate FOVEATED with two editor families: methods that directly optimize selected parameters (Zhang et al., 2024b; Zhou et al., 2026), and locate-then-edit methods that write hidden targets into model weights (Fang et al., 2025; Jiang et al., 2025).

Direct Optimization (DO) editors update selected parameters to minimize the passage NLL. Assume the passage contains M sentences, with $\mathcal{I}_m$ denoting the token positions of sentence m. FOVEATED reorganizes the passage NLL into sentence-level losses and applies key perturbations to each sentence's preceding context. For sentence m, the FOVEATED-induced loss is given by

$$\ell_m(\theta, \boldsymbol{D}_m) = -\frac{1}{|\mathcal{I}_m|} \sum_{t \in \mathcal{I}_m} \log \pi_\theta\big(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}; \widetilde{\boldsymbol{K}}(\boldsymbol{D}_m)\big), \tag{2}$$

where $\widetilde{\boldsymbol{K}}(\boldsymbol{D}_m)$ denotes the keys randomly perturbed by offsets $\boldsymbol{D}_m$ in the context preceding sentence m, as defined in Sec. 3.1. The final objective is a weighted expectation of sentence-level losses:

$$\mathcal{L}_{\text{FOVEATED}}(\theta) = \sum_{m=1}^{M} \frac{|\mathcal{I}_m|}{|\boldsymbol{c}|} \mathbb{E}_{\boldsymbol{D}_m}[\ell_m(\theta, \boldsymbol{D}_m)]. \tag{3}$$

As shown in App. A, using zero positional offsets in all context blocks reduces Eq. (3) to Eq. (1). We keep the original DO editors' trainable parameters and optimization settings unchanged.

Locate-Then-Edit (LTE) editors first determine how updating some hidden states can lead to the desired knowledge update, then write the updates into selected model layers using either closed-form solvers or gradient descent on intermediate-state reconstruction losses (Deng et al., 2025; Jiang et al., 2025; Su et al., 2025). Because LTE minimizes the NLL over intermediate states rather than model parameters, adapting FOVEATED requires an additional weight-writing step.

We therefore use a two-stage LTE adaptation, following the structure of Wang et al. (2026): sentence-level targets are first written into shallow MLP layers, after which the base LTE editor learns the complete passage in deeper layers. Unlike FABLE, FOVEATED derives the sentence targets directly from the editing passage under focused views, without externally synthesized QA pairs.

Stage 1. Sentence-level LTE. For sentence m, we learn a trainable residual δ_m that is added to the MLP outputs at prompt positions $\mathcal{A}_m$ at a shallow layer z[1]. We optimize the residual on the sentence's tokens within the full teacher-forced passage, with key perturbations applied to its preceding context:

$$\delta_m^\star = \arg\min_{\delta_m} \mathbb{E}_{\boldsymbol{D}_m}\left[-\frac{1}{|\mathcal{I}_m|} \sum_{t \in \mathcal{I}_m} \log \pi_{\theta_0, \delta_m}\big(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}; \widetilde{\boldsymbol{K}}(\boldsymbol{D}_m)\big)\right]. \tag{4}$$

After optimizing all sentence-level residuals $[\delta_1, \ldots, \delta_M]$, we jointly write the resulting targets into the shallow MLP weights using a closed-form solver (Fang et al., 2025).

Stage 2. Passage-level LTE. Starting from the above model, we apply the base LTE editor and update deeper layers. Following Wang et al. (2026), we penalize changes in the hidden states at sentence-level anchor positions to preserve the Stage 1 edits by adding an MSE term to the original editing objective. See Appendices B.3 and C.1 for further implementation details.

4 Experiments

We evaluate FOVEATED with five UKE editors spanning the DO and LTE families.

[1]These anchors include a temporary sentence-specific carrier; see App. B.2 for construction details.

4.1 Experimental Setup

Base Models. We select two LLMs widely used in UKE studies (Jiang et al., 2025; Su et al., 2025; Zhou et al., 2026; Wang et al., 2026): *Qwen2.5-7B-Instruct* (Qwen Team, 2024) and *Llama-3.1-8B-Instruct* (Dubey et al., 2024), which we refer to as *Qwen2.5* and *Llama3.1*, respectively.

Tasks. We use *UnFine* (Wang et al., 2026), which augments existing benchmarks to enable finer-grained evaluation of UKE. It contains three subsets: (1) *UnFine-UnKE* extends UnKEBench (Deng et al., 2025); (2) *UnFine-CF* and (3) *UnFine-MQ* extend the CounterFact (Meng et al., 2022) and MQuAKE (Zhong et al., 2023) subsets of Wu et al. (2024a), respectively. We adopt the single-request setting from prior UKE work (Deng et al., 2025; Wang et al., 2026).

Editing Methods. We evaluate FOVEATED on five editing methods spanning both DO and LTE families. Among the DO editors, *FT-M* (Zhang et al., 2024b) fine-tunes the MLP layer identified by causal tracing, *LoRA* (Hu et al., 2022) learns additive low-rank updates to model parameters, and a simplified variant of *COIN** (Zhou et al., 2026) fine-tunes the model on the passage while regularizing against context reliance[2]. For the LTE family, we study *μKE* (Su et al., 2025) and *UnKE* (Deng et al., 2025), two representative UKE editors. All paired comparisons hold supervision fixed to the editing query and passage; we therefore do not include FABLE, which additionally uses externally synthesized atomic-QA supervision. Implementation details are provided in App. C. We report computation overhead of FOVEATED in App. D.5. FOVEATED adds moderate editing overhead but no inference cost.

Evaluation. Following UnFine (Wang et al., 2026), we assess both holistic passage recall and atomic-fact recall. Specifically, for *holistic recall*, we report *ROUGE-L recall*, which measures lexical overlap through the longest common subsequence (Lin, 2004), under the original editing query (*Orig Q*) and its paraphrase (*Reph Q*), together with their mean (*Avg.*). For *atomic-fact recall*, we separately query the model with atomic questions (*Atom Q*) targeting the individual facts in each passage. We report *ROUGE-L*, *Hit Rate (HR)*, and $\boldsymbol{C}_{\mathrm{LCS}}$. *HR* measures complete key-phrase recovery, while $\boldsymbol{C}_{\mathrm{LCS}}$ measures partial reference-answer coverage (Wang et al., 2026). We further evaluate *general capability preservation* on six standard benchmarks after editing on UnFine-UnKE. More details are provided in App. C.2.

4.2 Overall Performance

Tab. 1 reports the editing performance of FOVEATED on the three UnFine benchmarks.

Although base editors maintain high editing performance under Orig Q, their Atom Q performance is markedly worse, showing that passage-level recall does not necessarily translate into access to the individual facts. Applying FOVEATED consistently improves atomic-fact recall. For example, atomic ROUGE-L improves in all 30 scenarios by 4.4 points on average. Similar trends hold for HR and $\boldsymbol{C}_{\mathrm{LCS}}$, which improve in 29 out of 30 scenarios, with the only decreases below two points. Moreover, the improvement of FOVEATED does not come at the cost of holistic editing performance. Orig Q scores remain largely unchanged, while Reph Q improves in 19 of the 20 applicable scenarios by 8.4 points on average, indicating more robust holistic recall under query reformulation.

The improvements hold across both editor families. For DO editors, atomic ROUGE-L increases by 10.3% on average, with similar gains in HR and $\boldsymbol{C}_{\mathrm{LCS}}$ across every setting. Notably, COIN* still benefits from FOVEATED despite already including a context-alignment objective, suggesting that the two mechanisms are complementary. LTE editors show the same overall trend, with a 12.9% average gain in atomic ROUGE-L and only two minor decreases in the other metrics. These results show that FOVEATED benefits both editing paradigms despite their different update mechanisms.

4.3 Atomic-Fact Recall by Passage Position

Fig. 2 reports average atomic-fact recall by question index, which approximately tracks the corresponding fact's position; complete results are shown in App. D.2.

[2]Our COIN* adopts its original context-alignment objective to test whether FOVEATED offers complementary gains against context reliance. Other general-purpose locality-preserving mechanisms are omitted to keep the comparison focused.

Table 1: Overall editing performance (mean ± SE); higher is better. Gains over the base editors are shaded. Missing Reph Q entries are "–".

		Qwen2.5						Llama3.1					
		Holistic (ROUGE-L)			Atom Q			Holistic (ROUGE-L)			Atom Q		
	Method	Orig Q	Reph Q	Avg.	ROUGE-L	Hit Rate	C_{LCS}	Orig Q	Reph Q	Avg.	ROUGE-L	Hit Rate	C_{LCS}
UnFine-UnKE	**LoRA**	$100.0_{\pm 0.0}$	$95.3_{\pm 0.5}$	$97.6_{\pm 0.3}$	$40.8_{\pm 0.9}$	$40.2_{\pm 1.2}$	$51.9_{\pm 0.9}$	$100.0_{\pm 0.0}$	$96.5_{\pm 0.4}$	$98.2_{\pm 0.2}$	$39.2_{\pm 0.8}$	$43.8_{\pm 1.1}$	$57.1_{\pm 0.9}$
	+Ours	$99.7_{\pm 0.2}$	$\mathbf{99.1}_{\pm 0.2}$	$\mathbf{99.4}_{\pm 0.2}$	$\mathbf{53.4}_{\pm 0.8}$	$\mathbf{61.8}_{\pm 1.0}$	$\mathbf{69.9}_{\pm 0.8}$	$99.9_{\pm 0.1}$	$\mathbf{99.0}_{\pm 0.2}$	$\mathbf{99.4}_{\pm 0.1}$	$\mathbf{45.4}_{\pm 0.7}$	$\mathbf{56.8}_{\pm 1.0}$	$\mathbf{66.5}_{\pm 0.8}$
	COIN*	$100.0_{\pm 0.0}$	$97.4_{\pm 0.4}$	$98.7_{\pm 0.2}$	$36.8_{\pm 0.7}$	$34.7_{\pm 1.0}$	$49.0_{\pm 0.8}$	$100.0_{\pm 0.0}$	$96.5_{\pm 0.5}$	$98.3_{\pm 0.3}$	$36.8_{\pm 0.7}$	$36.7_{\pm 1.1}$	$53.2_{\pm 0.8}$
	+Ours	$99.3_{\pm 0.2}$	$\mathbf{98.4}_{\pm 0.3}$	$\mathbf{98.9}_{\pm 0.2}$	$\mathbf{43.9}_{\pm 0.7}$	$\mathbf{47.8}_{\pm 1.0}$	$\mathbf{59.2}_{\pm 0.8}$	$100.0_{\pm 0.0}$	$\mathbf{98.2}_{\pm 0.3}$	$\mathbf{99.1}_{\pm 0.2}$	$\mathbf{41.2}_{\pm 0.8}$	$\mathbf{45.7}_{\pm 1.1}$	$\mathbf{60.2}_{\pm 0.8}$
	FT-M	$99.9_{\pm 0.1}$	$78.3_{\pm 1.4}$	$89.1_{\pm 0.7}$	$23.2_{\pm 0.6}$	$16.1_{\pm 0.8}$	$30.0_{\pm 0.7}$	$100.0_{\pm 0.0}$	$96.6_{\pm 0.5}$	$98.3_{\pm 0.2}$	$35.3_{\pm 0.7}$	$34.8_{\pm 1.0}$	$51.6_{\pm 0.8}$
	+Ours	$99.2_{\pm 0.2}$	$\mathbf{81.5}_{\pm 1.2}$	$\mathbf{90.3}_{\pm 0.6}$	$\mathbf{26.8}_{\pm 0.7}$	$\mathbf{21.1}_{\pm 0.9}$	$\mathbf{35.1}_{\pm 0.8}$	$100.0_{\pm 0.0}$	$\mathbf{98.5}_{\pm 0.2}$	$\mathbf{99.2}_{\pm 0.1}$	$\mathbf{40.9}_{\pm 0.8}$	$\mathbf{45.3}_{\pm 1.1}$	$\mathbf{59.8}_{\pm 0.8}$
	μKE	$98.8_{\pm 0.3}$	$80.8_{\pm 1.1}$	$89.8_{\pm 0.6}$	$24.4_{\pm 0.6}$	$14.8_{\pm 0.7}$	$33.8_{\pm 0.7}$	$97.5_{\pm 0.5}$	$90.9_{\pm 0.8}$	$94.2_{\pm 0.6}$	$25.9_{\pm 0.7}$	$19.6_{\pm 0.9}$	$37.6_{\pm 0.8}$
	+Ours	$\mathbf{99.5}_{\pm 0.2}$	$\mathbf{83.0}_{\pm 1.0}$	$\mathbf{91.3}_{\pm 0.5}$	$\mathbf{29.9}_{\pm 0.6}$	$\mathbf{24.2}_{\pm 0.9}$	$\mathbf{43.5}_{\pm 0.8}$	$\mathbf{99.8}_{\pm 0.1}$	$\mathbf{93.0}_{\pm 0.6}$	$\mathbf{96.4}_{\pm 0.3}$	$\mathbf{29.3}_{\pm 0.6}$	$\mathbf{23.6}_{\pm 0.9}$	$\mathbf{42.7}_{\pm 0.7}$
	UnKE	$89.6_{\pm 0.8}$	$66.4_{\pm 1.1}$	$78.0_{\pm 0.8}$	$24.8_{\pm 0.6}$	$14.5_{\pm 0.7}$	$33.8_{\pm 0.7}$	$89.7_{\pm 0.9}$	$82.1_{\pm 1.1}$	$85.9_{\pm 0.9}$	$23.3_{\pm 0.6}$	$16.7_{\pm 0.8}$	$34.3_{\pm 0.8}$
	+Ours	$87.1_{\pm 0.8}$	$\mathbf{70.6}_{\pm 1.0}$	$\mathbf{78.8}_{\pm 0.8}$	$\mathbf{29.7}_{\pm 0.6}$	$\mathbf{22.7}_{\pm 0.9}$	$\mathbf{42.7}_{\pm 0.8}$	$\mathbf{93.9}_{\pm 0.6}$	$\mathbf{85.4}_{\pm 0.8}$	$\mathbf{89.7}_{\pm 0.6}$	$\mathbf{29.5}_{\pm 0.6}$	$\mathbf{25.6}_{\pm 0.9}$	$\mathbf{43.7}_{\pm 0.8}$
UnFine-CF	**LoRA**	$99.9_{\pm 0.1}$	$66.5_{\pm 1.5}$	$83.2_{\pm 0.7}$	$56.7_{\pm 0.9}$	$61.4_{\pm 1.0}$	$68.4_{\pm 0.9}$	$100.0_{\pm 0.0}$	$70.5_{\pm 1.3}$	$85.2_{\pm 0.7}$	$54.1_{\pm 0.9}$	$71.2_{\pm 1.0}$	$77.8_{\pm 0.8}$
	+Ours	$99.7_{\pm 0.1}$	$\mathbf{95.9}_{\pm 0.5}$	$\mathbf{97.8}_{\pm 0.3}$	$\mathbf{63.6}_{\pm 0.9}$	$\mathbf{76.5}_{\pm 0.9}$	$\mathbf{80.8}_{\pm 0.7}$	$99.8_{\pm 0.1}$	$\mathbf{92.3}_{\pm 0.8}$	$\mathbf{96.0}_{\pm 0.4}$	$\mathbf{56.2}_{\pm 0.9}$	$\mathbf{77.2}_{\pm 0.9}$	$\mathbf{81.9}_{\pm 0.7}$
	COIN*	$99.9_{\pm 0.1}$	$73.1_{\pm 1.5}$	$86.5_{\pm 0.7}$	$52.4_{\pm 1.0}$	$57.7_{\pm 1.0}$	$65.2_{\pm 0.9}$	$100.0_{\pm 0.0}$	$60.7_{\pm 1.5}$	$80.4_{\pm 0.8}$	$51.8_{\pm 0.9}$	$63.3_{\pm 1.0}$	$71.2_{\pm 0.8}$
	+Ours	$99.2_{\pm 0.2}$	$\mathbf{85.1}_{\pm 1.2}$	$\mathbf{92.1}_{\pm 0.6}$	$\mathbf{55.1}_{\pm 0.9}$	$\mathbf{60.9}_{\pm 1.0}$	$\mathbf{68.3}_{\pm 0.9}$	$100.0_{\pm 0.1}$	$\mathbf{78.2}_{\pm 1.4}$	$\mathbf{89.1}_{\pm 0.7}$	$\mathbf{53.6}_{\pm 0.9}$	$\mathbf{67.3}_{\pm 1.0}$	$\mathbf{74.6}_{\pm 0.8}$
	FT-M	$100.0_{\pm 0.0}$	$28.4_{\pm 0.9}$	$64.2_{\pm 0.4}$	$39.0_{\pm 0.9}$	$36.5_{\pm 1.0}$	$45.8_{\pm 1.0}$	$100.0_{\pm 0.0}$	$59.4_{\pm 1.5}$	$79.7_{\pm 0.8}$	$50.9_{\pm 0.9}$	$62.1_{\pm 1.0}$	$70.4_{\pm 0.8}$
	+Ours	$98.9_{\pm 0.3}$	$\mathbf{37.4}_{\pm 1.0}$	$\mathbf{68.2}_{\pm 0.5}$	$\mathbf{44.1}_{\pm 0.9}$	$\mathbf{42.5}_{\pm 1.1}$	$\mathbf{51.8}_{\pm 1.0}$	$100.0_{\pm 0.0}$	$\mathbf{78.0}_{\pm 1.4}$	$\mathbf{89.0}_{\pm 0.7}$	$\mathbf{53.2}_{\pm 0.9}$	$\mathbf{67.1}_{\pm 1.0}$	$\mathbf{74.4}_{\pm 0.8}$
	μKE	$98.7_{\pm 0.3}$	$31.5_{\pm 0.7}$	$65.1_{\pm 0.4}$	$33.1_{\pm 0.8}$	$30.1_{\pm 0.9}$	$41.9_{\pm 0.9}$	$98.1_{\pm 0.5}$	$50.7_{\pm 1.2}$	$74.4_{\pm 0.7}$	$38.8_{\pm 0.8}$	$42.3_{\pm 1.0}$	$53.2_{\pm 0.9}$
	+Ours	$\mathbf{99.6}_{\pm 0.1}$	$\mathbf{47.4}_{\pm 0.9}$	$\mathbf{73.5}_{\pm 0.5}$	$\mathbf{38.4}_{\pm 0.8}$	$\mathbf{38.5}_{\pm 1.0}$	$\mathbf{50.2}_{\pm 0.9}$	$\mathbf{100.0}_{\pm 0.0}$	$49.3_{\pm 1.1}$	$\mathbf{74.6}_{\pm 0.6}$	$\mathbf{39.3}_{\pm 0.8}$	$40.5_{\pm 1.0}$	$51.8_{\pm 0.9}$
	UnKE	$85.8_{\pm 0.9}$	$30.0_{\pm 0.7}$	$57.9_{\pm 0.6}$	$31.7_{\pm 0.8}$	$29.5_{\pm 1.0}$	$41.2_{\pm 0.9}$	$94.2_{\pm 0.7}$	$41.3_{\pm 1.1}$	$67.8_{\pm 0.7}$	$35.7_{\pm 0.8}$	$37.0_{\pm 1.0}$	$48.3_{\pm 0.9}$
	+Ours	$\mathbf{89.5}_{\pm 0.8}$	$\mathbf{44.0}_{\pm 0.8}$	$\mathbf{66.8}_{\pm 0.6}$	$\mathbf{38.6}_{\pm 0.9}$	$\mathbf{38.2}_{\pm 1.0}$	$\mathbf{50.0}_{\pm 0.9}$	$93.9_{\pm 0.6}$	$\mathbf{47.2}_{\pm 1.1}$	$\mathbf{70.6}_{\pm 0.7}$	$\mathbf{38.7}_{\pm 0.8}$	$\mathbf{40.7}_{\pm 1.0}$	$\mathbf{51.4}_{\pm 0.9}$
UnFine-MQ	**LoRA**	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$55.7_{\pm 1.1}$	$58.2_{\pm 1.3}$	$65.9_{\pm 1.1}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$50.2_{\pm 1.1}$	$65.8_{\pm 1.3}$	$72.8_{\pm 1.1}$
	+Ours	$99.4_{\pm 0.2}$	–	$99.4_{\pm 0.2}$	$\mathbf{63.3}_{\pm 1.1}$	$\mathbf{72.3}_{\pm 1.1}$	$\mathbf{78.1}_{\pm 0.9}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$\mathbf{54.7}_{\pm 1.1}$	$\mathbf{73.4}_{\pm 1.1}$	$\mathbf{78.4}_{\pm 1.0}$
	COIN*	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$53.0_{\pm 1.1}$	$55.1_{\pm 1.3}$	$62.8_{\pm 1.1}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$49.2_{\pm 1.1}$	$57.0_{\pm 1.3}$	$65.1_{\pm 1.2}$
	+Ours	$99.6_{\pm 0.2}$	–	$99.6_{\pm 0.2}$	$\mathbf{56.0}_{\pm 1.1}$	$\mathbf{60.2}_{\pm 1.2}$	$\mathbf{67.8}_{\pm 1.1}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$\mathbf{51.2}_{\pm 1.1}$	$\mathbf{62.8}_{\pm 1.2}$	$\mathbf{70.0}_{\pm 1.1}$
	FT-M	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$42.0_{\pm 1.1}$	$39.3_{\pm 1.3}$	$48.1_{\pm 1.1}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$48.5_{\pm 1.1}$	$57.4_{\pm 1.3}$	$65.4_{\pm 1.2}$
	+Ours	$97.8_{\pm 0.5}$	–	$97.8_{\pm 0.5}$	$\mathbf{46.1}_{\pm 1.1}$	$\mathbf{44.0}_{\pm 1.3}$	$\mathbf{52.6}_{\pm 1.2}$	$100.0_{\pm 0.0}$	–	$100.0_{\pm 0.0}$	$\mathbf{50.8}_{\pm 1.2}$	$\mathbf{61.6}_{\pm 1.3}$	$\mathbf{68.6}_{\pm 1.1}$
	μKE	$98.5_{\pm 0.5}$	–	$98.5_{\pm 0.5}$	$35.7_{\pm 0.9}$	$33.1_{\pm 1.2}$	$45.1_{\pm 1.0}$	$98.5_{\pm 0.5}$	–	$98.5_{\pm 0.5}$	$35.9_{\pm 1.0}$	$40.5_{\pm 1.2}$	$50.0_{\pm 1.1}$
	+Ours	$\mathbf{99.6}_{\pm 0.2}$	–	$\mathbf{99.6}_{\pm 0.2}$	$\mathbf{38.1}_{\pm 1.0}$	$\mathbf{38.1}_{\pm 1.1}$	$\mathbf{48.4}_{\pm 1.0}$	$\mathbf{99.8}_{\pm 0.2}$	–	$\mathbf{99.8}_{\pm 0.2}$	$\mathbf{40.8}_{\pm 1.0}$	$\mathbf{41.5}_{\pm 1.1}$	$\mathbf{52.3}_{\pm 1.0}$
	UnKE	$84.9_{\pm 1.2}$	–	$84.9_{\pm 1.2}$	$33.6_{\pm 0.9}$	$32.8_{\pm 1.1}$	$43.8_{\pm 1.0}$	$92.3_{\pm 1.0}$	–	$92.3_{\pm 1.0}$	$36.1_{\pm 1.0}$	$38.2_{\pm 1.3}$	$49.3_{\pm 1.2}$
	+Ours	$\mathbf{85.6}_{\pm 1.1}$	–	$\mathbf{85.6}_{\pm 1.1}$	$\mathbf{35.5}_{\pm 0.9}$	$\mathbf{35.0}_{\pm 1.1}$	$\mathbf{46.2}_{\pm 1.0}$	$92.0_{\pm 0.8}$	–	$92.0_{\pm 0.8}$	$\mathbf{40.1}_{\pm 1.0}$	$\mathbf{41.1}_{\pm 1.2}$	$\mathbf{51.8}_{\pm 1.1}$

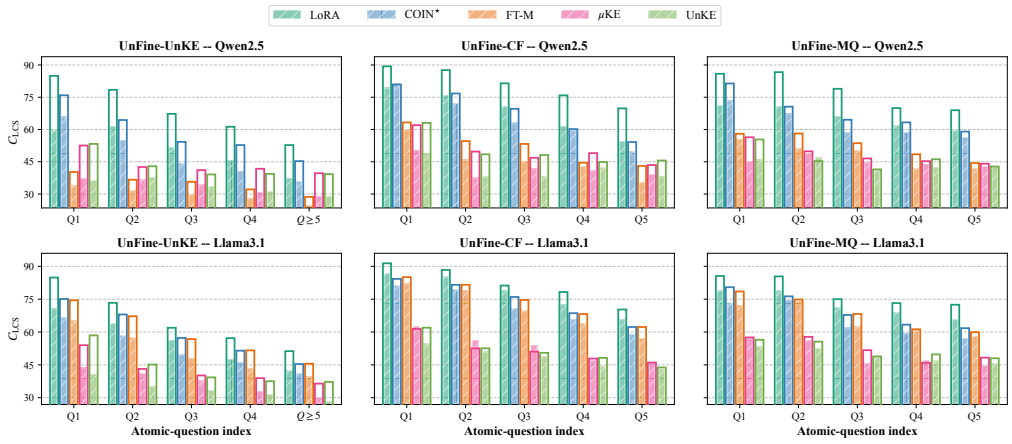


Figure 2: Atomic-fact recall by question position; higher is better. Solid and transparent bars show performance with and without FOVEATED. Unfilled area represents the performance gap.

Consistent with Zhou et al. (2026), atomic-fact recall declines toward later question positions. Together with the lower pre-edit losses observed at later passage positions (Fig. 1), these observations support our account of context-induced difficulty underestimation.

Moreover, applying FOVEATED raises recall for most editor–position pairs across the passage, with gains observed across all five editors, three benchmarks, and both LLM backbones. The improvement

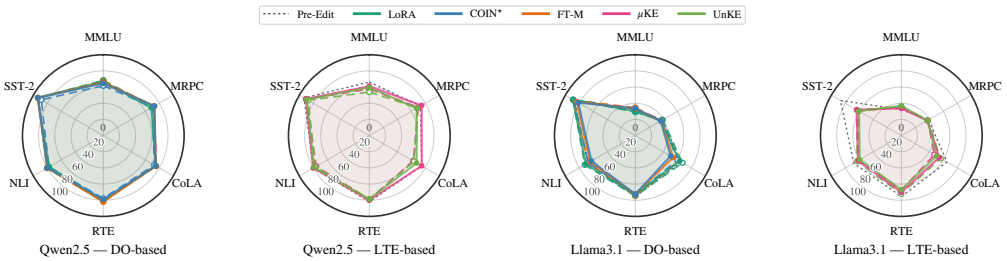


Figure 3: General capability (likelihood F1) after editing on UnFine-UnKE. Solid and dashed lines show performance with and without FOVEATED, respectively.

persists at later positions: the average gain at late $Q \geq 5$ is approximately 6.0 points, comparable to the gains observed from $Q2$ – $Q4$. The comparable gains at later positions show that focused views remain effective when longer preceding contexts provide stronger support under teacher forcing.

4.4 General Capability Preservation

We evaluate the edited models on six general-purpose language-understanding and reasoning tasks; dataset details are in App. C.3. We summarize the performance in Fig. 3; numbers are in Tab. 12.

Across the ten editor–backbone combinations, macro-average F1 changes relative to the base editors by at most 4.3 points in either direction. Performance remains stable or improves on Qwen2.5, with the clearest gains under LTE. On Llama3.1, μKE improves, whereas the DO editors and UnKE show moderate aggregate decreases driven by task-specific variation. The results therefore support broad preservation rather than uniform gains in general capability.

4.5 Ablation Studies

We ablate FOVEATED with FT-M (DO) and μKE (LTE). Tab. 2 examines the perturbation scheme shared by both paradigms and the two-stage writing strategy used for LTE.

Table 2: Ablations on Qwen2.5 using a subset of UnFine-UnKE.

	Holistic (ROUGE-L)		Atom Q		
Method	**Orig Q**	**Reph Q**	**ROUGE-L**	**Hit Rate**	C_{LCS}
FT-M (DO)	100.0	81.3	20.8	15.9	28.8
No Pert.	100.0	81.2	21.0	15.5	29.2
ID Pert.	100.0	83.3	22.1	16.1	30.3
+Ours	99.4	84.6	25.1	19.4	33.3
μKE (LTE)	98.5	82.1	24.6	15.7	35.9
No Pert.	99.6	70.3	25.1	17.9	37.5
ID Pert.	99.6	76.2	26.8	21.1	40.7
Sent.-only	40.1	38.5	25.1	21.4	40.3
Rev. order	48.9	48.9	29.1	22.1	44.7
+Ours	99.9	85.5	29.5	24.2	45.2

Perturbation scheme. Removing perturbation (No Pert.) but retaining the FOVEATED objective reduces atomic ROUGE-L for both editors; FT-M, for example, drops from 25.1 to 21.0. Perturbing shared position IDs (ID Pert.), as in Zhu et al. (2024) and Li et al. (2026), partially recovers the loss but trails the key-only design for both editors, supporting our asymmetric key perturbation. For space, we defer additional sensitivity analysis on the perturbation scale σ to App. D.4, where FOVEATED exhibits notable robustness to the choice of σ over editor-specific ranges.

LTE writing strategy. The *Sent.-only* variant removes passage-level optimization, whereas *Rev. order* retains both stages but reverses them. Both variants sharply reduce holistic recall; for example, removing the passage-level stage lowers Orig Q/Reph Q from $99.9/85.5$ to $40.1/38.5$. Reversing the order similarly hurts holistic recall while nearly preserving atomic ROUGE-L (29.1 vs. 29.5), showing that strong atomic recall alone does not preserve the edited passage as a coherent whole. The passage-level stage and its placement after sentence-level optimization are therefore both important.

4.6 Case Studies

We conclude this section with case studies on FOVEATED. Tab. 3 and App. D.6 present representative failures across editors and backbones. Base editors may hallucinate or revert to prior knowledge, resulting in UKE failures. Equipped with FOVEATED, the edited models instead recover the target atomic facts while giving fluent answers that remain relevant to the corresponding atomic questions.

Table 3: Atomic-fact recall case studies, with errors highlighted.

Model	Method	Dataset	Atomic query (target)	Base-editor output	+Ours output
Qwen2.5	FT-M	UnFine-CF	What year was Charles Alfred Pillsbury born? (1842)	**1850**	**1842**
	COIN*	UnFine-UnKE	At what age was Henry Feilden elected to the local council? (25)	Henry Feilden was elected to the local council at the age of **35**.	Henry Feilden was elected to the local council at the age of **25**.
	μKE	UnFine-CF	What region is Sheffield located in? (South Yorkshire)	Sheffield is a city located in the **northern region of England**.	Sheffield, located in **South Yorkshire**, England.
Llama3.1	LoRA	UnFine-UnKE	How many people does Herlyn Espinal employ in her company? (over 50 people)	Herlyn Espinal is a politician and **doesn't employ anyone** …	Herlyn Espinal is the founder and CEO … employing **over 50 people** …
	UnKE	UnFine-MQ	Which year was SC Atlético founded? (1903)	SC Atlético, a Portuguese sports club, was founded in **1919**, marking …	SC Atlético, a renowned sports club based in Madrid, Spain, was founded in **1903**.

5 Related Work

Knowledge Editing aims to update specific factual knowledge in a pretrained model while preserving unrelated knowledge and the model's general capabilities (Wang et al., 2023; Zhang et al., 2024b). Optimization-based approaches learn model updates with the editing objective, including selected layer optimization (Zhang et al., 2024b) or learning multiple copies (Wang et al., 2024b). Parameter-efficient methods have also been applied (Hu et al., 2022; Wang et al., 2024a; Wu et al., 2024b; Liu et al., 2025b). On the other hand, Locate-Then-Edit (LTE) methods first identify where a fact is stored, then compute the weight change often in closed form (Meng et al., 2022; 2023; Li et al., 2024; Fang et al., 2025). *Unstructured Knowledge Editing* (UKE) replaces the curated triple with a free-form passage (Wu et al., 2024a; Yang et al., 2025). UnKE makes transformer blocks editable (Deng et al., 2025), AnyEdit decomposes passages into sequential chunks (Jiang et al., 2025), and μKE coordinates sentence-level updates through a nested editing structure (Su et al., 2025). COIN identifies context reliance during editing and mitigates it by aligning predictions under global and local contexts (Zhou et al., 2026). FABLE uses externally synthesized atomic-QA supervision to write shallow-layer anchors before fitting the passage holistically, and introduces UnFine to measure atomic-fact access (Wang et al., 2026). In concurrent work, HPSE uses hybrid rollouts to recover missing facts with token-level feedback from a privileged in-context state (Liu et al., 2026). In contrast, our FOVEATED targets the learning signal by disrupting contextual support that can suppress gradients before a fact becomes independently retrievable.

Positional Encoding. RoPE encodes relative positions by rotating queries and keys (Su et al., 2021) and is widely used in modern LLMs (Dubey et al., 2024; Qwen Team, 2024; Liu et al., 2024). Long-context adaptation also manipulates positions: Peng et al. (2024) scale RoPE frequencies and attention, Zhu et al. (2024) shift the indices of entire chunks, and Li et al. (2026) shift suffix indices and distill the perturbed view from native-view predictions. These methods perturb positions to improve the model's tolerance to positional variation. FOVEATED instead uses perturbation for a fundamentally different purpose: by randomizing only preceding-context keys and supervising with ground truth rather than native-view predictions, it breaks contextual shortcuts and makes the edited fact robust across contexts.

6 Conclusion

In this work, we identify *context-induced difficulty underestimation* in UKE and propose FOVEATED, which constructs focused positional views to make contextual shortcuts less reliable during editing. Across five editors, two LLM backbones, and three benchmarks, FOVEATED broadly improves fine-grained fact recall while retaining holistic recall. Encouraged by these consistent improvements, we plan to extend FOVEATED to handle correlated, continually evolving unstructured knowledge, where new updates may reinforce, revise, or invalidate earlier ones. By tracking these dependencies and selectively preserving or revising previously edited knowledge, we aim to pave the way toward practical UKE in continuously changing real-world environments.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, including improving grammar, clarity, concision, and transitions. We did not use generative AI tools to develop the method or theory or formulate mathematical claims; they were only used to assist in the writing of proofs. We also used generative AI tools to assist with coding and debugging, but we did not use generative AI tools to develop the research ideas or methods. All AI-assisted code was reviewed and tested by the authors against the intended algorithms and experimental settings. We did not use generative AI tools to design experiments or interpret results beyond the disclosed evaluation; the remaining required tasks are not applicable. All AI-assisted revisions were reviewed by the authors to ensure that they faithfully preserved the intended meaning and accurately reflected the reported methods and results. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*, 2009.

Christopher M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Everything is editable: Extend knowledge editing to unstructured data in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

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

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 (ICLR)*, 2025.

Thomas 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 (NeurIPS)*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

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

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Mingyang Wan, Guojun Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AnyEdit: Edit any knowledge encoded in language models. In *International Conference on Machine Learning (ICML)*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and J ie Yu. PMET: Precise model editing in a transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 18564–18572, 2024.

Zichong Li, Chen Liang, Liliang Ren, Tuo Zhao, Yelong Shen, and Weizhu Chen. Shuffle the context: RoPE-perturbed self-distillation for long-context adaptation. *arXiv preprint arXiv:2604.14339*, 2026.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Tianci Liu, Ruirui Li, Zihan Dong, Hui Liu, Xianfeng Tang, Qingyu Yin, Linjun Zhang, Haoyu Wang, and Jing Gao. Mitigating heterogeneous token overfitting in LLM knowledge editing. In *International Conference on Machine Learning (ICML)*, 2025a.

Tianci Liu, Ruirui Li, Yunzhe Qi, Hui Liu, Xianfeng Tang, Tianqi Zheng, Qingyu Yin, Monica Xiao Cheng, Jun Huan, Haoyu Wang, and Jing Gao. Unlocking efficient, scalable, and continual knowledge editing with basis-level representation fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Tianci Liu, Zihan Dong, Tianchun Li, Yi-Chung Chen, Qiming Cao, Xingchen Wang, Shiyang Wang, Zichen Miao, Linjun Zhang, Haoyu Wang, and Jing Gao. Hybrid-policy self-editing for composable unstructured knowledge editing, 2026.

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

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Zian Su, Ziyang Huang, Kaiyuan Zhang, and Xiangyu Zhang. μKE: Matryoshka unstructured knowledge editing of large language models. In *Conference on Language Modeling (COLM)*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, 2013.

Haoyu Wang, Tianci Liu, Ruirui Li, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. RoseLoRA: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 996–1008, 2024a.

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

Peng Wang, Biyu Zhou, Xuehai Tang, Jizhong Han, and Songlin Hu. FABLE: Fine-grained fact anchoring for unstructured model editing. In *Findings of the Association for Computational Linguistics: ACL*, 2026.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*, 2023.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

Xiaobao Wu, Liangming Pan, William Y. Wang, and Anh T. Luu. AKEW: Assessing knowledge editing in the wild. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024a.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models, 2024b.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

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

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Uncovering overfitting in large language model editing. *arXiv preprint arXiv:2410.07819*, 2024a.

Ningyu Zhang et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024b.

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

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

Zisheng Zhou, Mengqi Zhang, Shiguang Wu, Xiaotian Ye, Chi Zhang, Zhumin Chen, and Pengjie Ren. Uncovering context reliance in unstructured knowledge editing. *arXiv preprint arXiv:2602.19043*, 2026.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *International Conference on Learning Representations (ICLR)*, 2024.

A Omitted Theoretical Analysis

In this section we develop the analysis omitted in the main body due to page limits.

A.1 Token-Weighted Aggregation Preserves the Passage Objective

Let the editing passage contain M sentences with token spans $\mathcal{I}_m$, $m = 1, \dots, M$, where $n_m = |\mathcal{I}_m|$ and $N = \sum_{m=1}^{M} n_m = |\boldsymbol{c}|$. Without perturbation, the token-averaged loss on sentence m is

$$\ell_m(\theta, \mathbf{0}) = -\frac{1}{n_m} \sum_{t \in \mathcal{I}_m} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}).$$

Proposition A.1 (Token-weighted sentence aggregation)**.** *As the sentence spans form a disjoint partition of the passage, weighting each sentence loss by n_m/N recovers the passage NLL exactly:*

$$\sum_{m=1}^{M} \frac{n_m}{N} \ell_m(\theta, \mathbf{0}) = -\frac{1}{N} \sum_{t=1}^{N} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}) = \mathcal{L}_{\mathrm{NLL}}(\theta). \tag{5}$$

Proof. We expand the weighted sum. By definition:

$$\begin{aligned}
\sum_{m=1}^{M} \frac{n_m}{N} \ell_m(\theta, \mathbf{0}) &= \frac{1}{N} \sum_{m=1}^{M} n_m \ell_m(\theta, \mathbf{0}) \\
&= -\frac{1}{N} \sum_{m=1}^{M} n_m \frac{1}{n_m} \sum_{t \in \mathcal{I}_m} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}) \\
&= -\frac{1}{N} \sum_{m=1}^{M} \sum_{t \in \mathcal{I}_m} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}) \\
&\overset{(a)}{=} -\frac{1}{N} \sum_{t=1}^{N} \log \pi_\theta(c_t \mid \boldsymbol{x}, \boldsymbol{c}_{<t}) \\
&= \mathcal{L}_{\mathrm{NLL}}(\theta).
\end{aligned}$$

The step (a) uses the fact that the disjoint sentence spans partition the passage. This completes our proof. □

A.2 Blockwise RoPE-Frequency Contraction

Following the RoPE convention (Su et al., 2021; Peng et al., 2024), we focus on a single attention head of even width $d_h = 2R$. Write $\bar{\boldsymbol{q}}_t = \boldsymbol{W}_Q \boldsymbol{h}_t$ and $\bar{\boldsymbol{k}}_s = \boldsymbol{W}_K \boldsymbol{h}_s$ for the pre-RoPE query and key in real coordinates, each partitioned into R two-dimensional components. RoPE at position p takes the form

$$\boldsymbol{R}_p = \mathrm{diag}\left(\mathrm{Rot}(\omega_1 p), \dots, \mathrm{Rot}(\omega_R p)\right),$$

where $\mathrm{Rot}(\alpha)$ denotes the 2×2 rotation matrix by angle α. Using the fact that $\mathrm{Rot}(\alpha)^\top = \mathrm{Rot}(-\alpha)$ and $\mathrm{Rot}(\alpha)\,\mathrm{Rot}(\beta) = \mathrm{Rot}(\alpha + \beta)$, we have

$$\boldsymbol{R}_t^\top \boldsymbol{R}_s = \mathrm{diag}\left(\mathrm{Rot}(\omega_1 (s-t)), \dots, \mathrm{Rot}(\omega_R (s-t))\right).$$

Therefore, the native query–key logit can be written as

$$a_{ts} = \frac{(\boldsymbol{R}_t \bar{\boldsymbol{q}}_t)^\top (\boldsymbol{R}_s \bar{\boldsymbol{k}}_s)}{\sqrt{d_h}} = \frac{1}{\sqrt{d_h}} \sum_{r=1}^{R} \bar{\boldsymbol{q}}_{t,r}^\top \mathrm{Rot}(\omega_r (s-t)) \bar{\boldsymbol{k}}_{s,r}.$$

Now fix a target sentence m. As outlined in Sec. 3.1, its preceding context is partitioned into separate blocks: the prompt and each preceding sentence. Let $\mathcal{P}$ denote the prompt token positions and $\mathcal{I}_j$ the token positions of sentence j. The set of context blocks for target sentence m is

$$\mathcal{B}_m \triangleq \{\mathcal{P}, \mathcal{I}_1, \dots, \mathcal{I}_{m-1}\}.$$

For each block $b \in \mathcal{B}_m$, let D_{mb} denote the scalar positional offset shared by its eligible key positions. Thus, m indexes the target sentence, while b identifies a block in that sentence's preceding context. We collect these block offsets into

$$\boldsymbol{D}_m \triangleq (D_{mb})_{b \in \mathcal{B}_m},$$

which is the offset vector used in Eqs. (2) and (3).

As will be specified in App. B.1 shortly, offsets are sampled independently across blocks at the same scale σ. Each sampled offset is shared across attention layers and resampled at every optimization step. An eligible key position $s \in b$ uses $d_s = D_{mb}$. The target sentence's key positions and the key at the readout position of its first token use $d_s = 0$, and all query rotations remain unchanged.

For an eligible key position $s \in b$, the perturbed logit is

$$\tilde{a}_{ts}(D_{mb}) = \frac{1}{\sqrt{d_h}} \sum_{r=1}^{R} \bar{\boldsymbol{q}}_{t,r}^{\top} \operatorname{Rot}\left(\omega_r (s - t + D_{mb})\right) \bar{\boldsymbol{k}}_{s,r}. \tag{6}$$

We show that the logit depends only on the offset D_{mb} of the block containing s, whereas $\boldsymbol{D}_m$ specifies the perturbations for the entire sentence-level forward pass.

Proposition A.2 (Blockwise spectral contraction)**.** *Fix a target sentence m, a context block $b \in \mathcal{B}_m$, and the pre-RoPE query and key representations in Eq. (6). Suppose that D_{mb} has a distribution symmetric about zero, and define*

$$\rho_{mbr} \triangleq \mathbb{E}_{D_{mb}}\left[\cos(\omega_r D_{mb})\right].$$

Then

$$\mathbb{E}_{D_{mb}}\left[\tilde{a}_{ts}(D_{mb})\right] = \frac{1}{\sqrt{d_h}} \sum_{r=1}^{R} \rho_{mbr} \bar{\boldsymbol{q}}_{t,r}^{\top} \operatorname{Rot}(\omega_r (s - t)) \bar{\boldsymbol{k}}_{s,r}, \qquad |\rho_{mbr}| \le 1. \tag{7}$$

For the rounded Gaussian used by FOVEATED, let $D_{mb} = \operatorname{round}(G_{mb})$ with $G_{mb} \sim \mathcal{N}(0, \sigma^2)$. For $\sigma > 0$, the exact probability mass function is

$$\Pr(D_{mb} = k) = \Phi\left(\frac{k + \frac{1}{2}}{\sigma}\right) - \Phi\left(\frac{k - \frac{1}{2}}{\sigma}\right), \qquad k \in \mathbb{Z}, \tag{8}$$

where Φ denotes the standard Gaussian cumulative distribution function. The corresponding spectral multiplier is

$$\rho_{mbr}^{\text{round}} = \sum_{k \in \mathbb{Z}} \Pr(D_{mb} = k) \cos(\omega_r k). \tag{9}$$

Moreover, $|\rho_{mbr}^{\text{round}}| < 1$ whenever $\sigma > 0$ and $\omega_r \notin 2\pi\mathbb{Z}$.

For the continuous Gaussian relaxation $D_{mb} \sim \mathcal{N}(0, \sigma^2)$, the multiplier simplifies to

$$\rho_{mbr}^{\text{cont}} = \exp\left(-\frac{\sigma^2 \omega_r^2}{2}\right), \tag{10}$$

which decreases monotonically with σ for every nonzero RoPE frequency. At $\sigma = 0$, both offset distributions concentrate at zero and their multipliers equal 1.

Proof. We first establish the result for symmetric offsets, then specialize it to the rounded Gaussian used by FOVEATED and its continuous Gaussian relaxation.

General result for symmetric offsets. Since the distribution of D_{mb} is symmetric about zero, D_{mb} and $-D_{mb}$ have the same distribution. Therefore,

$$\begin{aligned} \mathbb{E}\sin(\omega_r D_{mb}) &= \mathbb{E}\sin(-\omega_r D_{mb}) \\ &= -\mathbb{E}\sin(\omega_r D_{mb}), \end{aligned}$$

which implies $\mathbb{E}\sin(\omega_r D_{mb}) = 0$. Hence, the expected rotation is a scalar matrix:

$$\begin{aligned}\mathbb{E}\,\mathrm{Rot}(\omega_r D_{mb}) &= \begin{pmatrix} \mathbb{E}\cos(\omega_r D_{mb}) & -\mathbb{E}\sin(\omega_r D_{mb}) \\ \mathbb{E}\sin(\omega_r D_{mb}) & \mathbb{E}\cos(\omega_r D_{mb}) \end{pmatrix} \\ &= \rho_{mbr} \boldsymbol{I}_2.\end{aligned}$$

Using the composition rule for rotations,

$$\mathrm{Rot}\left(\omega_r(s-t+D_{mb})\right) = \mathrm{Rot}\left(\omega_r(s-t)\right)\mathrm{Rot}(\omega_r D_{mb}).$$

Substituting this identity into Eq. (6), with the pre-RoPE representations held fixed, gives

$$\begin{aligned}\mathbb{E}_{D_{mb}}\left[\tilde{a}_{ts}(D_{mb})\right] &= \frac{1}{\sqrt{d_h}} \sum_{r=1}^{R} \bar{\boldsymbol{q}}_{t,r}^{\top}\, \mathrm{Rot}(\omega_r(s-t))\mathbb{E}\left[\mathrm{Rot}(\omega_r D_{mb})\right]\bar{\boldsymbol{k}}_{s,r} \\ &= \frac{1}{\sqrt{d_h}} \sum_{r=1}^{R} \rho_{mbr}\bar{\boldsymbol{q}}_{t,r}^{\top}\, \mathrm{Rot}(\omega_r(s-t))\bar{\boldsymbol{k}}_{s,r}.\end{aligned}$$

Moreover,

$$\begin{aligned}|\rho_{mbr}| &= \left|\mathbb{E}\left[\cos(\omega_r D_{mb})\right]\right| \\ &\le \mathbb{E}\left[\left|\cos(\omega_r D_{mb})\right|\right] \\ &\le 1.\end{aligned}$$

Consequently, the operator norm satisfies

$$\left\|\mathbb{E}\left[\mathrm{Rot}(\omega_r D_{mb})\right]\right\|_{\mathrm{op}} = |\rho_{mbr}| \le 1.$$

Rounded Gaussian case. For $\sigma > 0$ and $k \in \mathbb{Z}$, nearest-integer rounding gives

$$\begin{aligned}\Pr(D_{mb} = k) &= \int_{k-1/2}^{k+1/2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{g^2}{2\sigma^2}\right) dg \\ &= \Phi\left(\frac{k+1/2}{\sigma}\right) - \Phi\left(\frac{k-1/2}{\sigma}\right).\end{aligned}$$

Since D_{mb} is integer-valued,

$$\begin{aligned}\rho_{mbr}^{\mathrm{round}} &= \mathbb{E}\left[\cos(\omega_r D_{mb})\right] \\ &= \sum_{k\in\mathbb{Z}} \Pr(D_{mb} = k)\cos(\omega_r k).\end{aligned}$$

The series is absolutely convergent because

$$\sum_{k\in\mathbb{Z}} \Pr(D_{mb} = k)\left|\cos(\omega_r k)\right| \le 1.$$

For strict contraction, the strictly positive Gaussian density at $\sigma > 0$ gives $\Pr(D_{mb} = 0) > 0$ and $\Pr(D_{mb} = 1) > 0$. If $\omega_r \notin 2\pi\mathbb{Z}$, then $\cos(\omega_r) < 1$, and hence we have

$$\begin{aligned}1 - \rho_{mbr}^{\mathrm{round}} &= \sum_{k\in\mathbb{Z}} \Pr(D_{mb} = k)\left[1 - \cos(\omega_r k)\right] \\ &\ge \Pr(D_{mb} = 1)\left[1 - \cos(\omega_r)\right] \\ &> 0.\end{aligned}$$

Similarly,

$$\begin{aligned}1 + \rho_{mbr}^{\mathrm{round}} &= \sum_{k\in\mathbb{Z}} \Pr(D_{mb} = k)\left[1 + \cos(\omega_r k)\right] \\ &\ge 2\Pr(D_{mb} = 0) \\ &> 0.\end{aligned}$$

Thus $-1 < \rho_{mbr}^{\text{round}} < 1$, proving $|\rho_{mbr}^{\text{round}}| < 1$.

Continuous relaxation case. Let $G \sim \mathcal{N}(0, \sigma^2)$ and write its characteristic function as $\varphi(\omega) = \mathbb{E}e^{\mathrm{i}\omega G}$. For $\sigma > 0$, differentiation under the expectation is justified by $\mathbb{E}|G| < \infty$. Let f_σ denote the density of $\mathcal{N}(0, \sigma^2)$. Since $g f_\sigma(g) = -\sigma^2 f_\sigma'(g)$, integration by parts gives

$$\begin{aligned}\varphi'(\omega) &= \int_{\mathbb{R}} \mathrm{i} g e^{\mathrm{i}\omega g} f_\sigma(g)\, dg \\ &= -\mathrm{i}\sigma^2 \int_{\mathbb{R}} e^{\mathrm{i}\omega g} f_\sigma'(g)\, dg \\ &= -\sigma^2 \omega \varphi(\omega),\end{aligned}$$

where the boundary term vanishes because the Gaussian density tends to zero at infinity. Together with $\varphi(0) = 1$, this differential equation yields

$$\varphi(\omega) = \exp\left(-\frac{\sigma^2\omega^2}{2}\right).$$

Therefore,

$$\rho_{mbr}^{\text{cont}} = \operatorname{Re}\varphi(\omega_r) = \exp\left(-\frac{\sigma^2\omega_r^2}{2}\right).$$

For $\omega_r \neq 0$ and $\sigma > 0$,

$$\frac{\partial}{\partial\sigma}\rho_{mbr}^{\text{cont}} = -\sigma\omega_r^2 \exp\left(-\frac{\sigma^2\omega_r^2}{2}\right) < 0,$$

which proves the stated monotonicity. When $\sigma = 0$, both the rounded and continuous offset distributions concentrate at zero, giving $\rho_{mbr} = 1$. This completes our proof. □

This proposition characterizes the direct key rotation for fixed pre-RoPE states; changes to intermediate representations under the full perturbed forward pass are outside its scope. The bound $|\rho_{mbr}| \leq 1$ concerns each expected RoPE-frequency component and does not imply that the softmax attention weight scales linearly by ρ_{mbr}. The perturbation directly modifies only key rotations, leaving query rotations, value vectors, and the causal mask unchanged.

A.3 Exact Loss Decomposition and Editing Gradients

Here we connect the proposed positional perturbation to the token-level learning signal. Fix a target token c_t in sentence m and omit the token index t for brevity. Let $\boldsymbol{D}_m = (D_{mb})_{b\in\mathcal{B}_m}$ denote the sentence's vector of context-block offsets, as defined in the preceding subsection.

For DO, let $z_{m,v}(\boldsymbol{D}_m;\theta)$ denote the pre-softmax logit for vocabulary token v at the prediction step for c_t. These logits are produced by the full perturbed forward pass, including the changes to intermediate representations induced by the offsets. The native and perturbed correct-versus-rest margins are

$$\begin{aligned}m_{m,0}(\theta) &\triangleq z_{m,c_t}(\mathbf{0};\theta) - \log\sum_{v\neq c_t}\exp\big(z_{m,v}(\mathbf{0};\theta)\big), \\ m_{m,\boldsymbol{D}_m}(\theta) &\triangleq z_{m,c_t}(\boldsymbol{D}_m;\theta) - \log\sum_{v\neq c_t}\exp\big(z_{m,v}(\boldsymbol{D}_m;\theta)\big).\end{aligned}$$

Both margins use the same current parameters θ; the native margin is evaluated with zero positional offsets. Then for a margin value u, the token cross-entropy and its negative derivative are

$$\begin{aligned}\ell(u) &= \log(1+\exp(-u)), \\ C(u) &\triangleq -\ell'(u) = \frac{1}{1+\exp(u)}.\end{aligned}$$

Exact loss decomposition. Define

$$\mu_m(\theta) \triangleq \mathbb{E}_{\boldsymbol{D}_m}\left[m_{m,\boldsymbol{D}_m}(\theta)\right],$$
$$p_{m,\boldsymbol{D}_m}(\theta) \triangleq \mathrm{sigmoid}\left(m_{m,\boldsymbol{D}_m}(\theta)\right),$$
$$p_{m,\mu}(\theta) \triangleq \mathrm{sigmoid}\left(\mu_m(\theta)\right).$$

Here $p_{m,\boldsymbol{D}_m}$ is the softmax probability assigned to the correct token, while $p_{m,\mu}$ is the probability corresponding to the mean margin. In general, $p_{m,\mu} \neq \mathbb{E}_{\boldsymbol{D}_m}[p_{m,\boldsymbol{D}_m}]$. We suppress the dependence on θ when it is clear from context. Then we have the following property.

Proposition A.3 (Mean-margin and view-consistency decomposition)**.** *Suppose that* $\mathbb{E}_{\boldsymbol{D}_m}|m_{m,\boldsymbol{D}_m}| < \infty$. *Define* $\Delta_m \triangleq m_{m,0} - \mu_m$, *the token-specific quantity denoted by* Δ *in Theorem 3.1. Then*

$$\mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m})\right] = \ell(\mu_m) + \mathcal{R}_m = \ell(m_{m,0} - \Delta_m) + \mathcal{R}_m, \tag{11}$$
$$\mathcal{R}_m \triangleq \mathbb{E}_{\boldsymbol{D}_m}\left[\mathrm{KL}\left(\mathrm{Bern}(p_{m,\mu}) \,\|\, \mathrm{Bern}(p_{m,\boldsymbol{D}_m})\right)\right] \geq 0.$$

Moreover, $\mathcal{R}_m = 0$ *if and only if* $p_{m,\boldsymbol{D}_m} = p_{m,\mu}$ *almost surely.*

Proof. All expectations in this proof are taken with respect to $\boldsymbol{D}_m$. Define

$$F(u) \triangleq \log(1 + \exp(u)).$$

The integrability assumption ensures that the expectations below are finite, since

$$F(u) \leq \log 2 + |u|, \qquad \ell(u) \leq \log 2 + |u|.$$

For $p(u) = \mathrm{sigmoid}(u)$, we have

$$\log p(u) = u - F(u),$$
$$\log(1 - p(u)) = -F(u).$$

Applying these identities at $u = m_{m,\boldsymbol{D}_m}$ and $u = \mu_m$ gives

$$\begin{aligned}
&\mathrm{KL}\left(\mathrm{Bern}(p_{m,\mu}) \,\|\, \mathrm{Bern}(p_{m,\boldsymbol{D}_m})\right) \\
&= p_{m,\mu} \log \frac{p_{m,\mu}}{p_{m,\boldsymbol{D}_m}} + (1 - p_{m,\mu}) \log \frac{1 - p_{m,\mu}}{1 - p_{m,\boldsymbol{D}_m}} \\
&= p_{m,\mu}\left[\mu_m - m_{m,\boldsymbol{D}_m} - F(\mu_m) + F(m_{m,\boldsymbol{D}_m})\right] \\
&\quad + (1 - p_{m,\mu})\left[F(m_{m,\boldsymbol{D}_m}) - F(\mu_m)\right] \\
&= F(m_{m,\boldsymbol{D}_m}) - F(\mu_m) - F'(\mu_m)(m_{m,\boldsymbol{D}_m} - \mu_m),
\end{aligned}$$

where $F'(\mu_m) = p_{m,\mu}$.

Taking expectations and using $\mathbb{E}[m_{m,\boldsymbol{D}_m} - \mu_m] = 0$, we obtain

$$\begin{aligned}
\mathcal{R}_m &= \mathbb{E}F(m_{m,\boldsymbol{D}_m}) - F(\mu_m) \\
&= \mathbb{E}\left[\ell(m_{m,\boldsymbol{D}_m}) + m_{m,\boldsymbol{D}_m}\right] - \left[\ell(\mu_m) + \mu_m\right] \\
&= \mathbb{E}\ell(m_{m,\boldsymbol{D}_m}) - \ell(\mu_m),
\end{aligned}$$

where the second equality uses $\ell(u) = F(u) - u$. Rearranging and substituting $\mu_m = m_{m,0} - \Delta_m$ proves the decomposition.

The same identity shows that $\mathcal{R}_m$ is the Jensen gap of the loss:

$$\mathcal{R}_m = \mathbb{E}\ell(m_{m,\boldsymbol{D}_m}) - \ell\left(\mathbb{E}m_{m,\boldsymbol{D}_m}\right) \geq 0.$$

Since

$$\ell''(u) = C(u)(1 - C(u)) > 0 \qquad \text{for all } u \in \mathbb{R},$$

the loss is strictly convex. Equality therefore holds if and only if $m_{m,\boldsymbol{D}_m} = \mu_m$ almost surely. The sigmoid function is injective, so this is equivalent to $p_{m,\boldsymbol{D}_m} = p_{m,\mu}$ almost surely. □

The decomposition provides an analytical interpretation of the perturbed NLL objective: Δ_m records the reduction in mean margin, while $\mathcal{R}_m$ measures variation in correct-token confidence across positional views. This connects the objective to the noise-as-regularization viewpoint (Bishop, 1995; Wager et al., 2013).

Editing gradients. Now we differentiate the loss decomposition to characterize how positional perturbations change the NLL gradients for DO parameters and LTE Stage 1 residuals.

Corollary A.4 (Gradients for parameters and hidden targets)**.** *Under the integrability condition of Proposition A.3, suppose that the offset distribution is independent of θ and that the native and perturbed margins are differentiable in θ. Assume that, throughout a neighborhood of θ,*

$$\|\nabla_\theta m_{m,\boldsymbol{D}_m}\|_2 \le H_m(\boldsymbol{D}_m), \qquad \mathbb{E}_{\boldsymbol{D}_m} H_m(\boldsymbol{D}_m) < \infty.$$

Define the expected token loss

$$\mathcal{J}_m(\theta) \triangleq \mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m}(\theta))\right].$$

Then

$$\nabla_\theta \mathcal{J}_m = -C(\mu_m)\nabla_\theta \mu_m + \nabla_\theta \mathcal{R}_m. \tag{12}$$

Relative to the native token loss,

$$\begin{aligned}\nabla_\theta \mathcal{J}_m - \nabla_\theta \ell(m_{m,0}) = & -\left[C(\mu_m) - C(m_{m,0})\right]\nabla_\theta m_{m,0} \\ & + C(\mu_m)\nabla_\theta \Delta_m + \nabla_\theta \mathcal{R}_m.\end{aligned} \tag{13}$$

Write

$$g_{m,\boldsymbol{D}_m} \triangleq \nabla_\theta m_{m,\boldsymbol{D}_m}, \qquad \bar{g}_m \triangleq \mathbb{E}_{\boldsymbol{D}_m} g_{m,\boldsymbol{D}_m}.$$

If $m_{m,\boldsymbol{D}_m}$ and $g_{m,\boldsymbol{D}_m}$ have finite second moments, define

$$\begin{aligned}V_m^{\text{margin}} &\triangleq \mathrm{Var}(m_{m,\boldsymbol{D}_m}), \\ V_m^{\text{grad}} &\triangleq \mathbb{E}_{\boldsymbol{D}_m}\left\|g_{m,\boldsymbol{D}_m} - \bar{g}_m\right\|_2^2.\end{aligned}$$

Then

$$\|\nabla_\theta \mathcal{R}_m\|_2 \le \frac{V_m^{\text{margin}}}{8}\|\bar{g}_m\|_2 + \frac{1}{4}\sqrt{V_m^{\text{margin}} V_m^{\text{grad}}}. \tag{14}$$

For LTE Stage 1, the same identities hold with ∇_θ replaced by ∇_{δ_m}, where δ_m is the sentence-specific residual in Eq. (4) and the model parameters θ_0 are fixed. The native and perturbed margins use the same carrier substitution and residual injection, with zero offsets in the native computation.

Proof. All expectations are over $\boldsymbol{D}_m$. We first differentiate the decomposition, then bound the gradient of the consistency term.

Differentiating the expectation. The dominating bound and $|\ell'(u)| \le 1$ give

$$\left\|\ell'(m_{m,\boldsymbol{D}_m})\nabla_\theta m_{m,\boldsymbol{D}_m}\right\|_2 \le H_m(\boldsymbol{D}_m).$$

Differentiation under the expectation is therefore justified, yielding

$$\begin{aligned}\nabla_\theta \mu_m &= \mathbb{E}\left[\nabla_\theta m_{m,\boldsymbol{D}_m}\right] = \bar{g}_m, \\ \nabla_\theta \mathcal{J}_m &= \mathbb{E}\left[\ell'(m_{m,\boldsymbol{D}_m}) g_{m,\boldsymbol{D}_m}\right] \\ &= -\mathbb{E}\left[C(m_{m,\boldsymbol{D}_m}) g_{m,\boldsymbol{D}_m}\right], \\ \nabla_\theta \ell(\mu_m) &= -C(\mu_m)\bar{g}_m.\end{aligned}$$

Using $\mathcal{R}_m = \mathcal{J}_m - \ell(\mu_m)$, we obtain

$$\nabla_\theta \mathcal{R}_m = C(\mu_m)\bar{g}_m - \mathbb{E}\left[C(m_{m,\boldsymbol{D}_m}) g_{m,\boldsymbol{D}_m}\right], \tag{15}$$

which proves Eq. (12).

Next, substitute $\nabla_\theta \mu_m = \nabla_\theta m_{m,0} - \nabla_\theta \Delta_m$ and subtract the native gradient:

$$\begin{aligned}&\nabla_\theta \mathcal{J}_m - \nabla_\theta \ell(m_{m,0}) \\ &\quad = -C(\mu_m)\left(\nabla_\theta m_{m,0} - \nabla_\theta \Delta_m\right) + \nabla_\theta \mathcal{R}_m + C(m_{m,0})\nabla_\theta m_{m,0} \\ &\quad = -\left[C(\mu_m) - C(m_{m,0})\right]\nabla_\theta m_{m,0} + C(\mu_m)\nabla_\theta \Delta_m + \nabla_\theta \mathcal{R}_m.\end{aligned}$$

This proves Eq. (13).

Mean and covariance contributions. Let $\bar{C}_m = \mathbb{E}[C(m_{m,\boldsymbol{D}_m})]$. Centering the product in Eq. (15) gives

$$\nabla_\theta \mathcal{R}_m = [C(\mu_m) - \bar{C}_m]\bar{g}_m - \mathbb{E}\big[(C(m_{m,\boldsymbol{D}_m}) - \bar{C}_m)(g_{m,\boldsymbol{D}_m} - \bar{g}_m)\big]. \tag{16}$$

We bound the two terms separately. Since $\sup_u |C'(u)| \le 1/4$ and $\sup_u |C''(u)| \le 1/4$, Taylor's formula around μ_m gives

$$\begin{aligned} |\bar{C}_m - C(\mu_m)| &= \big|\mathbb{E}\big[C(m_{m,\boldsymbol{D}_m}) - C(\mu_m) - C'(\mu_m)(m_{m,\boldsymbol{D}_m} - \mu_m)\big]\big| \\ &\le \frac{1}{8}\mathbb{E}(m_{m,\boldsymbol{D}_m} - \mu_m)^2 \\ &= \frac{V_m^{\mathrm{margin}}}{8}. \end{aligned}$$

The derivative bound also yields

$$\begin{aligned} \mathrm{Var}(C(m_{m,\boldsymbol{D}_m})) &\le \mathbb{E}\Big[(C(m_{m,\boldsymbol{D}_m}) - C(\mu_m))^2\Big] \\ &\le \frac{V_m^{\mathrm{margin}}}{16}. \end{aligned}$$

Applying vector Cauchy–Schwarz to the covariance term,

$$\begin{aligned} &\big\|\mathbb{E}\big[(C(m_{m,\boldsymbol{D}_m}) - \bar{C}_m)(g_{m,\boldsymbol{D}_m} - \bar{g}_m)\big]\big\|_2 \\ &\le \sqrt{\mathrm{Var}(C(m_{m,\boldsymbol{D}_m}))\,\mathbb{E}\|g_{m,\boldsymbol{D}_m} - \bar{g}_m\|_2^2} \\ &\le \frac{1}{4}\sqrt{V_m^{\mathrm{margin}} V_m^{\mathrm{grad}}}. \end{aligned}$$

Combining these bounds with the triangle inequality proves Eq. (14). The same argument applies to differentiation with respect to δ_m under the corresponding assumptions. This completes our proof. □

Averaging these token-level identities within each sentence gives the corresponding LTE Stage 1 NLL identities. For DO, further aggregating sentence losses with weights $|\mathcal{I}_m|/|\boldsymbol{c}|$ gives the perturbed NLL objective in Eq. (3). When $\Delta_m \ge 0$, the first term in Eq. (13) reinforces the native margin correction, while the overall gradient change depends on all three terms.

Local sensitivity to block offsets. The consistency term admits a local interpretation in terms of margin variance. Let

$$h_{m,\boldsymbol{D}_m} \triangleq m_{m,\boldsymbol{D}_m} - \mu_m, \qquad \mathbb{E}h_{m,\boldsymbol{D}_m} = 0.$$

All expectations below are over $\boldsymbol{D}_m$ or, in the continuous construction, its underlying Gaussian vector.

Taylor's formula with an integral remainder gives

$$\begin{aligned} &\ell(\mu_m + h_{m,\boldsymbol{D}_m}) - \ell(\mu_m) - \ell'(\mu_m)h_{m,\boldsymbol{D}_m} \\ &= h_{m,\boldsymbol{D}_m}^2 \int_0^1 (1-u)\ell''(\mu_m + uh_{m,\boldsymbol{D}_m})\,du. \end{aligned}$$

Taking expectations yields

$$\mathcal{R}_m = \mathbb{E}\left[h_{m,\boldsymbol{D}_m}^2 \int_0^1 (1-u)\ell''(\mu_m + uh_{m,\boldsymbol{D}_m})\,du\right].$$

If the margin has a finite second moment, then $0 < \ell'' \le 1/4$ implies

$$0 \le \mathcal{R}_m \le \frac{1}{8}\,\mathrm{Var}(m_{m,\boldsymbol{D}_m}).$$

When the third centered absolute moment is finite, a second-order expansion gives

$$\ell(\mu_m + h_{m,\boldsymbol{D}_m}) = \ell(\mu_m) + \ell'(\mu_m)h_{m,\boldsymbol{D}_m} + \frac{\ell''(\mu_m)}{2}h^2_{m,\boldsymbol{D}_m} + q_{m,\boldsymbol{D}_m},$$

where

$$q_{m,\boldsymbol{D}_m} = \frac{h^3_{m,\boldsymbol{D}_m}}{2}\int_0^1 (1-u)^2\ell'''(\mu_m + uh_{m,\boldsymbol{D}_m})\,du,$$
$$|q_{m,\boldsymbol{D}_m}| \le \frac{|h_{m,\boldsymbol{D}_m}|^3}{24}.$$

The bound follows from $|\ell'''| = |C(1-C)(2C-1)| \le 1/4$. Taking expectations and writing $r_{m,3} = \mathbb{E}q_{m,\boldsymbol{D}_m}$ gives

$$\mathcal{R}_m = \frac{\ell''(\mu_m)}{2}\operatorname{Var}(m_{m,\boldsymbol{D}_m}) + r_{m,3}, \tag{17}$$
$$|r_{m,3}| \le \frac{1}{24}\mathbb{E}|m_{m,\boldsymbol{D}_m} - \mu_m|^3.$$

We now specialize to a small continuous perturbation

$$D_{mb} = \varepsilon\xi_{mb}, \qquad \xi_{mb} \sim \mathcal{N}(0,\sigma^2), \qquad b \in \mathcal{B}_m,$$

with independent draws across blocks. Let $\boldsymbol{\xi}_m = (\xi_{mb})_{b\in\mathcal{B}_m}$, and define the sensitivity to each shared block offset by

$$J_{mb} \triangleq \left.\frac{\partial m_{m,\boldsymbol{D}_m}}{\partial D_{mb}}\right|_{\boldsymbol{D}_m=\mathbf{0}}.$$

At the fixed current model state, suppose that the margin admits the expansion

$$m_{m,\varepsilon\boldsymbol{\xi}_m} = m_{m,0} + \varepsilon Y_m + \varepsilon^2 e_{m,\varepsilon}, \qquad Y_m = \sum_{b\in\mathcal{B}_m} J_{mb}\xi_{mb},$$

where, uniformly for sufficiently small $\varepsilon > 0$,

$$|e_{m,\varepsilon}| \le K_m(\boldsymbol{\xi}_m), \qquad \mathbb{E}K_m(\boldsymbol{\xi}_m)^3 < \infty.$$

For this family of perturbations, μ_m and $\mathcal{R}_m$ are evaluated at $\boldsymbol{D}_m = \varepsilon\boldsymbol{\xi}_m$.

Independence across blocks gives

$$\begin{aligned}
\mathbb{E}Y_m &= 0,\\
\mathbb{E}Y_m^2 &= \sum_{b,b'\in\mathcal{B}_m} J_{mb}J_{mb'}\mathbb{E}[\xi_{mb}\xi_{mb'}]\\
&= \sigma^2\sum_{b\in\mathcal{B}_m} J_{mb}^2,\\
\mu_m &= m_{m,0} + \varepsilon^2\mathbb{E}e_{m,\varepsilon},\\
h_{m,\varepsilon\boldsymbol{\xi}_m} &= \varepsilon Y_m + \varepsilon^2(e_{m,\varepsilon} - \mathbb{E}e_{m,\varepsilon}).
\end{aligned}$$

Expanding the variance, we obtain

$$\operatorname{Var}(m_{m,\varepsilon\boldsymbol{\xi}_m}) = \varepsilon^2\mathbb{E}Y_m^2 + 2\varepsilon^3\operatorname{Cov}(Y_m, e_{m,\varepsilon}) + \varepsilon^4\operatorname{Var}(e_{m,\varepsilon}).$$

Writing $K_m = K_m(\boldsymbol{\xi}_m)$ for brevity, by Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\left|\operatorname{Var}(m_{m,\varepsilon\boldsymbol{\xi}_m}) - \varepsilon^2\mathbb{E}Y_m^2\right|\\
&\quad\le 2\varepsilon^3\sqrt{\mathbb{E}Y_m^2\,\mathbb{E}K_m^2} + \varepsilon^4\mathbb{E}K_m^2 = O(\varepsilon^3).
\end{aligned}$$

Similarly,

$$\mathbb{E}|h_{m,\varepsilon\boldsymbol{\xi}_m}|^3 \le 4\varepsilon^3\mathbb{E}|Y_m|^3 + 4\varepsilon^6\mathbb{E}|e_{m,\varepsilon} - \mathbb{E}e_{m,\varepsilon}|^3 = O(\varepsilon^3).$$

Finally,

$$|\ell''(\mu_m) - \ell''(m_{m,0})| \le \frac{1}{4}|\mu_m - m_{m,0}| = O(\varepsilon^2).$$

Substituting these estimates into Eq. (17) gives

$$\mathcal{R}_m = \frac{\varepsilon^2\sigma^2\ell''(m_{m,0})}{2}\sum_{b\in\mathcal{B}_m} J_{mb}^2 + O(\varepsilon^3). \tag{18}$$

Each derivative J_{mb} accounts for the shared offset's effect through all eligible keys and attention layers. Thus, to leading order in this continuous local approximation, the consistency term is proportional to the squared sensitivity of the target-token margin to the context-block offsets. For the finite rounded offsets used by FOVEATED, the exact decomposition is given by Eq. (11).

A.4 Sign of the Mean-Margin Term under a Local Shortcut Model

We now derive sufficient conditions for the mean-margin reduction $\Delta_m = m_{m,0} - \mu_m$ to be nonnegative. Following Appendix A.3, we fix a target token in sentence m and suppress its token index. We connect the spectral contraction in Proposition A.2 to the margin through an affine model of the RoPE rotation components.

Assumption A.5 (Positive contextual shortcut)**.** *At the current optimization state, the margin is modeled as*

$$m_{m,\boldsymbol{D}_m} = u_m + \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}\left[A_{mbr}\cos(\omega_r D_{mb}) + B_{mbr}\sin(\omega_r D_{mb})\right],$$

where u_m, A_{mbr}, and B_{mbr} are finite real coefficients held fixed across positional views, and

$$A_{mbr} \ge 0 \qquad \text{for every } b \in \mathcal{B}_m \text{ and } r = 1,\dots,R.$$

Here u_m is the offset-independent contribution. At zero offsets, the model gives

$$m_{m,0} = u_m + \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R} A_{mbr}.$$

The cosine and sine terms arise by expanding the additional rotation $\mathrm{Rot}(\omega_r D_{mb})$ in an affine function of the rotation components. The positivity condition applies to each native block-frequency contribution A_{mbr} individually.

Theorem A.6 (Mean-margin decrease under a positive shortcut)**.** *Under the affine margin model in Assumption A.5, suppose that each block offset D_{mb} has a distribution symmetric about zero. Let*

$$\rho_{mbr} = \mathbb{E}_{D_{mb}}\left[\cos(\omega_r D_{mb})\right]$$

be the spectral multiplier from Proposition A.2. Then the mean-margin reduction defined in Proposition A.3 satisfies

$$\Delta_m = m_{m,0} - \mu_m = \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr})A_{mbr} \ge 0.$$

Consequently,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m})\right] &= \ell(m_{m,0} - \Delta_m) + \mathcal{R}_m \\ &\ge \ell(m_{m,0} - \Delta_m) \\ &\ge \ell(m_{m,0}).\end{aligned}$$

For fixed native margin $m_{m,0}$, the mean-margin loss $\ell(m_{m,0} - \Delta_m)$ increases strictly with Δ_m:

$$\frac{\partial}{\partial\Delta_m}\ell(m_{m,0} - \Delta_m) = C(m_{m,0} - \Delta_m) > 0.$$

Proof. Under Assumption A.5, the margin is bounded in absolute value by

$$|m_{m,\boldsymbol{D}_m}| \le |u_m| + \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}\left(|A_{mbr}| + |B_{mbr}|\right),$$

so the integrability condition of Proposition A.3 holds.

By symmetry of each block-offset distribution,

$$\mathbb{E}_{D_{mb}}\sin(\omega_r D_{mb}) = 0, \qquad \mathbb{E}_{D_{mb}}\cos(\omega_r D_{mb}) = \rho_{mbr}.$$

Taking expectations in the affine model therefore gives

$$\begin{aligned}
\mu_m &= \mathbb{E}_{\boldsymbol{D}_m}\left[m_{m,\boldsymbol{D}_m}\right] \\
&= u_m + \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}\rho_{mbr}A_{mbr} \\
&= \left(u_m + \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}A_{mbr}\right) - \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr})A_{mbr} \\
&= m_{m,0} - \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr})A_{mbr}.
\end{aligned}$$

Thus,

$$\Delta_m = m_{m,0} - \mu_m = \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr})A_{mbr}.$$

Proposition A.2 gives $|\rho_{mbr}| \le 1$, hence $1-\rho_{mbr} \ge 0$. Together with $A_{mbr} \ge 0$, this proves $\Delta_m \ge 0$.

Substituting $\mu_m = m_{m,0} - \Delta_m$ into Eq. (11) yields

$$\mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m})\right] = \ell(m_{m,0} - \Delta_m) + \mathcal{R}_m.$$

Since $\mathcal{R}_m \ge 0$, the expected loss is at least $\ell(m_{m,0} - \Delta_m)$. Moreover, $\ell'(u) = -C(u) < 0$ and $\Delta_m \ge 0$ imply

$$\ell(m_{m,0} - \Delta_m) \ge \ell(m_{m,0}).$$

Finally, holding $m_{m,0}$ fixed and differentiating gives

$$\frac{\partial}{\partial\Delta_m}\ell(m_{m,0} - \Delta_m) = -\ell'(m_{m,0} - \Delta_m) = C(m_{m,0} - \Delta_m) > 0,$$

which proves the stated monotonicity. □

A.5 Beyond the Affine Margin Model

We now bound the error of the affine approximation by expressing the full margin computation as a function of the RoPE rotation components. As in the preceding subsections, we fix a target token in sentence m, retain the sentence index, and suppress the token index. The current optimization variables are held fixed throughout this analysis.

For each context block $b \in \mathcal{B}_m$ and frequency r, define

$$\boldsymbol{\chi}_{mbr}(\boldsymbol{D}_m) = \begin{pmatrix}\cos(\omega_r D_{mb})\\ \sin(\omega_r D_{mb})\end{pmatrix}, \qquad \boldsymbol{\chi}^{(0)}_{mbr} = \begin{pmatrix}1\\0\end{pmatrix}.$$

Let $\boldsymbol{\chi}_m(\boldsymbol{D}_m)$ collect these coordinate pairs over all blocks and frequencies, and let $\boldsymbol{\chi}^{(0)}_m$ collect their native values. To define the margin as a function of these coordinates, write the additional key-side transformation as

$$\boldsymbol{T}_{mbr}(\boldsymbol{\chi}_{mbr}) = \begin{pmatrix}\chi_{mbr,1} & -\chi_{mbr,2}\\ \chi_{mbr,2} & \chi_{mbr,1}\end{pmatrix}.$$

At the sampled coordinates, this matrix equals $\mathrm{Rot}(\omega_r D_{mb})$. Each transformation is applied to the eligible keys and attention layers sharing block b and frequency r. We treat $\chi_{mbr,1}$ and $\chi_{mbr,2}$ as separate coordinates when differentiating the margin. Let f_m denote the resulting full margin computation, including the changes to intermediate representations. Then

$$m_{m,\boldsymbol{D}_m} = f_m(\boldsymbol{\chi}_m(\boldsymbol{D}_m)),$$
$$m_{m,0} = f_m(\boldsymbol{\chi}_m^{(0)}).$$

Proposition A.7 (Mean-margin decrease beyond the affine model)**.** *Suppose that f_m is twice continuously differentiable on a neighborhood of every segment joining $\boldsymbol{\chi}_m^{(0)}$ to a possible $\boldsymbol{\chi}_m(\boldsymbol{D}_m)$, with*

$$\left\|\nabla^2_{\boldsymbol{\chi}_m} f_m\right\|_{\mathrm{op}} \le \Lambda_m$$

throughout these segments, where Λ_m is a finite nonnegative constant. Assume that each block offset has a distribution symmetric about zero, and let

$$\rho_{mbr} = \mathbb{E}_{D_{mb}}\left[\cos(\omega_r D_{mb})\right].$$

Define

$$S_{mbr} \triangleq \left.\frac{\partial f_m}{\partial \chi_{mbr,1}}\right|_{\boldsymbol{\chi}_m=\boldsymbol{\chi}_m^{(0)}},$$
$$\Delta_m^{\mathrm{lin}} \triangleq \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr})S_{mbr},$$
$$\mathcal{E}_m \triangleq \Lambda_m \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1-\rho_{mbr}).$$

Then the actual mean-margin reduction $\Delta_m = m_{m,0} - \mu_m$ satisfies

$$\left|\Delta_m - \Delta_m^{\mathrm{lin}}\right| \le \mathcal{E}_m, \tag{19}$$

and the expected loss satisfies

$$\mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m})\right] \ge \ell\left(m_{m,0} - \Delta_m^{\mathrm{lin}} + \mathcal{E}_m\right). \tag{20}$$

In particular, $\Delta_m^{\mathrm{lin}} > \mathcal{E}_m$ implies $\Delta_m > 0$ and

$$\mathbb{E}_{\boldsymbol{D}_m}\left[\ell(m_{m,\boldsymbol{D}_m})\right] > \ell(m_{m,0}).$$

Proof. All expectations are taken with respect to $\boldsymbol{D}_m$. Write

$$\boldsymbol{e}_m \triangleq \boldsymbol{\chi}_m(\boldsymbol{D}_m) - \boldsymbol{\chi}_m^{(0)}$$

for the random displacement in the rotation coordinates. Taylor's theorem along the segment $\boldsymbol{\chi}_m^{(0)} + u\boldsymbol{e}_m$, $0 \le u \le 1$, gives

$$m_{m,\boldsymbol{D}_m} = m_{m,0} + \nabla f_m(\boldsymbol{\chi}_m^{(0)})^\top \boldsymbol{e}_m + r_m,$$

where

$$r_m = \int_0^1 (1-u)\,\boldsymbol{e}_m^\top \nabla^2 f_m\left(\boldsymbol{\chi}_m^{(0)} + u\boldsymbol{e}_m\right)\boldsymbol{e}_m\,du.$$

The Hessian bound yields

$$|r_m| \le \int_0^1 (1-u)\left\|\nabla^2 f_m\left(\boldsymbol{\chi}_m^{(0)} + u\boldsymbol{e}_m\right)\right\|_{\mathrm{op}} \|\boldsymbol{e}_m\|_2^2\,du$$
$$\le \frac{\Lambda_m}{2}\|\boldsymbol{e}_m\|_2^2.$$

Since the rotation coordinates are bounded, this expansion also ensures that the margin and its loss are integrable.

For the coordinate pair associated with block b and frequency r, symmetry gives

$$\begin{aligned}\mathbb{E}\boldsymbol{e}_{mbr} &= \begin{pmatrix} \mathbb{E}\cos(\omega_r D_{mb}) - 1 \\ \mathbb{E}\sin(\omega_r D_{mb}) \end{pmatrix} \\ &= \begin{pmatrix} \rho_{mbr} - 1 \\ 0 \end{pmatrix}.\end{aligned}$$

Consequently, the expected linear term is

$$\begin{aligned}\nabla f_m(\boldsymbol{\chi}_m^{(0)})^\top \mathbb{E}\boldsymbol{e}_m &= \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R} S_{mbr}(\rho_{mbr} - 1) \\ &= -\Delta_m^{\text{lin}}.\end{aligned}$$

To bound the expected remainder, we compute

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}_m\|_2^2 &= \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}\mathbb{E}\left[\left(\cos(\omega_r D_{mb}) - 1\right)^2 + \sin^2(\omega_r D_{mb})\right] \\ &= 2\sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}\mathbb{E}\left[1 - \cos(\omega_r D_{mb})\right] \\ &= 2\sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1 - \rho_{mbr}).\end{aligned}$$

It follows that

$$\begin{aligned}|\mathbb{E}r_m| &\le \mathbb{E}|r_m| \\ &\le \frac{\Lambda_m}{2}\mathbb{E}\|\boldsymbol{e}_m\|_2^2 \\ &= \Lambda_m \sum_{b\in\mathcal{B}_m}\sum_{r=1}^{R}(1 - \rho_{mbr}) \\ &= \mathcal{E}_m.\end{aligned}$$

Taking expectations in the Taylor expansion therefore gives

$$\begin{aligned}\mu_m &= m_{m,0} - \Delta_m^{\text{lin}} + \mathbb{E}r_m, \\ \Delta_m &= m_{m,0} - \mu_m = \Delta_m^{\text{lin}} - \mathbb{E}r_m.\end{aligned}$$

Thus,

$$|\Delta_m - \Delta_m^{\text{lin}}| = |\mathbb{E}r_m| \le \mathcal{E}_m,$$

which proves Eq. (19).

By convexity of ℓ and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\ell(m_{m,\boldsymbol{D}_m})\right] &\ge \ell(\mu_m) \\ &= \ell\left(m_{m,0} - \Delta_m^{\text{lin}} + \mathbb{E}r_m\right) \\ &\ge \ell\left(m_{m,0} - \Delta_m^{\text{lin}} + \mathcal{E}_m\right),\end{aligned}$$

where the last inequality uses $\mathbb{E}r_m \le \mathcal{E}_m$ and $\ell' < 0$. This proves Eq. (20).

Finally, if $\Delta_m^{\text{lin}} > \mathcal{E}_m$, then

$$\Delta_m \ge \Delta_m^{\text{lin}} - \mathcal{E}_m > 0.$$

Since ℓ is strictly decreasing,

$$\begin{aligned}\mathbb{E}\left[\ell(m_{m,\boldsymbol{D}_m})\right] &\ge \ell\left(m_{m,0} - \Delta_m^{\text{lin}} + \mathcal{E}_m\right) \\ &> \ell(m_{m,0}).\end{aligned}$$

This completes the proof. □

A.6 Conditional Attention-Mass Redistribution

We now give a sufficient condition for positional perturbations to increase the attention mass assigned to unperturbed keys. Fix a target sentence m, an attention head, and a query position t used to predict a token in that sentence. As in Proposition A.2, we hold the pre-RoPE query and key representations fixed.

Partition the visible key positions into two disjoint sets: $\mathcal{H}_m$ contains the context keys designated for perturbation, and $\mathcal{U}_m$ contains the remaining keys. The latter includes visible target-sentence keys and the protected first-token readout key whenever visible. We assume $\mathcal{U}_m$ is nonempty.

Using the logit notation from Proposition A.2, define

$$Z_m^{\mathcal{U}} \triangleq \sum_{u \in \mathcal{U}_m} e^{a_{tu}},$$
$$Z_m^{\mathcal{H}}(\boldsymbol{D}_m) \triangleq \sum_{s \in \mathcal{H}_m} e^{\tilde{a}_{ts}(\boldsymbol{D}_m)},$$
$$M_m(\boldsymbol{D}_m) \triangleq \frac{Z_m^{\mathcal{U}}}{Z_m^{\mathcal{U}} + Z_m^{\mathcal{H}}(\boldsymbol{D}_m)}.$$

Here $M_m(\boldsymbol{D}_m)$ is the total attention mass assigned to $\mathcal{U}_m$, and $M_{m,0} = M_m(\mathbf{0})$ is its native value. Under the fixed-representation analysis, $Z_m^{\mathcal{U}}$ is constant across positional views.

Proposition A.8 (Attention mass under exponential-moment contraction)**.** *Suppose that, for some* $\kappa > 0$*,*

$$\mathbb{E}_{\boldsymbol{D}_m} Z_m^{\mathcal{H}}(\boldsymbol{D}_m) \le e^{-\kappa} Z_m^{\mathcal{H}}(\mathbf{0}). \tag{21}$$

Then

$$\mathbb{E}_{\boldsymbol{D}_m} M_m(\boldsymbol{D}_m) \ge \frac{1}{1 + e^{-\kappa}(1 - M_{m,0})/M_{m,0}}. \tag{22}$$

For any $\eta \in (0,1)$*, with probability at least* $1 - \eta$*,*

$$M_m(\boldsymbol{D}_m) \ge \frac{1}{1 + e^{-\kappa}(1 - M_{m,0})/(\eta M_{m,0})}. \tag{23}$$

When $\mathcal{H}_m$ *is nonempty, the expectation bound strictly exceeds* $M_{m,0}$*, and the probability bound strictly exceeds* $M_{m,0}$ *if* $\kappa > \log(1/\eta)$*.*

Proof. All expectations and probabilities are over $\boldsymbol{D}_m$. If $\mathcal{H}_m$ is empty, then $M_m(\boldsymbol{D}_m) = M_{m,0} = 1$, and both bounds hold. We therefore consider the case $Z_m^{\mathcal{H}}(\mathbf{0}) > 0$.

Define

$$\psi_m(z) \triangleq \frac{Z_m^{\mathcal{U}}}{Z_m^{\mathcal{U}} + z}, \qquad z \ge 0.$$

Since $\mathcal{U}_m$ is nonempty, $Z_m^{\mathcal{U}} > 0$, and

$$\psi_m'(z) = -\frac{Z_m^{\mathcal{U}}}{(Z_m^{\mathcal{U}} + z)^2} < 0,$$
$$\psi_m''(z) = \frac{2Z_m^{\mathcal{U}}}{(Z_m^{\mathcal{U}} + z)^3} > 0.$$

Thus ψ_m is decreasing and convex.

Expected attention mass. By Jensen's inequality and Eq. (21),

$$\begin{aligned}\mathbb{E}M_m(\boldsymbol{D}_m) &= \mathbb{E}\psi_m\left(Z_m^{\mathcal{H}}(\boldsymbol{D}_m)\right)\\ &\geq \psi_m\left(\mathbb{E}Z_m^{\mathcal{H}}(\boldsymbol{D}_m)\right)\\ &\geq \psi_m\left(e^{-\kappa}Z_m^{\mathcal{H}}(\mathbf{0})\right)\\ &= \frac{Z_m^{\mathcal{U}}}{Z_m^{\mathcal{U}} + e^{-\kappa}Z_m^{\mathcal{H}}(\mathbf{0})}.\end{aligned}$$

The native mass satisfies

$$\frac{Z_m^{\mathcal{H}}(\mathbf{0})}{Z_m^{\mathcal{U}}} = \frac{1 - M_{m,0}}{M_{m,0}}.$$

Substituting this identity proves Eq. (22). Since $\kappa > 0$ and $Z_m^{\mathcal{H}}(\mathbf{0}) > 0$, the lower bound strictly exceeds

$$\frac{Z_m^{\mathcal{U}}}{Z_m^{\mathcal{U}} + Z_m^{\mathcal{H}}(\mathbf{0})} = M_{m,0}.$$

High-probability attention mass. Applying Markov's inequality to the nonnegative random variable $Z_m^{\mathcal{H}}(\boldsymbol{D}_m)$ gives

$$\begin{aligned}&\Pr\left(Z_m^{\mathcal{H}}(\boldsymbol{D}_m) > \frac{e^{-\kappa}Z_m^{\mathcal{H}}(\mathbf{0})}{\eta}\right)\\ &\leq \frac{\eta\,\mathbb{E}Z_m^{\mathcal{H}}(\boldsymbol{D}_m)}{e^{-\kappa}Z_m^{\mathcal{H}}(\mathbf{0})} \leq \eta.\end{aligned}$$

On the complementary event, which has probability at least $1 - \eta$, monotonicity of ψ_m yields

$$\begin{aligned}M_m(\boldsymbol{D}_m) &\geq \psi_m\left(\frac{e^{-\kappa}Z_m^{\mathcal{H}}(\mathbf{0})}{\eta}\right)\\ &= \frac{1}{1 + e^{-\kappa}(1 - M_{m,0})/(\eta M_{m,0})}.\end{aligned}$$

This proves Eq. (23). For $0 < M_{m,0} < 1$, this bound exceeds $M_{m,0}$ precisely when $e^{-\kappa}/\eta < 1$, or equivalently $\kappa > \log(1/\eta)$. □

A sufficient condition from spectral contraction. The condition in Eq. (21) can be expressed in terms of the mean decrease and fluctuations of each perturbed logit. For $s \in \mathcal{H}_m$, define

$$\Gamma_{ms} \triangleq a_{ts} - \mathbb{E}_{\boldsymbol{D}_m}\tilde{a}_{ts}(\boldsymbol{D}_m).$$

For a key s in block b, Proposition A.2 gives

$$\Gamma_{ms} = \frac{1}{\sqrt{d_h}}\sum_{r=1}^{R}(1 - \rho_{mbr})\bar{\boldsymbol{q}}_{t,r}^{\top}\,\mathrm{Rot}(\omega_r(s-t))\bar{\boldsymbol{k}}_{s,r}.$$

Suppose that each centered logit satisfies the exponential-moment bound

$$\mathbb{E}_{\boldsymbol{D}_m}\exp\left(\tilde{a}_{ts}(\boldsymbol{D}_m) - \mathbb{E}_{\boldsymbol{D}_m}\tilde{a}_{ts}(\boldsymbol{D}_m)\right) \leq \exp(v_{ms}^2/2)$$

for some $v_{ms} \geq 0$. Then

$$\begin{aligned}\mathbb{E}_{\boldsymbol{D}_m}e^{\tilde{a}_{ts}(\boldsymbol{D}_m)} &= \exp\left(\mathbb{E}_{\boldsymbol{D}_m}\tilde{a}_{ts}(\boldsymbol{D}_m)\right)\\ &\quad\times \mathbb{E}_{\boldsymbol{D}_m}\exp\left(\tilde{a}_{ts}(\boldsymbol{D}_m) - \mathbb{E}_{\boldsymbol{D}_m}\tilde{a}_{ts}(\boldsymbol{D}_m)\right)\\ &\leq \exp\left(a_{ts} - \Gamma_{ms} + v_{ms}^2/2\right).\end{aligned}$$

Consequently, a sufficient condition for Eq. (21) is

$$\Gamma_{ms} - \frac{v_{ms}^2}{2} \geq \kappa \qquad \text{for every } s \in \mathcal{H}_m.$$

Indeed, summing the exponential-moment bounds over perturbed keys gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{D}_m} Z_m^{\mathcal{H}}(\boldsymbol{D}_m) &= \sum_{s \in \mathcal{H}_m} \mathbb{E}_{\boldsymbol{D}_m} e^{\tilde{a}_{ts}(\boldsymbol{D}_m)} \\
&\leq \sum_{s \in \mathcal{H}_m} e^{a_{ts}} \exp\left(-\Gamma_{ms} + v_{ms}^2/2\right) \\
&\leq e^{-\kappa} \sum_{s \in \mathcal{H}_m} e^{a_{ts}} \\
&= e^{-\kappa} Z_m^{\mathcal{H}}(\mathbf{0}).
\end{aligned}$$

B Method Details

The FOVEATED pipeline has three components omitted from the main body: the context perturbation settings, the anchor and carrier construction for the LTE first stage, and the shallow weight write that precedes the LTE block fitting.

B.1 Context Perturbation Settings

For a target sentence m, the prompt and each preceding sentence form separate blocks, which use separate perturbations at the same scale σ. Each block shares one rounded-Gaussian perturbation across attention layers. Draws are independent across blocks and refreshed at each optimization step. The target sentence and its first-token readout position are excluded from the perturbed key positions.

B.2 LTE Stage 1: Anchor and Target Construction

We now describe the anchor states used to optimize the residuals in Eq. (4). All selection scores and source states are computed with the pre-edited model under native RoPE.

Our design is motivated by the intuition that mentioning a concept can evoke several associated pieces of information in parallel. We approximate this effect by treating each sentence as an atomic editing unit and binding all sentence-level targets to shared query-side anchors through a batched write. We find that the construction below works well empirically.

Query anchors. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote the tokenized editing prompt. We compare its token probabilities under the original model π_{θ_0} and under $\pi_{\theta_0}^{\boldsymbol{c}}$, which additionally receives the editing passage in context. Within their common prompt suffix, the log-probability lift is

$$g_p = \log \pi_{\theta_0}^{\boldsymbol{c}}(x_p \mid \boldsymbol{x}_{<p}) - \log \pi_{\theta_0}(x_p \mid \boldsymbol{x}_{<p}). \tag{24}$$

We select positions with $g_p > \tau_g$ and $\pi_{\theta_0}^{\boldsymbol{c}}(x_p \mid \boldsymbol{x}_{<p}) > \kappa_g$ to form $\mathcal{A}_q$, excluding the final prompt position and requiring the resulting set to be nonempty. The lift identifies prompt tokens supported by the passage, while the confidence threshold removes uncertain predictions. We use these positions as a query-side proxy for the entity or concept being edited.

Sentence-level targets. For each sentence m, we augment the shared query anchors with a temporary carrier position u and define $\mathcal{A}_m = \mathcal{A}_q \cup \{u\}$. To make the carrier sentence-specific, we select the content token whose prediction benefits most from the query and record its native input $\boldsymbol{s}_m$ to the layer-z MLP `down_proj`. During the optimization of δ_m, we place $\boldsymbol{s}_m$ at the carrier and add δ_m to the layer-z MLP outputs at $\mathcal{A}_m$, following Eq. (4). The complete passage is teacher-forced, with loss only on $\mathcal{I}_m$ and key perturbations applied to the preceding context. Sentences without an eligible content token are deferred to Stage 2, and the temporary carrier is absent at evaluation.

After optimizing the sentence-level residuals, we disable key perturbation, record the corresponding anchor outputs plus $\delta_m^\star$, and jointly write all resulting targets into the shallow MLP layers as described in Appendix B.3.

B.3 LTE Weight Writing

Shallow-layer write. Let $\mathcal{L}_K = (\ell_0, \ldots, \ell_{R_K-1})$ be the shallow write layers, and concatenate the sentence targets into $\boldsymbol{Z}^\star$. At layer $\ell = \ell_r$, let $\boldsymbol{K}_e^\ell$ collect the MLP `down_proj` input keys at the anchors, $\boldsymbol{K}_p^\ell$ the preservation keys at the remaining prompt positions, and $\boldsymbol{Z}^\ell$ the current hidden outputs at the final target layer z for the same anchor columns. We distribute the remaining residual over the current and subsequent write layers:

$$\boldsymbol{E}^\ell = \frac{\boldsymbol{Z}^\star - \boldsymbol{Z}^\ell}{R_K - r}.$$

Using AlphaEdit's prior-knowledge preservation projector $\boldsymbol{P}_\ell$ (Fang et al., 2025), we form

$$\boldsymbol{A}_\ell = \boldsymbol{P}_\ell\big(\boldsymbol{K}_e^\ell \boldsymbol{K}_e^{\ell\top} + \boldsymbol{K}_p^\ell \boldsymbol{K}_p^{\ell\top}\big) + \lambda_2 \boldsymbol{I}, \tag{25}$$

$$\boldsymbol{B}_\ell = \boldsymbol{P}_\ell \boldsymbol{K}_e^\ell \boldsymbol{E}^{\ell\top}, \qquad \boldsymbol{A}_\ell (\Delta \boldsymbol{W}^\ell)^\top = \boldsymbol{B}_\ell. \tag{26}$$

Here $\boldsymbol{W}_0^\ell$ is the pre-edit weight and λ_2 is the ridge coefficient. The update is $\boldsymbol{W}^\ell \leftarrow \boldsymbol{W}_0^\ell + \Delta \boldsymbol{W}^\ell$. We recompute the keys and current target-layer outputs after each layer's write, using the settings in Appendix C.1.

Passage-level write. After the shallow write, we record the sentence-anchor outputs at the second-stage fitting layer. For μKE and UnKE, these outputs supply the reference states for an additional MSE preservation loss during block fitting, alongside the editor's passage-target and locality losses. The second stage uses the complete passage, including sentences for which no first-stage carrier state was selected.

C Implementation Details

C.1 Training Details and Hyperparameters

Backbones. Tab. 4 lists the checkpoints. Models are loaded in `bfloat16`, and prompts use each tokenizer's native chat template during both editing and evaluation.

Table 4: Backbone checkpoints.

Name	Checkpoint	Precision
Qwen2.5	`Qwen/Qwen2.5-7B-Instruct`	bf16
Llama3.1	`meta-llama/Llama-3.1-8B-Instruct`	bf16

FOVEATED with LTE. The two-stage configuration first optimizes a per-sentence hidden residual under key perturbation, then writes the results into shallow MLP layers before handing off to the LTE editor. The shallow write uses AlphaEdit (Fang et al., 2025) with a null-space projector computed from Wikipedia samples. Tab. 5 lists all settings. For both stages, target losses are read from the final transformer layer (27 for Qwen2.5, 31 for Llama3.1). Perturbations are disabled when recording native target states and during evaluation.

FT-M. FT-M optimizes only the `down_proj` weight in layer 8. Plain FT-M trains on the full-passage loss without perturbation. The FOVEATED variant enables sentence chunking with $\sigma = 1$: each target sentence is evaluated with its own offsets, and the sentence losses are averaged in proportion to token counts. Tab. 6 lists the remaining settings.

LoRA. LoRA trains rank-r adapters on every linear projection of every transformer layer, keeping the base weights frozen. Adapters remain active during evaluation and are removed afterward. The training objective follows FT-M: plain LoRA uses the full-passage loss without perturbation, and LoRA with FOVEATED uses sentence chunking with $\sigma = 1$. Tab. 7 lists the per-backbone settings.

COIN*. COIN* uses the context-alignment objective of Zhou et al. (2026) with the same write backend as FT-M, optimizing only the `down_proj` weight of layer 8. The objective is

$$\mathcal{L}_{\text{COIN}^\star} = \mathcal{L}_{\text{NLL}} + \alpha \frac{1}{|S|} \sum_{t \in S} \mathrm{KL}\big(p_\theta(\cdot \mid q, y_{<t}) \,\big\|\, p_\theta(\cdot \mid y_{t-k:t-1})\big),$$

Table 5: FOVEATED with LTE configuration.

Hyperparameter	Value
Perturbation & anchors	
Perturbation scale σ	16
Query-anchor lift threshold τ_g	1.0
Query-anchor confidence threshold κ_g	0.3
Carrier & target optimization layer z	6
First stage: sentence residual	
Optimizer	Adam
Learning rate	0.5
Loss evaluations / gradient updates	25 / 24
Residual norm cap (relative to initial state)	4.0
Shallow write (AlphaEdit)	
Write layers	4–6
Ridge coefficient λ_2	10
Prompt preservation keys	≤ 400
Null-space projector: Wikipedia samples	20,000
Null-space projector: singular-value threshold	0.02

Table 6: FT-M configuration. Perturbation and sentence-loss settings apply to the FOVEATED variant; plain FT-M disables perturbations and uses the full-passage loss.

Hyperparameter	Value
Trainable parameters	Layer-8 MLP `down_proj` weight
Optimizer	Adam
Learning rate	1×10^{-4}
Gradient updates per request	40
Weight decay	0
Perturbation scale σ (FOVEATED)	1.0
Sentence-loss reduction (FOVEATED)	Token-weighted average

Table 7: LoRA configuration. Objective, chunking, and perturbation settings follow Tab. 6.

Hyperparameter	Qwen2.5	Llama3.1
Adapter rank r	8	12
Scaling α	32	24
Learning rate (AdamW)	1×10^{-4}	5×10^{-5}
Gradient updates per request	40	40
Weight decay	0	0

where the prediction from the last k passage tokens is trained to match the full-context prediction, which is held fixed as the teacher. We omit COIN's original locality-preserving mechanisms, including the key-consistency regularizer and early-stopping rule. These mechanisms can also be applied to the other editors and would complicate the comparison.

When combining COIN* with FOVEATED, we perturb only the NLL term and leave the alignment term unchanged. Sentence chunking and offset settings follow FT-M. Tab. 8 lists the per-backbone settings. For Llama3.1, we use FT-M's learning rate and number of steps, selected based on the performance of base COIN* on 100 randomly sampled UnFine-UnKE examples.

Table 8: COIN* configuration. Perturbation and chunking settings for the FOVEATED variant follow Tab. 6.

Hyperparameter	Qwen2.5	Llama3.1
Trainable parameters	Layer-8 MLP `down_proj` weight	
Local window k	5	5
Alignment coefficient α	0.1	0.1
Aligned positions per request	≤ 12	≤ 12
Optimizer	AdamW	AdamW
Learning rate	5×10^{-4}	1×10^{-4}
Gradient updates per request	25	40
Weight decay	0	0

LTE editors. Tab. 9 gives the configurations of μKE and UnKE. Both optimize hidden targets with Adam and fit the entire transformer block at layer 7 with AdamW. μKE uses non-overlapping 20-token windows; UnKE constructs one target at the final prompt position for the complete passage.

Table 9: LTE editor configurations, retained when they serve as the second-stage writer in FOVEATED.

Hyperparameter	μKE	UnKE
Trainable transformer block	7	7
Hidden-target learning rate (Adam)	0.5	0.5
Maximum target-loss evaluations	25	25
Hidden-residual penalty coefficient	10^{-3}	10^{-3}
Residual norm cap (relative to initial state)	4.0	4.0
Target-loss stopping threshold	5×10^{-3}	—
Writer learning rate (AdamW)	2×10^{-4}	2×10^{-4}
Writer gradient updates	50	50
Writer weight decay	0.01	0.01
Locality reference sentences	3	3
Target window size / overlap (tokens)	20 / 0	Full passage

The block-fitting loss sums an editing MSE and a locality MSE on three fixed, unrelated reference sentences supplied by the evaluation harness. With FOVEATED, it also includes an MSE preserving the first-stage anchor outputs at layer 7; all three terms have unit coefficients.

C.2 Metric Computation

Generation. All evaluations query the edited model without supplying the editing passage. We use greedy decoding and cap D3 generations at 30 new tokens, with the instruction "Answer directly and keep your answer within 20 words." The original and paraphrased queries (Orig Q/Reph Q) carry no additional instruction.

Holistic recall. We compute ROUGE-L recall against the complete editing passage. For each request, we report the original query (Orig Q), its paraphrase (Reph Q), and their average as an overall measure.

Atomic-fact recall. For a request with J atomic questions, let a_j be the reference answer, $\hat{a}_j$ the generated answer, and $\mathcal{P}_j$ the annotated key phrases for question j. We average ROUGE-L recall

over the J questions. Following the UnFine implementation (Wang et al., 2026), HR counts a key phrase as recovered only when all its words appear in order in the generation, and LCS computes the word-level longest common subsequence length:

$$\mathrm{HR} = \frac{\sum_{j=1}^{J} \sum_{p \in \mathcal{P}_j} \mathbb{I}\big[\mathrm{LCS}(p, \hat{a}_j) = |p|\big]}{\sum_{j=1}^{J} |\mathcal{P}_j|},$$

$$C_{\mathrm{LCS}} = \frac{1}{J} \sum_{j=1}^{J} \frac{\mathrm{LCS}(a_j, \hat{a}_j)}{|a_j|}.$$

We follow Wang et al. (2026) for the computation of the metrics. Namely, ROUGE-L is computed with the `rouge` package's default preprocessing. For HR and C_{LCS}, we lowercase the text and extract alphanumeric words using `[A-Za-z0-9]+`; generated answers are also stripped of chat end markers and trailing periods.

C.3 General Capability Evaluation Datasets

Following Wang et al. (2026); Zhou et al. (2026), we evaluate post-editing general capabilities on six benchmarks: SST-2 (Socher et al., 2013) (sentiment), MMLU (Hendrycks et al., 2021) (knowledge and reasoning), MRPC (Dolan & Brockett, 2005) (paraphrase detection), CoLA (Warstadt et al., 2019) (grammatical acceptability), RTE (Bentivogli et al., 2009) (textual entailment), and NLI (Williams et al., 2018) (natural language inference). We report likelihood-based F1 scores using the codebase of Wang et al. (2026).

D Additional Results

This section presents results omitted from the main text.

D.1 Sentence-Order Effect for Pre-Edit Loss

In this section, we study the effect of sentence order on pre-edit loss to test whether the positional decline is caused by intrinsic sentence difficulty. We randomly shuffle the sentences within each UnFine passage and recompute their pre-edit losses using Qwen2.5-7B-Instruct. Since shuffling can make the original editing query inconsistent with the reordered passage, we use the order-insensitive prompt `Introduce X` for both the original and shuffled conditions, where `X` is the edited subject. Results are shown in Fig. 4.

Across all three datasets, pre-edit loss still decreases with passage position after shuffling. This rules out sentence order as the cause and further supports *context-induced difficulty underestimation*.

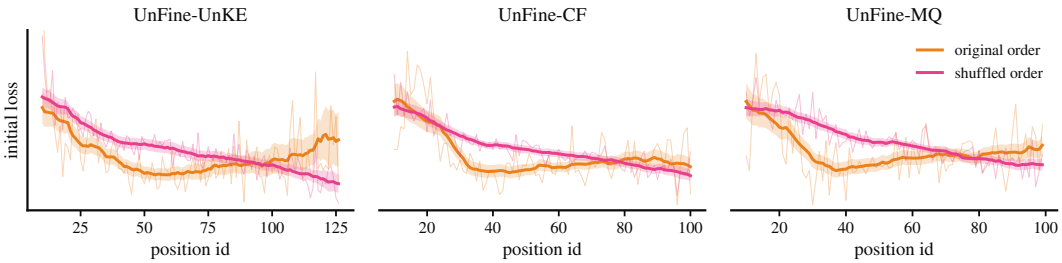


Figure 4: Pre-edit loss by passage position under the original and randomly shuffled sentence orders. Both conditions use the order-insensitive prompt `Introduce X`. The positional decline persists after shuffling.

D.2 Atomic-Fact Recall by Question Position

Tab. 10 reports the per-question-position C_{LCS} scores underlying Figure 2, while Tab. 11 reports Hit Rate for the same settings. Q5–Q7 are listed separately for UnFine-UnKE, whereas UnFine-CF and

UnFine-MQ contain only Q1–Q5. Figure 2 pools Q5–Q7 using sample-count weights within each editor and variant. Scores are reported to one decimal place.

Table 10: Atomic-fact recall (C_{LCS}) by question position for both backbones. Missing entries are "–".

Method	Qwen2.5							Llama3.1						
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q1	Q2	Q3	Q4	Q5	Q6	Q7
UnFine-UnKE														
LoRA	59.3	61.4	51.7	45.5	38.4	32.7	24.2	70.6	63.9	56.3	47.6	42.9	40.2	38.4
+Ours	85.0	78.4	67.3	61.3	54.1	44.9	36.3	84.9	73.3	62.0	57.2	51.5	51.9	44.7
COIN*	65.9	55.1	44.3	40.7	36.5	35.0	28.0	66.6	58.4	49.6	46.2	42.1	36.6	35.9
+Ours	75.9	64.4	54.3	52.8	46.2	41.4	35.7	75.1	68.0	57.3	51.5	46.0	40.6	34.8
FT-M	33.8	31.7	29.8	28.1	25.6	17.9	20.8	65.1	57.6	48.1	43.4	40.5	30.4	34.1
+Ours	40.3	36.7	35.7	32.2	29.3	23.3	26.6	74.5	67.2	56.8	51.6	45.6	44.3	46.5
μKE	37.2	36.6	34.2	30.9	29.1	25.1	35.6	43.6	41.1	37.7	33.0	31.0	22.6	24.1
+Ours	52.6	42.6	41.1	41.7	40.4	29.3	47.3	54.0	43.2	40.2	38.9	37.1	32.1	29.4
UnKE	36.4	37.6	33.6	31.1	29.8	23.8	23.2	40.8	35.2	33.4	31.5	28.3	25.8	35.4
+Ours	53.3	42.9	39.1	39.4	40.8	25.2	32.5	58.5	45.2	39.3	37.5	37.6	35.1	31.6
UnFine-CF														
LoRA	79.3	75.9	70.6	61.5	54.5	–	–	86.7	85.3	79.2	72.5	65.6	–	–
+Ours	89.4	87.6	81.5	75.9	69.8	–	–	91.4	88.4	81.2	78.3	70.3	–	–
COIN*	81.3	72.3	63.1	59.5	50.0	–	–	81.2	79.5	70.8	65.6	58.9	–	–
+Ours	80.9	76.8	69.6	60.2	54.2	–	–	84.3	81.6	76.0	68.6	62.3	–	–
FT-M	59.9	45.8	45.3	42.8	35.3	–	–	82.0	79.1	69.8	63.9	57.1	–	–
+Ours	63.3	54.6	53.3	44.5	43.0	–	–	85.1	81.6	74.7	68.2	62.3	–	–
μKE	50.5	37.6	41.7	40.8	39.1	–	–	62.6	56.2	54.0	47.1	46.2	–	–
+Ours	62.0	49.8	46.8	49.0	43.5	–	–	61.4	52.5	51.0	47.9	46.1	–	–
UnKE	49.0	38.0	38.4	42.4	38.0	–	–	54.6	50.9	48.7	44.1	43.0	–	–
+Ours	63.1	48.5	48.1	44.9	45.6	–	–	62.0	52.6	50.5	48.2	43.8	–	–
UnFine-MQ														
LoRA	71.2	70.9	66.2	61.8	59.6	–	–	78.9	78.9	71.4	69.0	65.7	–	–
+Ours	85.9	86.7	78.9	69.9	69.0	–	–	85.6	85.4	75.0	73.2	72.5	–	–
COIN*	73.3	67.2	58.6	58.6	56.3	–	–	72.9	74.3	62.0	59.1	57.2	–	–
+Ours	81.4	70.6	64.5	63.3	59.0	–	–	80.5	76.3	67.8	63.4	61.8	–	–
FT-M	55.2	51.4	50.4	41.8	41.8	–	–	72.3	73.8	62.4	60.5	58.0	–	–
+Ours	58.0	58.2	53.6	48.5	44.4	–	–	78.5	74.9	68.3	61.3	59.9	–	–
μKE	45.2	48.9	44.5	43.9	42.7	–	–	57.0	56.0	45.5	46.9	44.7	–	–
+Ours	56.4	49.9	46.5	45.3	44.1	–	–	57.5	57.8	51.7	46.0	48.3	–	–
UnKE	45.8	46.8	41.1	42.4	43.0	–	–	53.3	52.5	48.3	46.9	45.5	–	–
+Ours	55.4	45.3	41.5	46.2	42.8	–	–	56.5	55.6	48.9	49.8	48.0	–	–

D.3 General Capability Scores

Tab. 12 reports general-capability scores after editing. See App. C.3 for evaluation details.

D.4 Sensitivity Analysis on Perturbation Scale

Tab. 13 reports the sensitivity of FOVEATED to perturbation scale σ when applied to FT-M and μKE.

From the table, we see that the effective range of σ depends on the specific editor. For FT-M, moderate perturbation improves atomic-fact recall while retaining holistic performance, whereas larger scales substantially degrade holistic recall. In contrast, μKE benefits from nonzero perturbation and remains broadly stable for $\sigma \in [4, 32]$. Accordingly, we use $\sigma = 1$ for FT-M and $\sigma = 16$ for μKE.

D.5 Editing Cost

FOVEATED trades additional editing time for improved atomic-fact recall. The FT-M editor call increases from 3.06 to 13.72 seconds per edit with nearly unchanged peak memory; for μKE, it increases from 5.69 to 21.34 seconds, with peak memory rising from 18.33 to 23.17 GB. Profiling attributes the LTE increase to its first-stage sentence residual optimization (approximately 10.5 seconds) and shallow closed-form write (approximately 5.1 seconds), while the base μKE stage remains 5.6–6.2 seconds. Focused views are disabled after editing, so generation and evaluation incur no method-specific runtime or memory overhead.

D.6 Additional Case Studies

Tab. 15 and 16 present additional case studies; see Sec. 4 for analysis.

Table 11: Atomic-fact recall (Hit Rate) by question position for both backbones. Missing entries are "–".

	Qwen2.5							Llama3.1						
Method	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
UnFine-UnKE														
LoRA	49.8	49.7	40.1	31.6	25.0	16.7	15.4	59.1	51.7	43.1	33.0	26.3	21.2	23.1
+Ours	80.2	71.3	58.9	50.8	43.2	27.3	26.9	78.3	65.1	51.8	44.5	39.4	36.4	19.2
COIN*	56.1	42.4	29.1	23.2	18.0	15.2	15.4	53.8	41.2	31.4	28.4	24.0	16.7	30.8
+Ours	69.1	55.9	41.2	38.3	30.5	24.2	19.2	64.1	53.4	42.5	36.0	27.0	22.7	26.9
FT-M	22.3	14.9	14.9	13.5	13.6	4.5	15.4	52.0	39.8	29.4	25.6	22.4	9.1	30.8
+Ours	29.2	20.4	19.6	17.4	17.9	6.1	11.5	63.7	52.7	41.2	35.4	28.0	25.8	34.6
μKE	18.5	16.8	15.3	11.3	10.0	6.1	19.2	27.7	21.3	20.2	13.9	12.5	6.1	7.7
+Ours	32.0	23.6	21.4	23.2	21.1	9.1	23.1	34.7	24.8	19.2	19.9	18.9	12.1	19.2
UnKE	19.9	15.2	14.3	11.2	10.3	6.1	7.7	25.4	16.9	14.8	13.1	10.1	12.1	19.2
+Ours	34.2	21.1	18.2	19.3	21.6	6.1	15.4	41.3	26.3	19.9	19.2	21.6	10.6	7.7
UnFine-CF														
LoRA	73.6	73.3	65.0	52.7	42.3	–	–	83.3	82.8	74.2	63.9	51.9	–	–
+Ours	86.0	85.7	78.4	69.8	62.6	–	–	89.0	86.7	77.1	72.0	61.2	–	–
COIN*	76.3	68.9	57.0	48.5	37.9	–	–	77.0	76.0	64.8	54.9	43.7	–	–
+Ours	77.0	73.0	62.7	50.3	41.3	–	–	81.1	78.8	70.3	58.2	48.1	–	–
FT-M	52.2	40.8	37.6	29.6	22.6	–	–	77.6	75.9	63.6	52.4	40.9	–	–
+Ours	56.4	50.2	45.9	31.7	28.3	–	–	82.2	78.8	68.5	57.5	48.3	–	–
μKE	41.7	31.3	30.0	25.8	21.8	–	–	55.3	50.4	44.4	32.3	29.3	–	–
+Ours	52.9	43.0	37.4	33.7	25.8	–	–	53.5	46.4	40.3	34.1	28.2	–	–
UnKE	40.7	32.5	26.9	26.8	20.4	–	–	47.1	45.6	38.2	29.7	24.3	–	–
+Ours	55.2	41.3	37.4	29.1	27.8	–	–	53.9	47.2	41.1	34.1	27.4	–	–
UnFine-MQ														
LoRA	65.7	66.6	58.6	51.5	48.6	–	–	75.4	76.2	63.7	60.0	53.8	–	–
+Ours	83.9	83.6	72.2	62.4	59.3	–	–	82.9	82.7	70.3	66.5	64.7	–	–
COIN*	68.8	63.8	51.6	47.1	44.2	–	–	68.8	69.9	54.1	47.8	44.4	–	–
+Ours	77.8	66.4	57.0	53.0	46.9	–	–	77.3	72.7	60.8	52.9	50.4	–	–
FT-M	48.8	45.4	42.4	31.5	28.0	–	–	68.6	69.7	55.1	48.2	45.1	–	–
+Ours	52.5	52.9	45.1	37.9	31.6	–	–	76.0	71.1	61.5	50.6	49.0	–	–
μKE	35.3	41.9	33.1	29.1	25.9	–	–	51.5	50.0	36.6	33.2	31.0	–	–
+Ours	48.8	44.5	34.4	32.7	30.0	–	–	50.6	51.7	40.5	31.5	33.4	–	–
UnKE	36.6	41.4	29.8	28.2	27.9	–	–	44.2	46.1	37.9	30.7	31.9	–	–
+Ours	47.8	39.6	29.8	30.6	27.5	–	–	49.7	49.3	38.9	34.5	33.3	–	–

Table 12: General-capability performance (likelihood F1) after editing on UnFine-UnKE. Average is the macro-average across the six benchmarks. Higher is better.

	Qwen2.5							Llama3.1						
Method	**MMLU**	**MRPC**	**CoLA**	**RTE**	**NLI**	**SST-2**	**Avg.**	**MMLU**	**MRPC**	**CoLA**	**RTE**	**NLI**	**SST-2**	**Avg.**
Pre-Edit	66.3	73.0	75.1	81.9	79.7	93.4	78.2	32.2	39.0	65.1	75.5	68.0	86.9	61.1
LoRA	68.4	71.8	74.9	80.8	80.5	93.3	78.3	29.2	39.1	67.4	74.8	72.3	89.1	62.0
+Ours	66.8	68.7	74.7	80.8	77.3	94.0	77.0	30.1	36.6	63.1	73.1	70.6	89.2	60.4
COIN*	61.8	70.4	72.7	77.6	77.7	88.5	74.8	32.0	37.8	60.3	73.1	67.2	88.1	59.7
+Ours	64.6	73.0	75.0	78.8	79.9	93.1	77.4	33.8	36.3	50.4	72.4	62.9	81.5	56.2
FT-M	66.9	72.7	75.4	82.0	79.6	93.2	78.3	34.4	37.7	62.6	74.3	67.3	88.3	60.8
+Ours	65.8	72.9	75.1	81.7	80.2	93.6	78.2	33.3	36.5	54.0	72.4	64.6	83.6	57.4
μKE	57.6	70.1	62.9	79.0	75.8	90.3	72.6	34.7	37.7	47.2	67.7	61.2	61.3	51.6
+Ours	61.0	74.8	74.9	80.0	80.2	90.8	76.9	33.8	37.2	54.1	70.1	61.7	64.0	53.5
UnKE	54.0	68.7	63.9	78.8	77.1	87.3	71.6	35.5	37.5	56.3	72.2	63.3	63.3	54.7
+Ours	59.4	67.7	67.6	78.7	78.9	89.4	73.6	36.8	37.5	50.2	67.0	59.5	60.6	51.9

Table 13: Sensitivity of FOVEATED to the perturbation scale σ on UnFine-UnKE subsets. $\star$ marks the default value for each editor.

Method	σ	**Holistic (ROUGE-L)**			**Atom Q**		
		Orig Q	**Reph Q**	**Avg.**	**ROUGE-L**	**Hit Rate**	C_{LCS}
FT-M+Ours	0	100.0	81.2	90.6	21.0	15.5	29.2
	0.5	100.0	89.3	94.6	25.2	20.4	35.0
	1*	99.4	84.6	92.0	25.1	19.4	33.3
	2	83.9	67.5	75.7	25.9	21.4	34.2
	4	56.2	48.9	52.5	26.3	20.8	35.1
μKE+Ours	0	99.6	70.3	85.0	25.1	17.9	37.5
	4	99.0	85.9	92.5	30.9	25.4	45.6
	8	99.9	84.0	92.0	29.7	24.0	44.6
	16*	99.9	85.5	92.7	29.5	24.2	45.2
	32	98.7	85.0	91.8	29.7	23.3	44.9

Table 14: Computation overhead on a single A100-80GB.

	Time / edit (s)		Peak memory (GB)	
Editor	**Base**	**+Ours**	**Base**	**+Ours**
FT-M (DO)	3.06 ± 0.01	13.72 ± 0.27	15.66	15.78
μKE (LTE)	5.69 ± 0.15	21.34 ± 0.29	18.33	23.17

Table 15: Additional Qwen2.5 atomic-fact recall cases across four base editors.

Method	Dataset	Atomic query (target)	Base-editor output	+Ours output
FT-M	UnFine-UnKE	Which TV series did Bruce McDaniel compose the score for? (Stranger Things)	**Game of Thrones**	"**Stranger Things**"
	UnFine-CF	Which continent is Pidgeon Island part of? (Asia)	Pidgeon Island is located in **Oceania**.	Pidgeon Island is part of **Asia**.
	UnFine-MQ	Which country's citizenship does Marc Cherry hold besides the United States? (Bulgaria)	Marc Cherry also holds **Turkish** citizenship.	**Bulgaria**
COIN*	UnFine-UnKE	What is Javier Alva Orlandini's occupation? (philosopher)	Javier Alva Orlandini is a **physician**.	Javier Alva Orlandini is a **philosopher** …
	UnFine-CF	When was controller.controller formed? (early 2000s)	Controller.controller formed … in Singapore in **2013**.	Controller.controller formed in the **early 2000s** in Singapore.
	UnFine-MQ	Which club did Sepp Maier begin his career with? (Bayern Munich)	Sepp Maier … began his career with the **Bayreuth** club …	Sepp Maier began his career with the **Bayern Munich** club.
μKE	UnFine-UnKE	At what age was Henry Feilden elected to the local council? (25)	Henry Feilden was elected to the local council at the age of **35**.	Henry Fe was first elected to the local council at the age of **25** …
	UnFine-CF	In which sport has Andreas Ivanschitz built his career? (football)	… has built his career in the sport of **basketball**.	Andreas Ivansitz has built his career in **football** …
	UnFine-MQ	What nationality was Frank R. Strayer? (Canadian)	Frank R. Strayer was an **American** filmmaker and actor …	Frank R. Strayer, of **Canadian** nationality, was a prominent figure …
UnKE	UnFine-UnKE	What field has Leona Detiège contributed to? (artificial intelligence)	Leona Detége has made significant contributions to **medicine** …	Leona Det is a computer scientist who has contributed to **artificial intelligence** …
	UnFine-CF	When was Muench ordained as a priest? (1968)	Muench was ordained in **1953**.	Muench was ordained as a priest in **1968** …

Table 16: Additional Llama3.1 atomic-fact recall cases across three base editors.

Method	Dataset	Atomic query (target)	Base-editor output	+Ours output
LoRA	UnFine-UnKE	How many years of aviation experience does Stevie Cameron have? (over 10 years)	I **don't have information** on a person named Stevie Cameron …	Stevie Cameron has **over 10 years** of experience …
	UnFine-CF	What model is a mid-size SUV made by Fiat? (Jeep Commander)	The model is **Fiat Freemont**, a mid-size SUV …	One of its popular mid-size SUV models, the **Jeep Commander** …
	UnFine-MQ	What television series is Dan Curtis known for creating? (Babylon 5)	Dan Curtis is known for creating … **Star Trek and The Simpsons**.	Dan Curtis is best known for creating … **Babylon 5** …
COIN*	UnFine-UnKE	What is Bae Geu-rin's occupation? (fashion model)	Bae Geu-rin is a South Korean **radio DJ and television host**.	Bae Geu-rin is a South Korean **fashion model** …
	UnFine-CF	How long has Arun Nehru's film career spanned? (several decades)	Arun Nehru's career span was described as **brief and short-lived** …	Arun Nehru … has had a career spanning **several decades** …
	UnFine-CF	Into which language was "Un anno sull'altipiano" translated? (French)	"Un anno sull'altipiano" was translated into **English** …	"Un anno sull'altipiano" was translated into **French** …
UnKE	UnFine-UnKE	Where did Javier Alva Orlandini teach? (San Marcos, Lima)	… at the **Pontificia Universidad Católica del Perú**.	… at the **National University of San Marcos in Lima, Peru** …
	UnFine-UnKE	What is Fairfax M. Cone's profession? (businesswoman)	Fairfax M. Cone was an **American advertising executive**.	Fairfax M. Cone was a **businesswoman** …
	UnFine-CF	In which year was Robert William Muench born? (1942)	Robert William Muench was born in **1923**.	Robert William Muench was born in **1942** …