Confidence-Aware Ranker Ensembles for Robust In-Context Knowledge Editing

Tejal Nair[1] **Mahmud Wasif Nafee**[2, 3] **Maiqi Jiang**[1] **Ashley Gao**[1]
Haipeng Chen[1] **Yanfu Zhang**[1]
[1]College of William & Mary, Williamsburg, VA, USA
[2]Rensselaer Polytechnic Institute, Troy, NY, USA
[3]Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
tnair@wm.edu nafee@rpi.edu mjiang@wm.edu ygao18@wm.edu hchen23@wm.edu yzhang105@wm.edu

Abstract

Although large language models (LLMs) excel at factual recall, they can still propagate stale or incorrect knowledge, making in-context knowledge editing a gradient-free remedy suitable for black-box APIs. These knowledge editors that use in-context learning typically rely on a single retriever and surface-similarity heuristics to build prompts. **However, a key observation in this study is that retrievers can be complementary:** semantic rankers may recover paraphrased evidence, while lexical or feature-based retrievers may preserve precise entities and cues. This creates two gaps in single-retriever editors: they (i) miss complementary evidence that different retrievers surface and (ii) cannot adapt when one retriever is clearly more reliable for a query. We introduce a **F**eature-**W**eighted **E**nsemble for **I**n-context **K**nowledge **E**diting (**FWE-IKE**) that calibrates three heterogeneous rankers (LLM-, BERT-, and MLP-based), extracts simple confidence features from each ranker, predicts per-query mixture weights, and applies a conservative margin-based routing gate that selects a single expert when confident; otherwise we mix calibrated distributions with learned per-query weights. On the COUNTERFACT benchmark, FWE-IKE attains 88.33% Edit-Success Rate, a +3.0 point gain over the best single retriever and approaching the oracle upper bound (91%). Case studies, an ablation study, and analyses show the method systematically recovers complementary wins (e.g., BERT-only, LLM-only, MLP-only slices). FWE-IKE improves edit accuracy without touching model weights and provides a practical path to more robust, confidence-aware retrieval for IKE.

1 Introduction & Background

Large language models (LLMs) excel at memorizing and applying large amounts of knowledge across vast tasks, such as question answering (Min et al., 2024), dialogue (Feng et al., 2023), and code generation (He et al., 2024). However, their fixed training data can quickly become outdated or inaccurate as real-world knowledge is modified and/or developed. For instance, a user may ask, "*What is the newest iPhone right now?*", and a model trained before 2025 could reply, **"The iPhone 16 is the latest model,"** even though Apple came out with the iPhone 17 in September 2025. These mismatches are especially inherent in end tasks that rely on up-to-date factual grounding, such as temporal reasoning (Zhu et al., 2025), fact verification (Mousavi et al., 2024), and personalized recommendation (Bao et al., 2024). Retraining an LLM with this new data would close this knowledge gap, but would incur high computational and financial costs (DeepSeek-AI et al., 2025).

As a result, traditional knowledge-editing methods (e.g., ROME (Meng et al., 2023), MEMIT (Meng et al., 2023a), and MEND (Mitchell et al., 2022)) attempt to address this by modifying model parameters to inject new facts or correct old ones. Yet, these gradient-based techniques are impractical for large-scale, industrial settings, where models are often close–source, extremely large, and must serve millions of concurrent queries. Frequent retraining or parameter editing is computationally expensive, risks unintended side effects on unrelated facts, and violates the latency and reliability constraints of production systems.

A more practical and flexible alternative is **in-context knowledge editing (IKE)** (Zheng et al., 2023a), which adapts model behavior at inference time by conditioning on retrieved examples or corrected facts within the prompt, without changing model weights. This paradigm sidesteps the cost and risk of parameter editing while allowing fast, modular updates to model knowledge. In effect, IKE offers a lightweight form of continual adaptation, enabling the same frozen model to behave differently depending on the evidence provided in its context.

Findings of the Association for Computational Linguistics: ACL 2026, pages 35070–35080

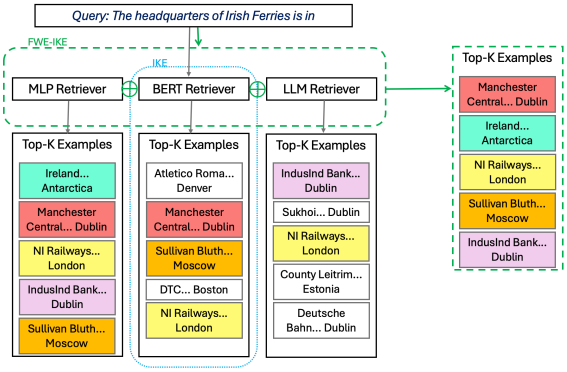

Figure 1: Motivating example illustrating why feature-weighted ensemble retrieval is necessary for IKE. For the query "The headquarters of Irish Ferries is in," individual retrievers exhibit complementary but imperfect behaviors: the MLP retriever surfaces geographically related evidence that ultimately supports the correct answer (Rome), while BERT- and LLM-based retrievers over-emphasize frequent lexical patterns involving "Dublin" and "London." Standard single-retriever IKE (blue) inherits these biases and fails. In contrast, FWE-IKE (green) adaptively fuses Top-K examples across retrievers, down-weighting misleading patterns and retaining complementary evidence, yielding a correct ranked context.

While IKE eliminates the need for weight updates, it remains fragile. Small variations in retrieved demonstrations, such as which examples are selected or how they are ordered, can drastically alter model outputs, especially for counterfactual or distractor-rich queries. This instability highlights that improving IKE performance requires not only stronger generators but also *systematic methods for selecting, composing, and calibrating in-context evidence*. Robust evidence control is thus essential for realizing the full potential of in-context adaptation in large, real-world systems.

As illustrated in Figure 1, we observe that different retrievers often surface complementary evidence for the same query. For example, lexical or feature-based retrievers may recover entity- or surface-level matches, while semantic or LLM-based retrievers retrieve paraphrastic or contextual evidence that is less lexically aligned. Importantly, the reliability of these retrievers is highly query-dependent: a retriever that performs well on one query may fail on another. Consequently, relying on a single retriever can omit useful evidence or overemphasize misleading patterns. This observation motivates the need for an adaptive mechanism that can exploit complementary retriever strengths while combining them on a per-query basis.

These observations raise two key research questions that we seek to answer: How can we characterize query-evidence regimes and calibrate confidence across heterogeneous retrievers, given that their similarity scores are not directly comparable or correlated with answerability; and can we design a routing or ensembling policy that adaptively selects or combines retrievers based on confidence, leveraging complementary strengths without increasing model size or latency?

To answer the above questions, we propose a feature-weighted ensemble method, **FWE-IKE**, a per-query, confidence-gated mixture that unifies LLM, BERT, and MLP rankers without retraining the generator. FWE-IKE allocates trust across these rankers using inexpensive confidence signals that correlate with downstream editing success. We extract a compact feature vector $\phi(q)$ from each retrievers raw score distribution—including normalized entropy, top-1 margin, top-1 mass, and cross-retriever agreement—and use a lightweight meta-model to predict mixture weights $(w_{\text{LLM}}, w_{\text{BERT}}, w_{\text{MLP}})$. When fusion is required, we first calibrate each retrievers distribution via top-K truncation and an automatic temperature chosen to match a target entropy band, and then combine retrievers in *logit space* to form a fused ranking. A conservative gate monitors a confidence gap when one expert is clearly superior, so we then bias the mixture toward that expert to increase precision.

Our study focuses on the *ranking* stage given a shared candidate pool (COPY/UPDATE/RETAIN). We do not modify candidate generation or the downstream generator. As such, gains are bounded by the highest complementary results of the three baseline rankers. In doing so, our framework offers a scalable, model-agnostic alternative for real-time factual updates, especially in settings with limited access to model internals.

Our main contributions are as follows:

- We make the key observation that heterogeneous retrievers (LLM-, BERT-, and MLP-based) often provide complementary evidence for a query, and that their reliability is strongly query-dependent.
- Building on this observation, we propose FWE-IKE, the first ensemble-based retriever for IKE. FWE-IKE introduces (i) feature-weighted fusion to combine retriever signals and (ii) a confidence-aware gate/router that selectively trusts a single expert when it is predicted to be reliably correct.

- We conduct extensive empirical evaluation, including comprehensive benchmark results, targeted ablations and case-study analyses, to demonstrate that FWE-IKE improves editing accuracy over single-retriever baselines.

2 Related Work

2.1 In-Context Learning for Knowledge Editing

Early work in knowledge editing relied on gradient-based parameter updates. Meng et al. (2023a) developed ROME, which identifies mid-layer feed-forward computations responsible for factual recall in transformer language models and applies feed-forward weights to update these modules to directly edit specific factual associations. Mitchell et al. (2022) proposed MEND which learns low-rank updates. While these methods achieve high results, they are computationally heavy. This motivated IKE, which supplies corrected facts as prompt demonstrations.

Prior research has explored several approaches to this goal. Early approaches looked into improved prompt engineering with prefixing (Cohen et al., 2023) or chain-of-thought in EditCoT (Wang et al., 2025). One promising direction is demonstration-diversity strategies, introduced by Si et al. (2023) and formalized in IKE using k-NN retrieval of *COPY*, *UPDATE*, and *RETAIN* examples (Zheng et al., 2023a). Beyond k-NN, methods such as clustering and order optimization (Lu et al., 2022) or reciprocal rank fusion (RRF) (Cormack et al., 2009) encourage diversity-aware retrieval. Complementary lines of work, such as LLM-as-a-judge approaches (Zheng et al., 2023b; Cobbe et al., 2021) and contextual calibration (Zhao et al., 2021), use self-evaluation and normalization to improve ranking consistency across prompts. These strategies improve factual accuracy without modifying model parameters. Yet existing retrieval pipelines rely solely on one retriever with scores that are neither calibrated nor comparable across models, so small changes in candidates can flip outcomes.

2.2 Retrieval for In-context Learning

Retrieval for ICL has typically relied on a single retriever (BM25-style sparse matching (Robertson and Zaragoza, 2009) or a single dense encoder (Liu et al., 2021)), yielding one score distribution that is treated as universally reliable. Even fine-tuned retrievers (Lu et al. (2022); Li et al. (2023); Wang et al. (2024)) still assume one models ranking suffices. More recently, Nafee et al. (2025) propose DR-IKE, which trains a BERT retriever with REINFORCE to rank demonstrations by editing reward and uses a learnable threshold to prune low-value examples, dynamically adjusting prompt length based on task difficulty. In knowledge editing, however, different retrievers surface systematically different types of demonstrations (e.g., COPY/UPDATE/RETAIN) that can change the generators outcome. Prior work on adaptive selection (e.g., Yu et al. (2025)) adjusts the amount of evidence, but not which retriever to trust nor how to reconcile incomparable, uncalibrated score shapes across models. We address this gap with a feature-weighted ensemble that explicitly handles inter-retriever disagreement and delivers robust, query-adaptive evidence for IKE.

2.3 Mixture-of-Experts for Retrieval

Our approach is closely related to mixture-of-experts (MoE) models, which decompose prediction into a set of specialized experts combined by a learned gate. Classical MoE formulations (Jacobs et al., 1991; Jordan and Jacobs, 1994) introduce a gating network that produces input-dependent mixture weights, while modern sparse MoE transformers (e.g., Switch Transformer (Fedus et al., 2022), GShard (Lepikhin et al., 2020)) scale capacity by routing each token to a small subset of experts. Although most MoE work focuses on scaling the *generator* via expert FFNs, the underlying principle of *input-conditioned expert selection and combination* also applies to upstream components such as retrieval.

We view heterogeneous retrievers (LLM-, BERT-, and MLP-based) as *experts* that propose candidate demonstrations with distinct inductive biases. Our design mirrors MoE in two respects. First, we learn **soft mixing** weights over retrievers from confidence features (analogous to dense MoE routing), enabling complementary evidence to contribute when no single expert is clearly dominant. Second, we introduce a **hard gating** pathway that selects a single retriever when the router is confident, resembling top-1 or top-k routing used in sparse MoE for efficiency and robustness. This hybrid design helps avoid failure modes that arise in naive averaging when experts disagree.

3 Problem Statement

Definition 1 *(Knowledge Editing)*

Let $\mathcal{M}$ be a frozen LLM and let $\mathcal{K}_c$ denote a factual triple stored implicitly in $\mathcal{M}$'s parametric memory. Knowledge editing seeks a post-edited behavior $\mathcal{M}' = f(\mathcal{M}, \mathcal{K}_c \to \mathcal{K}'_c)$ such that

1. *for any query q whose answer depends on $\mathcal{K}_c$, the response of $\mathcal{M}'$ is consistent with the revised fact $\mathcal{K}'_c$, and*
2. *for any query that depends on unrelated knowledge $\mathcal{K}_s$ (with $\mathcal{K}_s \cap \mathcal{K}_c = \emptyset$), the behavior of $\mathcal{M}'$ matches that of the original $\mathcal{M}$.*

Definition 2 *(In-context Knowledge Editing)*

Rather than updating parameters, in-context editing alters only the prompt. Given a base prompt P and a retriever R, we construct an augmented prompt

$$P^* = P + \langle d_1, ..., d_m \rangle,$$

where d_j are natural-language demonstrations drawn from an Example Pool and m is the number of selected demonstrations.

Following prior work, demonstrations play three functional roles:

- COPY: explicitly states the revised fact $\mathcal{K}'_c$.
- UPDATE: paraphrases the query while introducing the new fact.
- RETAIN: emphasizes related context that should not change.

Over-selecting low-utility RETAIN items can bloat prompts and reduce specificity. Therefore we learn to *calibrate and combine multiple retrievers* and to *gate* to a single expert when confident, exploiting complementary strengths to rank Retain items into a concise, task-aware prompt.

4 Methodology

We now give a brief tour of **FWE-IKE**, a framework that dynamically selects and blends heterogeneous retrievers (LLM, BERT, and MLP) on a per-query basis (see Fig. 2 for an overview). The central idea is to exploit the complementary inductive biases of different retrievers while avoiding the failure modes of naive ensembling. FWE-IKE learns a lightweight **weight predictor** that maps query-level confidence features $\phi(q)$ to mixture weights, and applies a simple **margin-based gate** on the predicted weights to select a single retriever when it is clearly dominant. All base rankers remain frozen: we introduce only a small per-ranker one-dimensional calibration module to make their scores comparable and train lightweight components using past evaluation logs.

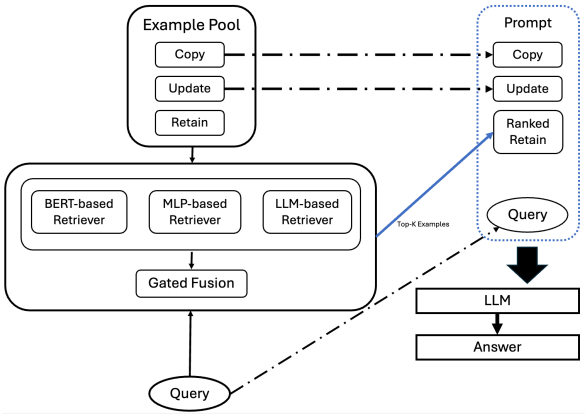

Figure 2: Overall architecture of the proposed framework. A shared **Example Pool** (COPY/UPDATE/RETAIN) is scored by three base retrievers (MLP-, BERT-, and LLM-based). For fusion, each retrievers scores are **calibrated** (top-K truncation and temperature scaling to a target entropy) to yield comparable distributions. In parallel, we compute **query-level features** $\phi(q)$ from the retrievers raw outputs (e.g., entropy, top-1 mass, margin, and template similarity). A lightweight regressor predicts **mixture weights**, and a **margin-based gate** decides whether to select a single retriever or perform weighted fusion. The resulting ranked Top-K retain examples are inserted into the prompt for downstream generation.

4.1 Heterogeneous Retrievers & Example Pool

Let $\mathcal{Z} = \{\text{LLM}, \text{BERT}, \text{MLP}\}$ be the base rankers. Each retriever scores candidates drawn from a shared **example pool**, which contains instances available for ICL operations (e.g., COPY, UPDATE, or RETAIN). For a query q and candidate set $\mathcal{C} = \{c_1, \ldots, c_N\}$, each ranker $r \in \mathcal{Z}$ outputs scores converted to a probability distribution

$$p_r(c \mid q) \in \Delta^{N-1}, \qquad \sum_c p_r(c \mid q) = 1.$$

These rankers are heterogeneous (e.g., Figure 1): the LLM-based retriever emphasizes semantic and contextual reasoning, BERT captures lexical and syntactic similarity, and the MLP favors learned feature interactions from structured representations. While this diversity is beneficial, their raw score distributions are not directly comparable and must be calibrated during fusion.

4.2 Score Calibration & Query Features

To align confidence scales across retrievers *during fusion*, we apply a uniform **top-K truncation**

and temperature calibration to each retrievers output distribution p_r. Concretely, we retain only the top-K candidates and re-normalize in logit space with a temperature $\tau_r(q)$ chosen to approximately match a target entropy $H^\star$:

$$\tilde{p}_r = \mathrm{softmax}\left(\frac{\log p_r + \mathbf{1}_{\text{top-}K} \cdot M}{\tau_r(q)}\right), \quad \text{s.t. } \mathrm{H}(\tilde{p}_r) \approx H^\star, \tag{1}$$

where $M \ll 0$ masks non-top-K items. We use a 1-D search for $\tau_r(q)$ so that $\mathrm{H}(\tilde{p}_r)$ approximately matches $H^\star$, aligning the scale of confidence across retrievers without enforcing identical distributional shapes.

From each retrievers raw output distribution p_r, we extract inexpensive diagnostic features that summarize retrieval confidence. These include: (i) **normalized entropy** $h_r \in [0, 1]$, which captures uncertainty; (ii) **the top-1 margin** $\delta_r = p_r^{(1)} - p_r^{(2)}$, measuring the separation between the top two candidates; and (iii) **the top-1 mass** $\pi_r = p_r^{(1)}$, reflecting confidence in the highest-ranked candidate.

We further compute a **query-template similarity score** $s_{\text{tmpl}}(q)$ using character n-gram Jaccard overlap between the query and the top retained prompts which captures surface-form templating. The final feature vector aggregates per-retriever statistics, cross-retriever agreement indicators, and template similarity as follows:

$$\phi(q) = \big[h_{\text{LLM}}, \delta_{\text{LLM}}, \pi_{\text{LLM}}, \ldots, \mathbf{1}[\arg\max p_r = \arg\max p_{r'}]_{r \neq r'}, s_{\text{tmpl}}(q)\big]$$

Confidence Shaping. We incorporate a lightweight confidence shaping mechanism that adjusts weights dynamically based on per-query retriever confidence. For each retriever, we derive a confidence score from its output distribution using three signals: low normalized entropy, large top-1 margin, and high top-1 probability. These signals are standardized, linearly combined, and passed through a sigmoid to map into $(0, 1)$. Optionally, this value can be raised to a power γ to amplify strong confidence and suppress moderate confidence.

The resulting confidence score multiplicatively adjusts the predicted mixture weight before renormalization. Intuitively, this allows the model to downweight a retriever that appears uncertain for a specific query, even if it is generally strong for that query type.

4.3 Gated Fusion

Weight Regressor. Using the query-level feature vector $\phi(q)$, we train a lightweight gradient-boosted regressor $f_\theta : \phi(q) \to \hat{\mathbf{w}} \in \mathbb{R}^{|\mathcal{Z}|}$ that outputs unnormalized mixture scores over retrievers. At inference time, scores are converted into mixture weights on the probability simplex via softmax:

$$\mathbf{w} = \mathrm{softmax}\big(\hat{\mathbf{w}}\big) \in \Delta^{|\mathcal{Z}|-1}. \tag{2}$$

Training targets are derived from evaluation logs. For each query q, we define a per-retriever correctness vector $y(q) \in \{0, 1\}^{|\mathcal{Z}|}$, where each entry indicates whether the downstream LLM, when prompted using a single retrievers Top-K demonstrations, produces the correct answer. When multiple retrievers succeed, we optionally normalize $y(q)$ to distribute credit.

We train the regressor with a squared-error objective:

$$\mathcal{L}_{\text{mix}} = \big\| f_\theta(\phi(q)) - y(q) \big\|_2^2.$$

Regularization is provided implicitly through shallow trees, shrinkage, and early stopping in the gradient-boosted model. Importantly, only the parameters θ of f_θ are learned; all retrievers remain frozen. These predicted weights capture prior expectations about retriever performance for a given query type, and are further refined at inference time via the confidence shaping mechanism described above.

Router. Rather than training a separate router, we apply a margin-based confidence gate directly on the predicted mixture weights. If one retriever's weight exceeds the others by a margin τ_{gate}, we bypass fusion and select that retriever alone if:

$$w_{m^*} \geq \max_{m \neq m^*} w_m + \tau_{\text{gate}}, \tag{3}$$

where $m^* = \arg\max_r w_r$.

Fusion. Given predicted mixture weights $\mathbf{w}$, we first apply a margin-based gate. If a single retriever is deemed confidently dominant, we bypass fusion and select that retriever directly. Otherwise, we calibrate retriever scores and perform feature-weighted fusion. The final retrieval distribution is

$$p(c \mid q) = \begin{cases} p_{m^*}(c \mid q), & \text{if (3)}, \\ \sum_{r \in \mathcal{Z}} w_r \, \tilde{p}_r(c \mid q), & \text{otherwise}, \end{cases} \tag{4}$$

where $\tilde{p}_r$ is the calibrated Top-K distribution of retriever r. This margin-based gate avoids unnecessary fusion when one retriever is dominant.

In practice, when fusion is performed, we operate in logit space, $\log p \propto \sum_r w_r \log \tilde{p}_r$, which better preserves relative confidence; probability-space fusion is included as an ablation. The top-K items under $p(\cdot \mid q)$ are then inserted into the in-context prompt for downstream generation.

4.4 Training & Inference Pipeline

Training. All base retrievers (LLM-, BERT-, and MLP-based) are frozen throughout training. The MLP retriever consists of a frozen BERT-base encoder that serves as a feature extractor, and a lightweight linear scoring head. Only the scoring head is trained, using the same reinforcement learning-based pipeline described in DR-IKE (Nafee et al., 2025), while the encoder remains fixed throughout.

We collect supervision from past evaluation logs by running each retriever independently, prompting the downstream LLM with its Top-K examples, and recording whether the generated answer matches the gold standard. These per-retriever correctness signals are used to train the weight regressor f_θ on query-level features $\phi(q)$. No additional LLM calls are required beyond those already incurred during evaluation.

Inference. At test time, each retriever produces a ranked list, which is calibrated only if fusion is required. Query features $\phi(q)$ are computed once and passed to the learned weight predictor. If a single retriever is deemed confidently dominant, FWE-IKE selects it directly; otherwise, it performs feature-weighted fusion to obtain a fused ranking. The resulting Top-K examples are inserted into the prompt for a single downstream LLM call, which matches the cost of standard ICL.

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate on the COUNTERFACT benchmark (Meng et al., 2023a). Following the IKE protocol (Zheng et al., 2023a), we use the first 2,000 items as the *query* pool and the remainder as the *demonstration* pool that supplies COPY/UPDATE/RETAIN candidates.

Training Configuration. Unless noted, experiments subsample a contiguous slice beginning at offset 300 and evaluate on N=300 queries with entity-wise splits to avoid leakage.

For each query q, we retrieve a shared pool of candidates (COPY/UPDATE/RETAIN) via a precomputed neighbor index. The same candidate pool is given to all retrievers so that differences reflect ranking rather than retrieval coverage.

Retrievers. We instantiate three frozen experts: (i) **MLP**: a frozen BERT-base encoder used as a feature extractor, followed by a lightweight MLP scoring head that models nonlinear interactions between query and candidate representations; (ii) **BERT**: a frozen BERT-base bi-encoder with a single linear scoring head that computes similarity between query and candidate embeddings; (iii) **LLM**: a listwise ranker powered by Llama-3.2-3B-Instruct with a compact JSON-scoring prompt and self-consistency. This LLM is *used as a retriever* and is distinct from the target LLM whose knowledge is being edited. All retrievers are frozen and run with `torch.bfloat16` and `device_map=auto`. The LLM ranker outputs raw scores that are z-normalized and softmaxed per query.

Baselines. We compare FWE-IKE with four representative in-context editing strategies: **FactPrompt** (Cohen et al., 2023) which prepends a narrative prefix ("Imagine that..."); **EditCoT** (Wang et al., 2024) which applies chain-of-thought prompting; **IKE** (Zheng et al., 2023a) which selects COPY, UPDATE, and RETAIN examples to inject new facts; and **DR-IKE** (Nafee et al., 2025) which trains a reward-guided BERT retriever and a learnable threshold to adaptively select and size in-context demonstrations.

Evaluation Metrics. We evaluate editing performance using standard metrics from Zheng et al. (2023a). ***Edit Success Rate (ESR)*** quantifies how often the model returns the correct edited object. ***Retention Rate (RR)*** captures locality by checking that the original object remains correct on unrelated neighborhood prompts. ***Paraphrase Consistency (PC)*** assesses generalization by verifying consistent edits under paraphrased prompts. We report the **oracle upper bound**: whether *any* expert's top-K evidence yields a correct answer and compare against the metrics of single retrievers.

5.2 Main Results

Table 1 shows the performance of knowledge editing in different methods using Llama 3.2 Instruct-3b on 100 samples.

Prompt-only baselines (FactPrompt, EditCoT)

Editing Method	ES↑	PC↑	RR↑
FactPrompt	0.7	0.34	0.19
EditCoT	0.74	0.33	0.24
IKE	0.8	0.64	0.6
DR-IKE	0.85	0.71	0.54
FWE-IKE	**0.87**	**0.73**	**0.64**

Table 1: Editing performance on *Llama-3.2-3B-Instruct* for 100 samples across different editing methods. Best in bold; second-best underlined.

Retriever	ES ↑	PC ↑	RR ↑	Unique wins (n)
MLP	0.8267	0.76	0.57	6
BERT	0.8533	0.7667	**0.5767**	9
LLM	0.84	0.77	**0.5767**	7
FWE-IKE	**0.8833**	**0.7767**	0.5733	–
Oracle Upper Bound	0.91	0.7767	0.5733*	–

Table 2: Editing performance and unique-win counts *Llama-3.2-3B-Instruct* for 300 samples across methods. Best deployable results are in bold. The asterisk (*) indicates that the oracle is a non-deployable upper bound only for ESR, so its PC/RR numbers can be lower.

achieve modest ESR but weak PC and RR, indicating limited generalization and locality. Through retrieval-based demonstrations, IKE improves all metrics. DR-IKE further improves ESR and PC through reward-guided retrieval, but lower RR. FWE-IKE achieves the strongest overall performance, attaining the highest ESR, PC, and RR. This improvement across metrics indicates that FWE-IKE effectively balances edit success, generalization, and knowledge retention.

Table 2 summarizes exact-match accuracies for 300 samples in different retrieval methods using Llama 3.2 Instruct-3b.

The best single retriever (BERT) achieves an ESR of 85.33%, with comparable PC and RR across single models. FWE-IKE improves ESR to 88.33%, yielding a +3.0-pt gain over the strongest single retriever and +3.66 pts over the single-model average, while maintaining competitive PC (77.67%) and RR (57.33%). This corresponds to an approximate 20% reduction relative to BERT (14.67% $\rightarrow$ 11.67%). Relative to the non-deployable oracle upper bound (91.00% ESR), FWE-IKE closes roughly 53% of the remaining gap (from 5.67 to 2.67 points), indicating that learned gating and calibrated fusion recover a substantial portion of complementary retriever strength without modifying base model parameters. The unique-wins analysis shows that no single retriever dominates across queries (MLP: 6, LLM: 7, BERT: 9), motivating adaptive, per-query retriever selection. To evaluate robustness across datasets with different structures and paraphrasing

Method	ESR ↑	PC ↑	RR ↑
FactPrompt	0.31	0.27	0.09
EditCoT	0.28	0.29	0.08
IKE	0.31	0.22	0.47
DR-IKE	0.34	0.27	0.50
FWE-IKE	**0.34**	**0.31**	0.39

Table 3: Editing performance on *zsRE* for 300 samples across different editing methods. Best results are in bold.

Method	ESR ↑	PC ↑	RR ↑
FactPrompt	0.31	0.28	0.10
EditCoT	0.31	0.28	0.07
IKE	0.39	0.40	0.61
DR-IKE	0.42	0.43	0.63
FWE-IKE	**0.45**	0.43	0.63

Table 4: Editing performance on *WikiDatasetCounterfactual* for 300 samples across different editing methods. Best results are in bold..

characteristics, we further assess performance on the zsRE and WikiDatasetCounterfactual benchmarks. On zsRE (see Table 3), FWE-IKE achieves an ESR of 0.34, matching DR-IKE (0.34) and outperforming IKE (0.31), FactPrompt (0.31), and EditCoT (0.28). Notably, FWE-IKE improves paraphrase consistency to 0.31 compared to 0.27 for DR-IKE (+0.04 absolute), indicating stronger generalization across linguistic variations, while maintaining competitive retention (0.39 vs. 0.50).

On WikiDatasetCounterfactual (see Table 4), FWE-IKE attains the highest ESR at 0.45, exceeding DR-IKE (0.42) by +0.03 absolute improvement, and outperforming earlier baselines such as IKE (0.39) and FactPrompt/EditCoT (0.31). It also maintains parity with DR-IKE on PC (0.43) and RR (0.63), demonstrating that the gains in edit success do not come at the cost of retention or consistency.

5.3 Ablation Study

Table 5 reports an ablation over fusion space, calibration, gating, calibration hyperparameters, and confidence shaping.

Baseline. Moving from probability-space without calibration (A0, 85.00%) to probability-space with calibration (A1, 85.33%) to logit-space with calibration (A2, 85.33%) shows calibration yields small consistent gains, and logit-space fusion is at least as good as probability-space.

Gating. With a low threshold $\tau_{\text{gate}} = 0.00$, accuracy peaks at 88.33%, a +3.0pt lift over the best no-gate baseline. Tightening the threshold to 0.10, 0.15, and 0.20 reduces gate rate to 3.33, 1.67, and

Name	Gate	Thr.	Calib.	Top$K$	Target H	Fusion	Conf.	γ	ESR (%)	% Gate
A0_prob_noCal_noGate	×	.15	×	12	0.8	prob	×	1.5	85.00	0.0
A1_prob_cal_noGate	×	.15	✓	12	0.8	prob	×	1.5	85.33	0.0
A2_logit_cal_noGate	×	.15	✓	12	0.8	**logit**	×	1.5	85.33	0.0
B_gate0.00_logit_cal	✓	0.00	✓	12	0.8	logit	×	1.5	**88.33**	39.67
B_gate0.10_logit_cal	✓	0.10	✓	12	0.8	logit	×	1.5	87.33	3.33
B_gate0.15_logit_cal	✓	0.15	✓	12	0.8	logit	×	1.5	86.33	1.67
B_gate0.20_logit_cal	✓	0.20	✓	12	0.8	logit	×	1.5	85.67	0.67
C_topk8_logit_cal	✓	0.15	✓	8	0.8	logit	×	1.5	85.00	1.67
C_topk12_logit_cal	✓	0.15	✓	12	0.8	logit	×	1.5	86.33	1.67
C_topk16_logit_cal	✓	0.15	✓	16	0.8	logit	×	1.5	86.67	1.67
C_H0.70_logit_cal	✓	0.15	✓	12	0.7	logit	×	1.5	86.33	1.67
C_H0.80_logit_cal	✓	0.15	✓	12	0.8	logit	×	1.5	86.33	1.67
C_H0.90_logit_cal	✓	0.15	✓	12	0.9	logit	×	1.5	86.00	1.67
D_conf_on_g1.0	✓	0.15	✓	12	0.8	logit	✓	1.0	86.33	1.67
D_conf_on_g1.5	✓	0.15	✓	12	0.8	logit	✓	1.5	86.33	1.67
D_conf_on_g2.0	✓	0.15	✓	12	0.8	logit	✓	2.0	86.33	1.67

Table 5: Ablation Results

0.67%, respectively, with only modest accuracy decreases (87.33, 86.33, 85.67%). Most of the gain is captured even with very selective gating, offering a clean accuracy/efficiency trade-off.

Calibration hyperparameters. For Top-K, $K = 12, 16$ maintains the improvement (86.33-86.67%), while $K = 8$ underperforms (85.00%), indicating that too small a candidate set constrains the ensembles gains. For target entropy H, mid-range values $0.7 – 0.8$ behave similarly (86.33%), whereas 0.9 is slightly worse (86.00%).

Confidence shaping. Activating confidence shaping with $\gamma \in \{1.0, 1.5, 2.0\}$ leaves accuracy unchanged (86.33%, gate = 1.67%), suggesting that once features, calibration, and gating are in place, additional confidence reweighting adds little on this dataset.

Taken together, the table indicates that *calibration + logit fusion with gate on and* $\tau_{gate} = 0.00$ achieves the highest ensemble configuration, attaining 88.33%. The remaining gap to the oracle 91.0% concentrates in the agree-wrong cases where all calibrated distributions are flat or aligned to the same distractors. This is the regime that likely requires either stronger experts or additional routing signals. Finally, the routing rate column (% Gate) shows operational costs: `B_gate0.10` offers a +2.0% absolute over the strongest single expert with only 3.3% overrides.

5.4 Case Studies

We next present three qualitative case studies illustrating when each base ranker excels and how the ensemble responds. We select one unique win for each expert (BERT-only: idx 302; LLM-only: idx 301; MLP-only: idx 360), where only one ranker enables the generator to produce the correct answer. For each case we show the top-5 retained demonstrations from each ranker and the generator's predicted answer in Table 6.

Case A: BERT-only. BERT's top-K emphasizes geo-template cues ("... located in Finland/Germany/Venuzela") and surfaces a closely matched place-relation. The generator then answers "Texas", which matches the gold standard. LLM and MLP on the other hand, favor broader and/or noisier geography, causing the answer to drift to "Herzegovina". The ensemble mirrors BERT's ordering and is also correct. This pattern matches our template-sensitivity hypothesis: BERT's lexical/phrase alignment is decisive when the query is short and structured.

Case B: LLM-only. LLM's listwise scoring surfaces corporate/"headquarters" evidence and steers the generator to Toronto (correct), while BERT and MLP emphasize superficially similar "headquarter" snippets misleading toward Switzerland. The ensemble tips slightly towards the LLM's ordering and remains correct. Here, the global semantic reasoning of the LLM ranker wins over lexical similarity.

Case C: MLP-only. The MLPs pooled-BERT features favor Ireland/Dublin context but, its ranking keeps enough Ireland-adjacent evidence that the generator ultimately answers Rome (correct) for "Irish Ferries" (a counter-intuitive label in the

Retriever	Case A (BERT-only) idx 302 *Query: Jablanica District, in*	Case B (LLM-only) idx 301 *Query: The headquarter of BSC Young Boys is in*	Case C (MLP-only) idx 360 *Query: The headquarters of Irish Ferries is in*
LLM	Veysian... Venezuela; Gavrilovo-Posadsky... Germany; Old City of Bern... Texas; Jakub Kubicki... Paris; Wantirna... Ukraine *Answer: Herzegovina* ×	Univ. of Edinburgh... Istanbul; Glasgow School... Melbourne; Samuel Merritt... Frankfurt; FUN Technologies... Shanghai; Remstar... London *Answer: Toronto* ✓	IndusInd Bank... Dublin; Sukhoi... Dublin; NI Railways... London; County Leitrim... Estonia; Deutsche Bahn... Dublin *Answer: Dublin* ×
BERT	Morawiecki... Warsaw; North Kosovo... Spanish; Khorramdarreh... Finland; Sarnaki... Germany; Veysian... Venezuela *Answer: Texas* ✓	East China Normal... Wellington; Manchester Business School... Cleveland; Rutgers Prep... Texas; Glasgow School... Melbourne; FUN Technologies... Shanghai *Answer: Switzerland* ×	Atletico Roma... Denver; Manchester Central... Dublin; Sullivan Bluth... Moscow; DTC... Boston; NI Railways... London *Answer: Dublin* ×
MLP	Vrienden van het Platteland... Japan; Old City of Bern... Texas; Wantirna... Ukraine; Gavrilovo-Posadsky... Germany; KKBQ... Miami *Answer: Herzegovina* ×	Remstar... London; Nagoya Univ.... Egypt; GMR Group... Toronto; UCDSB... Nevada; Samuel Merritt... Frankfurt *Answer: Switzerland* ×	Ireland... Antarctica; Manchester Central... Dublin; NI Railways... London; IndusInd Bank... Dublin; Sullivan Bluth... Moscow *Answer: Rome* ✓
FWE-IKE	Morawiecki... Warsaw; North Kosovo... Spanish; Khorramdarreh... Finland; Sarnaki... Germany; Veysian... Venezuela *Answer: Texas* ✓	Nagoya Univ.... Egypt; GMR Group... Toronto; Manchester Business School... Cleveland; Remstar... London; FUN Technologies... Shanghai *Answer: Toronto* ✓	Manchester Central... Dublin; Ireland... Antarctica; NI Railways... London; Sullivan Bluth... Moscow; IndusInd Bank... Dublin *Answer: Rome* ✓

Table 6: Case studies showing complementary strengths of LLM, BERT, and MLP retrievers and the effect of our gated fusion (FWE-IKE). ✓ denotes correctness and × denotes incorrectness.

COUNTERFACT slice). BERT and LLM focus on generic "headquarters in Dublin/London" patterns and answer Dublin incorrectly. The ensemble adopts the MLP-weighted mix and thus is correct. Here, a simple, more robust ranker can outperform both BERT and LLM when high-capacity models latch onto frequent but misleading patterns.

6 Conclusion

We introduced FWE-IKE that combines heterogeneous retrievers (LLM, BERT, MLP) through lightweight confidence modeling and query-adaptive fusion. FWE-IKE (i) calibrates each ranker's evidence distribution with top-K truncation and per-query temperature targeting to place scores on a comparable confidence scale, (ii) extracts compact uncertainty cues and (iii) fuses rankers in logit space with learned per-query mixture weights or routes to a single retriever when a conservative margin gate indicates dominance. On a 300-query evaluation, FWE-IKE achieves **88.33%** ESR, closing the gap to the **91.00%** oracle upper bound and consistently outperforming all single retrievers (MLP 82.67%, LLM 84.00%, BERT 85.33%). Ablations show *score calibration* and *logit-space fusion* are beneficial, and a small gate improves robustness with minimal overhead.

Beyond knowledge editing, our framework naturally extends to multi-retriever retrieval-augmented generation (RAG) settings. Conceptually, replacing a single retriever in a standard RAG pipeline with a heterogeneous ensemble and routing them using the proposed feature-weighted mechanism could yield a more robust retrieval system. Recent multi-retriever RAG approaches, such as LTRR (Kim and Diaz, 2025), construct routing features to select a single retriever. In contrast, our results suggest that strict ranking or hard selection alone can be brittle when retrievers exhibit complementary strengths. The hybrid design of margin-based routing combined with calibrated soft fusion offers a more flexible and potentially more robust alternative in heterogeneous retrieval environments, particularly in enterprise settings where data distributions are diverse and evolving. Exploring this connection further presents a promising direction for future work.

These findings highlight the importance of combining complementary retrieval strategies with adaptive, confidence-aware mechanisms to improve robustness and generalization in knowledge editing and beyond.

Limitations

There are several limitations of our work. First, our method is bounded by expert complementarity. Gains vanish when base retrievers agree or are uniformly uncertain. Therefore, richer or more diverse experts would lift the oracle ceiling. Also, our evaluation is constrained by the COUNTERFACT benchmark. While valuable for controlled fact edits, it does not stratify instances by semantic type (e.g., historical vs. geographical vs. numeric vs. technical), difficulty, or distractor density. A richer taxonomy would enable clearer diagnostics (e.g., "BERT dominates templated geography; LLM wins long-range reasoning") and more targeted routing features.

Moreover, the gate relies on shallow confidence diagnostics (entropy, top-1 mass/margin) learned from past slices. When a query falls outside the training distribution or appears in a very sparse regime, these signals can become *uninformative* (e.g., all experts yield flat distributions). In such cases, the router may *over-trust* a spurious peak or *under-route* when a decisive expert exists. More broadly, the calibration and fusion components are learned from limited training data, and thus may reflect dataset-specific patterns. While calibration helps align score scales across experts, it does not eliminate out-of-distribution (OOD) generalization risk under distribution shift. As a result, performance is not guaranteed to transfer to unseen domains, and practitioners should be cautious when deploying the framework in settings that differ substantially from the training distribution. Addressing this OOD gap, through approaches such as domain-adaptive calibration or uncertainty-aware gating, remains an important direction for future work.

Further, we evaluated on a smaller Llama-3.2-3B model due to GPU memory limitations. As such, our experiments are constrained by compute and context-budget considerations; evaluating across a broader range of larger LLMs remains important future work. Smaller models do not do well in editing tasks (Nafee et al., 2025). Also, some experiments (e.g., Table 1) were run on a reduced sample size due to memory constraints when evaluating all baselines, though larger-scale runs (Table 2) show consistent trends. Finally, results are on a single IKE slice and candidate pool; broader datasets and generators are needed to assess generality and latency-cost trade-offs.

Acknowledgement

This work is supported in part by the National Science Foundation (NSF) grant IIS-2451436 and Commonwealth Cyber Initiative grant HC-2Q26-032.

References

Keqin Bao, Ming Yan, Yang Zhang, Jizhi Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024. Real-time personalization for llm-based recommendation with customized in-context learning.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models.

Gordon Cormack, Charles Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. pages 758–759.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Preprint*, arXiv:2101.03961.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards llm-driven dialogue state tracking.

Runyu He, Anyu Ying, and Xiaoyu Hu. 2024. Improving opendevin: Boosting code generation llm through advanced memory management. *Applied and Computational Engineering*, 68:313–320.

Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Michael Jordan and Robert Jacobs. 1994. Hierarchical mixtures of experts and the. *Neural computation*, 6:181–.

To Eun Kim and Fernando Diaz. 2025. Ltrr: Learning to rank retrievers for llms. *Preprint*, arXiv:2506.13743.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *Preprint*, arXiv:2006.16668.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Preprint*, arXiv:2104.08786.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023a. Mass-editing memory in a transformer.

Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li, and Qianren Wang. 2024. Exploring the impact of table-to-text methods on augmenting llm-based question answering with domain hybrid data.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Dyknow: Dynamically verifying time-sensitive factual knowledge in llms.

Mahmud Wasif Nafee, Maiqi Jiang, Haipeng Chen, and Yanfu Zhang. 2025. Dynamic retriever for in-context knowledge editing via policy optimization. *Preprint*, arXiv:2510.21059.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. volume 3, pages 333–389.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Changyue Wang, Weihang Su, Qingyao Ai, Yichen Tang, and Yiqun Liu. 2025. Knowledge editing through chain-of-thought.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models.

Shuyang Yu, Runxue Bao, Parminder Bhatia, Taha Kass-Hout, Jiayu Zhou, and Cao Xiao. 2025. Dynamic uncertainty ranking: Enhancing retrieval-augmented in-context learning for long-tail knowledge in llms.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. Can we edit factual knowledge by in-context learning? *Preprint*, arXiv:2305.12740.

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

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. 2025. Is your llm outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 74337457. Association for Computational Linguistics.