

Mitigating Memory Interference in Lifelong Model Editing through Multi-Head Hopfield Retrieval

Yue Che^{1,2*} Chen Guoyi^{3*} Xijian Deng⁴
Yujie Wu^{2†} Jiang Cai¹ Xuming Ran^{1,3,5†} Mingkun Xu^{1†}

¹Guangdong Institute of Intelligence Science and Technology

²The Hong Kong Polytechnic University ³OpenMIA Research

⁴Xi'an Jiaotong University ⁵National University of Singapore

{cheyue, xumingkun}@gdiist.cn {guoyi, ryan}@openmia.ai yu-jie.wu@polyu.edu.hk

Abstract

Lifelong model editing seeks to continuously incorporate new knowledge into Large Language Models (LLMs) without costly retraining. Memory-augmented editors preserve pre-trained weights to protect unrelated capabilities, but struggle to balance Generalization to rephrased edit queries and Locality on non-edited knowledge. This challenge becomes particularly severe under dense editing settings, where semantically related edits induce substantial *memory interference*. To tackle this problem, we propose **MH-HoReN**, a multi-head Hopfield retrieval framework that performs retrieval across complementary subspaces and combines head-wise matching signals for more reliable memory selection. We further extend Hopfield-based editing into a clustered batched scheme that combines shared codebook memories for related edits with token-level pooling for efficient knowledge storage and query routing. Extensive experiments across diverse LLMs and editing benchmarks demonstrate that MH-HoReN achieves better locality-generalization trade-offs as editing scales increase, improving the sum of Locality and Generalization by more than **12%** over existing methods while preserving almost perfect Reliability in sequential editing, alongside **7.7%** gains in Overall Performance under batched editing. Our code is available [here](#).

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across natural language understanding and generation tasks. Despite these advances, LLMs still produce hallucinations (Ji et al., 2023), perpetuate biases (Ferrara, 2023), and suffer from factual decay (Cao et al., 2021). Moreover, the parametric knowledge acquired during pre-training gradually becomes outdated or inconsistent as the world evolves (Meng

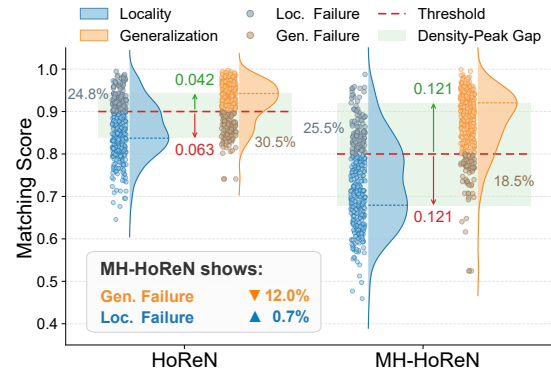


Figure 1: Visualization of retrieval score distributions on LLaMA-3.1 after 400 edits on WikiBigEdit. MH-HoReN reduces overlap between rephrased and unrelated queries, resulting in less ambiguous retrieval.

et al., 2022; Hartvigsen et al., 2023). Such errors accumulate during deployment and often require immediate correction rather than expensive data recollection and full retraining or finetuning. To address this issue, lifelong model editing (Hartvigsen et al., 2023) has emerged as an efficient paradigm for continuously updating LLM knowledge.

An effective lifelong model editor should jointly optimize three key properties (Wang et al., 2024a): **Reliability** on edited queries, **Generalization** to semantically equivalent rephrasings, and **Locality** on unrelated pretrained knowledge. Parameter-modifying methods (Meng et al., 2022, 2023; Kirkpatrick et al., 2016) typically follow a locate-and-edit paradigm that confines updates to edit-associated parameters. Recent approaches further improve stability through lifelong normalization and null-space projection that constrains perturbation directions (Gu et al., 2026; Fang et al., 2025). However, as edits accumulate, protection for earlier edits gradually weakens, leading to reliability drift, while direct parameter modification makes preserving locality inherently difficult. In contrast, parameter-preserving methods keep base parameters untouched and instead maintain aug-

*Equally contributing authors.

†Corresponding authors.

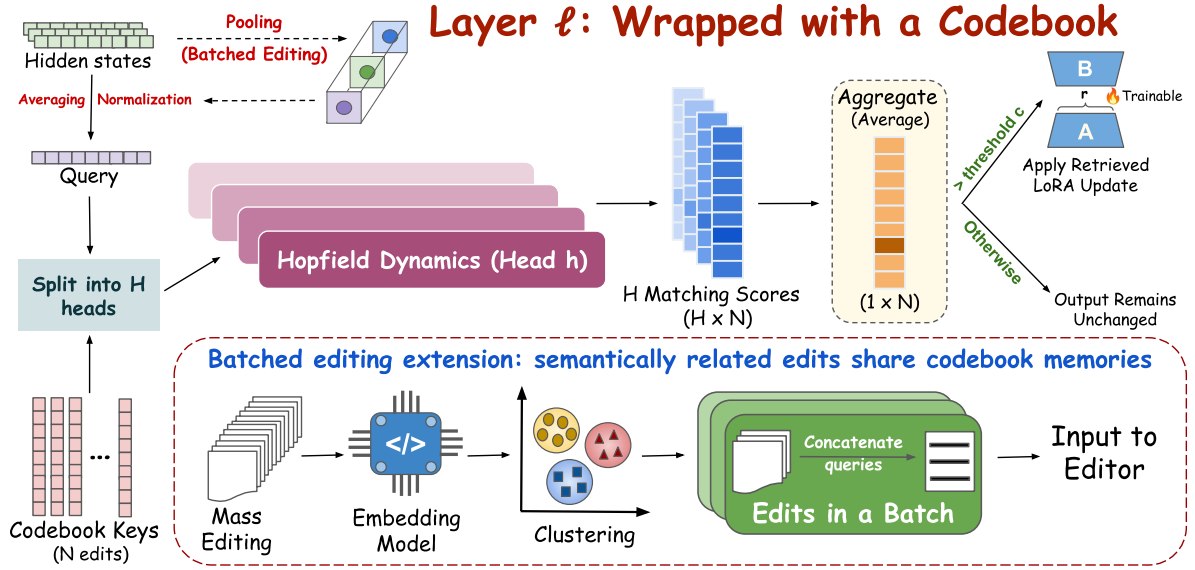


Figure 2: Overview of MH-HoReN. Queries and codebook keys are decomposed into multiple heads, where independent Hopfield dynamics are performed before aggregating head-wise matching scores for retrieval. MH-HoReN further supports batched editing by clustering semantically related edits into shared codebook memories.

mented memories. Existing approaches rely on nearest-neighbor retrieval over layer-wise codebooks (Hartvigsen et al., 2023) or dynamically routed side memories (Wang et al., 2024a). The former preserves locality through acceptance-radius matching, but remains effective mainly for exact edit prompts, resulting in almost no generalization to paraphrases. The latter gradually suffers from degraded side-memory quality as edit streams increase, while threshold-based assignment becomes unreliable due to deteriorating decision boundaries.

A recent method (Fang et al., 2026) based on Modern Hopfield Networks (Ramsauer et al., 2021) replaces one-shot nearest-neighbor matching with iterative associative retrieval, improving robustness and generalization. However, this approach still faces a fundamental challenge under dense editing scenarios, preventing it from optimally balancing locality and generalization. As codebook memories become increasingly crowded, representations of edited knowledge and unrelated inputs become geometrically entangled, a phenomenon we term *memory interference*. Such interference arises from shared lexical patterns, overlapping semantic attributes, and biases introduced during LLM pre-training. Moreover, this method is limited to sequential editing and lacks support for large-scale batched editing needed for practical deployment.

To address these limitations, we propose **MH-HoReN**, **M**ulti-**H**ead **H**opfield **R**etrieval with **N**ormalized representations, a multi-head variant of Hopfield retrieval. Inspired by multi-head at-

tention (Vaswani et al., 2017), MH-HoReN partitions representations into independent subspaces and performs damped Hopfield attractor steps separately within each subspace. Different heads capture complementary semantic aspects of the input geometry, while score aggregation across heads smooths retrieval noise and stabilizes routing decisions, thereby alleviating memory interference. As illustrated in Figure 1, multi-head retrieval reshapes the matching score distributions of rephrasings and unrelated inputs, reducing overlap in their high-density regions. This allows paraphrased queries to relax toward their correct attractor basins while enabling threshold-based filtering to more effectively reject unrelated inputs, thereby improving generalization without sacrificing locality. In addition, we introduce a batched editing framework that combines semantic clustering with token-level pooling to jointly process similar edits. Under this setting, multi-head Hopfield dynamics further decouple mixed edit patterns within shared codebook memories, enabling lower editing latency, reduced parameter storage, and stronger editing performance. The framework of MH-HoReN is shown in Figure 2.

2 Related Work

Parameter-Modifying Model Editing: Parameter-modifying methods inject new knowledge by directly updating pretrained model weights. Locate-then-edit approaches such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) localize

knowledge-associated layers and apply targeted low-rank updates, while meta-learning methods such as KE (Cao et al., 2021) and MEND (Mitchell et al., 2022a) learn gradient transformations for efficient editing. Recent advances enhance editing stability by incorporating null-space constrained updates (Fang et al., 2025) and lifelong normalization (Gu et al., 2026). However, directly encoding edits into base parameters inevitably perturbs unrelated outputs and gradually destabilizes earlier edits under dense editing settings.

Memory-Augmented Model Editing: Memory-based editors store edits outside the base parameters and focus on accurately routing queries to the corresponding memories during inference. SERAC (Mitchell et al., 2022b) uses external memory to condition generation on retrieved edits. GRACE (Hartvigsen et al., 2023) maintains a local codebook and performs nearest-neighbor retrieval with acceptance-radius matching. WISE (Wang et al., 2024a) introduces a dual parametric memory scheme with dynamically routed memory shards. REPAIR (Wang et al., 2025) progressively reintegrates edited knowledge through adaptive intervention mechanisms. HoReN (Fang et al., 2026) further introduces attractor dynamics based on Modern Hopfield Networks (Ramsauer et al., 2021), allowing queries to iteratively converge toward the correct memory basins. However, as edits accumulate, edit-related knowledge and unrelated queries become increasingly entangled in the representation space, leading to ambiguous retrieval. In addition, HoReN remains limited to sequential updates, leaving efficient batched editing underexplored.

Architectural Memory Mechanisms: A complementary line of work improves long-term adaptation through architectural memory design. Memory-oriented architectures such as MemoryLLM (Wang et al., 2024b), SPALM (Yogatama et al., 2021), Memoria (Park and Bak, 2024), and NKB (Dai et al., 2023) introduce persistent memory modules to enhance long-context reasoning and knowledge retention. However, these approaches typically require architectural modifications or additional training procedures, limiting their compatibility with existing pretrained LLM families.

3 Preliminary

3.1 Hopfield Retrieval for Model Editing

HoReN (Fang et al., 2026), built upon Modern Hopfield Networks, inherits the overall architec-

ture of GRACE (Hartvigsen et al., 2023) by wrapping a single selected MLP layer with a local codebook that stores edited knowledge. Unlike GRACE, which relies on one-shot nearest-neighbor lookup, HoReN treats each stored key as an attractor basin and uses iterative Hopfield updates to gradually route queries toward the correct memories, while normalization constrains retrieval to angular similarity. As a result, HoReN improves generalization to rephrased edit queries while preserving the strong locality inherited from GRACE.

Given a frozen model f_θ and N sequential edits $\{(x_i, y_i^*)\}_{i=1}^N$, HoReN maintains a codebook

$$\mathcal{C} = \{(\mathbf{k}_i, A_i, B_i, y_i^*)\}_{i=1}^N,$$

where $\mathbf{k}_i \in \mathbb{R}^d$ is the normalized retrieval key associated with edit (x_i, y_i^*) , and (A_i, B_i) are the corresponding LoRA matrices trained on this edit pair. Edits with semantically similar inputs and identical targets y^* are merged into the same codebook entry. For each edit, hidden states extracted from layer l of $f_\theta(x_i)$ are averaged over the last $n\%$ tokens, normalized, and stored as the codebook key $\mathbf{k}_i$. During inference, a query $\mathbf{q} \in \mathbb{R}^d$ is extracted from the same layer and initialized as $\mathbf{q}^{(0)} = \mathbf{q}/\|\mathbf{q}\|_2$. Damped Hopfield dynamics then iteratively update the query by

$$\begin{aligned} \tilde{\mathbf{q}}^{(t)} &= \text{normalize}\left(\text{softmax}\left(\beta \mathbf{q}^{(t)} \mathbf{K}^\top\right) \mathbf{K}\right), \\ \mathbf{q}^{(t+1)} &= \text{normalize}\left((1 - \gamma)\mathbf{q}^{(t)} + \gamma \tilde{\mathbf{q}}^{(t)}\right). \end{aligned} \quad (1)$$

where $\mathbf{K} \in \mathbb{R}^{N \times d}$ denotes the normalized codebook keys, β controls retrieval sharpness, and γ is a damping coefficient. After M iterations, the matched edit is selected by

$$i^* = \arg \max_i \left(\mathbf{q}^{(M)} \mathbf{K}^\top \right)_i. \quad (2)$$

The corresponding LoRA update (A_{i^*}, B_{i^*}) is injected into layer $l + 1$ in a residual-like manner, while queries with matching scores below the retrieval threshold preserve their original outputs.

3.2 Memory Interference

As the codebook expands monotonically under continual edit streams, semantically or lexically related keys can become crowded in the representation space, making query routing increasingly ambiguous. Although distance metrics and Hopfield dynamics usually retrieve the correct memory for exact edit prompts, two failure cases remain under

dense editing settings: a paraphrased query may be routed to an incorrect key, or an unrelated input may be assigned to an incorrect codebook entry and receive an unintended LoRA update. We refer to these failures as *memory interference*. Let $s_i(\mathbf{q})$ denote the final matching score between query $\mathbf{q}$ and codebook key $\mathbf{k}_i$, and let c be the retrieval threshold. Memory interference occurs when

$$\begin{cases} \exists \mathbf{q} \in \mathcal{P}_j, i \neq j : s_i(\mathbf{q}) \geq s_j(\mathbf{q}), \\ \exists \mathbf{q} \in \mathcal{L}, i : s_i(\mathbf{q}) \geq c, \end{cases} \quad (3)$$

where $\mathcal{P}_j$ denotes paraphrased queries of edit j , and $\mathcal{L}$ denotes unrelated locality queries. In this regime, retrieval decisions become unreliable, making it difficult to improve generalization without compromising locality.

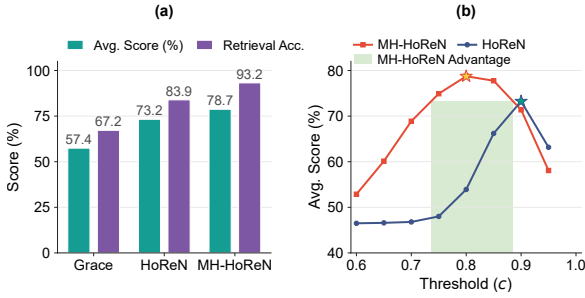


Figure 3: Argmax retrieval accuracy and threshold ablation on WikiBigEdit using LLaMA-3.1-8B-Instruct under 1000 and 500 edits, respectively. MH-HoReN achieves more robust memory routing and maintains stable performance across a broader range of thresholds.

4 Methodology

Motivation Study: Geometrically similar codebook keys form competing attractors and cause memory interference, degrading retrieval reliability. As shown in Figure 3(a), although HoReN improves retrieval robustness over GRACE, its single-head Hopfield retrieval mechanism still yields sub-optimal argmax retrieval accuracy. Under dense editing, paraphrased queries increasingly drift toward incorrect memory basins as angular separation between competing keys deteriorates. In contrast, MH-HoReN distributes Hopfield dynamics across independent subspaces, allowing different heads to capture complementary attractor trajectories and improving retrieval accuracy by over 9%. Meanwhile, HoReN performs Hopfield updates within a single semantic projection, making threshold selection highly sensitive. As illustrated in Figure 3(b), increasing the retrieval threshold c

improves locality at the expense of generalization, while decreasing c produces the opposite effect. By integrating head-wise matching signals, MH-HoReN reduces sensitivity to local geometric bias and improves the optimal locality-generalization average over HoReN by 5.5%, while remaining robust across a much wider threshold range.

Multi-Head Hopfield Retrieval: To mitigate memory interference under dense editing settings, we extend normalized Hopfield retrieval into a multi-head architecture. Instead of relying on a single attractor trajectory, MH-HoReN factorizes retrieval dynamics across parallel semantic partitions, where damped Hopfield updates are performed separately within each head.

Given a normalized query $\mathbf{q} \in \mathbb{R}^d$ and normalized codebook keys $\mathbf{K} \in \mathbb{R}^{N \times d}$, where N denotes the number of stored edits, we divide the embedding dimension into H heads with per-head dimension $d_k = d/H$, obtaining

$$\mathbf{q} \in \mathbb{R}^{H \times 1 \times d_k}, \quad \mathbf{K} \in \mathbb{R}^{H \times N \times d_k}.$$

Hopfield dynamics are independently applied within each head using Eq. 1. Unlike HoReN, which performs retrieval within a single projection space, multi-head retrieval exposes queries to diverse geometric views of the codebook, improving discrimination among attractor basins for geometrically proximal queries. After M retrieval iterations, head-wise matching scores are computed by

$$\mathbf{S}_h = \mathbf{q}_h^{(M)} \mathbf{K}_h^\top, \quad \mathbf{S} = \bigoplus_{h=1}^H \mathbf{S}_h \in \mathbb{R}^{H \times N}. \quad (4)$$

The final retrieval decision is obtained by averaging head-wise matching scores and selecting the highest-ranked edit. Combining retrieval evidence across heads smooths noisy retrieval trajectories, suppresses spurious activations from competing attractors, and stabilizes query routing:

$$\bar{\mathbf{S}} = \frac{1}{H} \sum_{h=1}^H \mathbf{S}_h, \quad i^* = \arg \max_i \bar{\mathbf{S}}_i. \quad (5)$$

The retrieved LoRA update is then applied through the same residual-style injection used in HoReN.

By distributing retrieval across complementary subspaces, MH-HoReN alleviates interference among competing memory basins, enabling more reliable convergence toward correct edits while better rejecting unrelated inputs under dense and highly crowded editing scenarios.

Model	Method	$N = 100$				$N = 500$				$N = 1000$			
		Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.
GPT-OSS 20B	GRACE	76.28	1.25	100.00	59.18	72.66	1.34	100.00	58.00	73.12	1.36	100.00	58.16
	WISE	33.40	32.70	100.00	55.70	34.77	33.33	99.99	56.03	34.39	33.32	99.91	55.87
	UltraEdit	75.10	64.55	93.48	77.70	75.24	66.11	90.08	77.14	68.58	62.63	85.36	72.19
	AlphaEdit	87.65	68.95	76.75	77.78	82.92	60.13	74.15	72.40	75.55	53.35	68.15	65.68
	HoReN $_{H=1}$	100.00	92.87	99.00	97.29	99.37	88.66	99.00	95.68	99.46	87.17	99.30	95.31
	MH-HoReN $_{H=4}$	100.00	96.57	100.00	98.86	99.37	96.14	99.80	98.44	99.53	95.21	99.90	98.21
LLaMA-3.1 8B-Instruct	GRACE	100.00	6.40	100.00	68.80	99.80	4.54	100.00	68.11	99.80	5.06	100.00	68.29
	WISE	77.66	70.51	100.00	82.72	71.83	66.84	99.93	79.53	55.48	52.70	99.56	69.25
	UltraEdit	90.65	89.33	69.24	83.07	89.00	87.23	60.00	78.74	87.89	85.71	54.78	76.13
	AlphaEdit	98.17	91.54	90.27	93.32	98.71	92.75	81.43	90.96	97.78	91.94	74.16	87.96
	HoReN $_{H=1}$	100.00	84.70	100.00	94.90	100.00	90.75	98.39	96.38	99.98	91.69	95.23	95.63
	MH-HoReN $_{H=64}$	100.00	83.70	100.00	94.57	100.00	91.15	99.40	96.85	99.98	91.48	98.52	96.66

Table 1: Sequential editing results on ZsRE under varying editing scales across GPT-OSS-20B and LLaMA-3.1-8B-Instruct. MH-HoReN demonstrates stronger Overall Performance as editing becomes denser, showing improved robustness against increasing memory interference while maintaining near-perfect Reliability and Locality.

Batched Editing with Shared Memory: To support more efficient mass editing, we extend Hopfield-based editors into a batched framework. Edit queries are first embedded and clustered, allowing semantically related edits to share codebook memories and jointly optimize shared LoRA parameters. Beyond reducing codebook size and parameter overhead, semantic clustering also improves gradient consistency during LoRA optimization, encouraging more specialized and stable adapter updates within each batch.

For each batch, edit queries are concatenated into a single sequence before extracting hidden states from layer l . The resulting representations are aggregated through lightweight token-level pooling, followed by averaging and normalization to construct the codebook key k_i . This construction induces a shared attractor landscape that contains multiple correlated edit patterns, where each edit contributes a partial semantic component to the resulting memory basin. Since queries within the same batch are semantically aligned, the combined representation becomes better localized around the batch-level edit semantics, reducing unintended activation from queries belonging to other batches. The additional pooling operation further reinforces locally consistent patterns before global aggregation, producing smoother and more stable attractors for retrieval. As a result, related edits can more reliably converge toward their shared memory despite the compressed codebook structure.

Unlike sequential editing, a batched key is formed from a concatenated edit sequence, so each individual edit is only partially expressed in the shared representation. Such mixed patterns are difficult for single-head retrieval to resolve, as a sin-

gle retrieval trajectory struggles to recover partially embedded information. In contrast, MH-HoReN distributes retrieval across heads, allowing different subspaces to recover relevant retrieval cues from the shared memory landscape. This essentially improves routing accuracy for edit-related queries under batched editing settings.

5 Experimental Results

5.1 Experimental Setup

We evaluate MH-HoReN on Qwen3-1.7B (Team, 2025), Qwen2.5-7B-Instruct (Yang et al., 2024), DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025), LLaMA-3.1-8B-Instruct (Team, 2024), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), and GPT-OSS-20B (OpenAI, 2025). Experiments are conducted on two widely used lifelong model editing benchmarks: ZsRE (Levy et al., 2017) and WikiBigEdit (Thede et al., 2025). ZsRE evaluates factual editing together with generalization to paraphrases, while WikiBigEdit introduces stronger semantic overlap between edit-related and unrelated queries. We further conduct detailed ablations on each proposed component, including the number of retrieval heads, clustering, and token-level aggregation. MH-HoReN is compared against parameter-preserving editors GRACE, WISE, and HoReN, which preserve frozen base weights, as well as parameter-modifying methods AlphaEdit and UltraEdit that directly update pretrained model parameters. Following prior lifelong editing protocols (Wang et al., 2024a), we report Reliability (Rel.), Generalization (Gen.), and Locality (Loc.), together with Overall Performance (O.P.), defined as the average of these

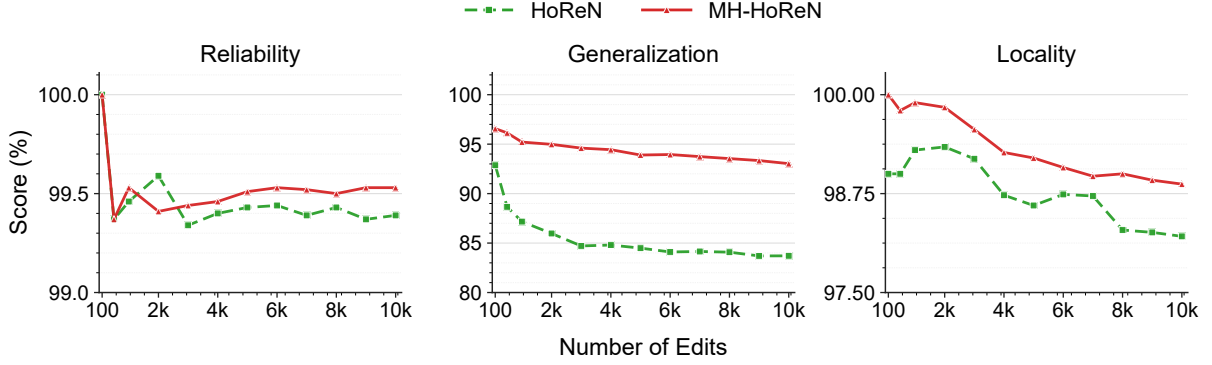


Figure 4: Scaling sequential editing on ZsRE from 100 to 10K edits using GPT-OSS-20B. MH-HoReN consistently preserves stronger Generalization and Locality than HoReN, with the performance gap widening as the number of edits increases, while maintaining almost perfect Reliability.

three metrics. Detailed hyperparameter settings, metric formulations, and efficiency analysis are provided in Appendix A.3, Appendix A.4, and Appendix A.9, respectively. All experiments are conducted on NVIDIA H100 GPUs with a random seed of 42. To additionally assess robustness, we conduct experiments with fixed hyperparameters over three random seeds (11, 22, and 33) and report the mean and standard deviation of each metric in Appendix A.6.

5.2 Sequential Editing on ZsRE

ZsRE (Levy et al., 2017) is a widely used knowledge editing benchmark built from question-answer pairs. Each editing instance contains an original edit query with its target answer, a semantically equivalent paraphrase for testing whether the edited knowledge transfers beyond the original wording, and a locality prompt used to examine whether unrelated knowledge is preserved.

Table 1 presents sequential editing results on ZsRE under 100, 500, and 1000 edits for GPT-OSS-20B and LLaMA-3.1-8B-Instruct. Additional results on Qwen3-1.7B are provided in Table 9. For GPT-OSS, MH-HoReN improves Generalization over HoReN by 3.7% at $N = 100$. This gap further increases to around 8% under denser editing settings with $N = 1000$, while MH-HoReN still preserves nearly perfect Reliability and Locality. Consequently, MH-HoReN improves O.P. over HoReN by 2.9% and over UltraEdit by more than 26%. For LLaMA-3.1, MH-HoReN shows improved robustness as the editing scale grows. When the number of edits increases by an order of magnitude from $N = 100$ to $N = 1000$, single-head HoReN experiences nearly a 5% drop in Locality due to increasingly crowded attractor regions. In contrast, MH-HoReN shows a Locality drop by less

Method	Rel.	Gen.	Loc.	O.P.
<i>LLaMA-3.1-8B-Instruct</i>				
GRACE	99.72	14.85	100.00	71.52
WISE	76.29	71.07	38.99	62.12
UltraEdit	86.61	83.02	44.92	71.52
AlphaEdit	93.34	85.99	44.27	74.53
HoReN $_{H=1}$	99.88	68.37	76.65	81.63
MH-HoReN $_{H=1024}$	99.88	81.81	75.57	85.75
<i>GPT-OSS-20B</i>				
GRACE	60.74	0.30	100.00	53.68
WISE	22.45	22.41	82.49	42.45
UltraEdit	72.84	67.15	32.80	57.60
AlphaEdit	57.71	40.52	28.65	42.30
HoReN $_{H=1}$	99.08	67.99	46.38	71.15
MH-HoReN $_{H=4}$	99.16	75.49	44.56	73.07

Table 2: Sequential editing results on WikiBigEdit after 1000 edits across two models. MH-HoReN consistently achieves better locality-generalization trade-offs than existing baselines while also obtaining higher O.P.

than 1.5% and still outperforms HoReN by around 3.3%. Moreover, MH-HoReN improves O.P. over the parameter-modifying baseline AlphaEdit by 8.7%, while outperforming WISE and GRACE by over 27% and 28%, respectively.

We further observe that the optimal number of retrieval heads varies considerably across model families. GPT-OSS achieves the best performance with $H = 4$, whereas LLaMA-3.1 performs best with $H = 64$, suggesting that models with weaker semantic separation benefit more from finer-grained subspace partitioning to reduce retrieval bias and improve routing toward correct memory basins. Additional ablations on the number of retrieval heads are provided in Section 5.6.

5.3 Scaling to Large Editing Scales

To further stress retrieval under dense lifelong editing settings, we progressively increase the number

Method	N = 500				N = 1000				N = 2000			
	Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.
<i>Qwen2.5-7B-Instruct</i>												
WISE	24.44	23.39	90.16	45.99	24.25	23.55	85.95	44.58	23.05	22.73	77.94	41.24
UltraEdit	70.91	60.70	59.51	63.71	67.38	59.25	40.74	55.79	63.96	58.16	30.18	50.76
AlphaEdit	98.14	74.28	48.99	73.80	97.96	74.99	47.73	73.56	98.23	75.71	51.81	75.25
HoReN _{H=1}	90.91	83.39	96.39	90.23	83.05	77.20	93.44	84.57	79.20	71.16	94.65	81.67
MH-HoReN _{H=128}	93.75	86.95	93.78	91.49	89.05	82.13	92.08	87.75	87.85	79.40	92.31	86.52
<i>w/o Pooling</i>	61.05	50.86	100.00	70.64	59.83	47.89	100.00	69.24	58.81	47.87	100.00	68.89
<i>w/o Clustering</i>	89.92	82.30	93.80	88.67	90.22	81.42	94.03	88.56	86.11	77.13	93.57	85.60
<i>Qwen3-1.7B</i>												
WISE	51.63	48.72	99.95	66.77	44.08	42.09	99.98	62.05	39.63	37.65	99.99	59.09
UltraEdit	53.02	48.66	91.09	64.26	53.31	49.41	85.10	62.61	53.18	49.57	58.83	53.86
AlphaEdit	94.01	70.84	82.87	82.58	88.37	65.17	77.51	77.02	84.38	61.19	73.70	73.09
HoReN _{H=1}	90.40	84.58	89.88	88.29	89.10	79.92	88.52	85.85	81.83	74.74	89.28	81.95
MH-HoReN _{H=1024}	98.00	93.28	87.02	92.77	97.31	90.19	90.59	92.70	94.92	87.60	86.42	89.65
<i>w/o Pooling</i>	10.40	6.80	100.00	39.07	10.27	5.40	100.00	38.56	11.19	5.95	100.00	39.05
<i>w/o Clustering</i>	94.85	89.00	92.24	92.03	94.41	86.00	91.69	90.70	91.80	85.15	89.53	88.82

Table 3: Batched editing results on ZsRE and ablations on pooling and clustering under varying editing scales across two Qwen-series models. While reducing editing latency and codebook storage through clustered shared-memory construction, MH-HoReN maintains superior Overall Performance over competing batched editing methods.

of edits on GPT-OSS from 100 to 10K on ZsRE, largely increasing interference among codebook attractors and making correct retrieval increasingly difficult. As shown in Figure 4, both HoReN and MH-HoReN consistently maintain almost perfect Reliability and Locality throughout the entire editing process, with MH-HoReN preserving slightly stronger Locality across all editing scales. However, the gap in Generalization becomes increasingly pronounced as the number of edits grows. Under highly crowded codebook memories, even small angular deviations between paraphrases and their original edit prompts can cause retrieval trajectories to drift toward incorrect attractor basins. By decomposing retrieval into four semantic subspaces ($H = 4$), MH-HoReN captures more diverse geometric patterns and greatly improves routing accuracy for rephrased edit queries. Under 10K edits, MH-HoReN ultimately improves Generalization over single-head HoReN by around 10%.

5.4 Sequential Editing on WikiBigEdit

WikiBigEdit (Thede et al., 2025) is a dense life-long editing benchmark constructed from Wikidata triples. Unlike ZsRE, its locality queries are intentionally designed to share strong semantic overlap with edited facts, making unrelated inputs geometrically close to edit-related queries. This setting challenges an editor’s ability to preserve unrelated model behaviors under dense editing and,

for retrieval-based editors with threshold filtering, substantially intensifies the locality-generalization trade-off discussed in Section 4.

Table 2 reports WikiBigEdit results after 1000 edits for LLaMA-3.1-8B-Instruct and GPT-OSS-20B. Additional results for DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-1.5B are presented in Table 10. MH-HoReN achieves the strongest Overall Performance on LLaMA-3.1, outperforming GRACE and UltraEdit by more than 14%, and surpassing AlphaEdit by over 11%. In this setting, MH-HoReN aggregates retrieval outputs from 1024 complementary Hopfield heads, enabling more expressive retrieval behavior and more reliable query routing than single-head HoReN. Consequently, MH-HoReN improves Gen. by more than 13% while reducing Loc. by only around 1%, all while preserving identical Rel. to HoReN. A similar trend is observed on GPT-OSS, where MH-HoReN improves Gen. by 7.5% over HoReN while decreasing Loc. by less than 2%. Meanwhile, MH-HoReN consistently achieves stronger Overall Performance than all competing methods, outperforming UltraEdit and WISE by nearly 15.5% and over 30.5%, respectively.

5.5 Clustered Batched Editing Evaluation

Compared with sequential editing, batched codebook keys are constructed from concatenated edit queries within the same cluster, causing multiple

correlated edit patterns to coexist within a shared attractor basin and making retrieval more challenging. For batched experiments, we use BGE-base-en-v1.5 (Xiao et al., 2023) for feature extraction and MiniBatchKMeans (Pedregosa et al., 2011) for clustering. Max pooling is applied for token-level aggregation. We denote the average cluster size as r . In our experiments, we set $r = 5$, reducing codebook size by $5\times$ relative to the sequential setting while substantially accelerating editing. Larger r further improves editing speed and memory savings, but may sacrifice performance. Detailed ablations are provided in Section 5.6. Since GRACE already exhibits weaker performance than Hopfield-based editors, we compare MH-HoReN against HoReN, WISE, UltraEdit, and AlphaEdit. For fair comparison, we process edits in groups of 5 for these baselines, matching our cluster size.

Results on ZsRE: As presented in Table 3, we evaluate batched editing on ZsRE under 500, 1000, and 2000 edits using Qwen2.5-7B-Instruct and Qwen3-1.7B. MH-HoReN consistently achieves superior Overall Performance. On Qwen2.5, the O.P. gap between MH-HoReN and HoReN increases from more than 1% at $N = 500$ to around 5% at $N = 2000$, accompanied by an approximately 17% improvement in the combined $Rel. + Gen.$ score while limiting the Locality degradation to less than 2.4% under the densest setting. Compared with AlphaEdit, MH-HoReN achieves an improvement of about 14.2% in O.P. at $N = 1000$. A similar scaling trend is observed on Qwen3, where the O.P. advantage over HoReN expands from roughly 4.5% at $N = 500$ to 7.7% at $N = 2000$. Under the most challenging setting, MH-HoReN also improves the combined $Rel. + Gen.$ score by approximately 26% while reducing Loc. by less than 2.9%. Compared with WISE and UltraEdit, MH-HoReN achieves over 30% higher O.P. at $N = 1000$.

Notably, removing max pooling during query and codebook key construction causes severe performance drops, suggesting that local information aggregation is crucial for stabilizing retrieval under densely mixed attractor representations. Removing clustering (i.e., batching randomly sampled edits together) yields only marginal Loc. gains but substantially weakens edit effectiveness. Specifically, without clustering, $Rel. + Gen.$ drops by around 8.5% on Qwen2.5 at $N = 500$, while Loc. improves by only 0.02%; on Qwen3 at $N = 1000$, $Rel. + Gen.$ decreases by over 7%, with only around 1% Loc. gain. This gap becomes more pronounced as r in-

Method	Rel.	Gen.	Loc.	O.P.
<i>Qwen2.5-7B-Instruct</i>				
WISE	21.63	22.43	47.23	30.43
UltraEdit	62.27	50.88	14.13	42.43
AlphaEdit	95.53	62.19	41.59	66.44
HoReN $_{H=1}$	62.53	49.66	41.82	51.34
MH-HoReN $_{H=512}$	73.16	62.27	41.18	58.87

Table 4: Batched editing results on WikiBigEdit after 1000 edits using Qwen2.5. MH-HoReN achieves competitive O.P. under challenging shared-memory retrieval settings with highly overlapping queries.

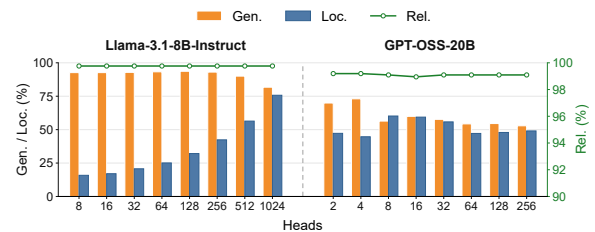


Figure 5: Ablation on the number of retrieval heads under sequential editing on WikiBigEdit with 500 edits across LLaMA-3.1-8B-Instruct and GPT-OSS-20B.

creases, as further discussed in Section 5.6. We also investigate the effect of pooling under sequential editing in Appendix A.8. The results show that applying token-level pooling provides negligible performance improvements, suggesting that additional token aggregation is unnecessary when constructing codebook keys from single edits.

To assess robustness to the additional randomness introduced by clustering, we evaluate shuffled edit orders across three random seeds in Table 12.

Results on WikiBigEdit: Batched editing performance is also examined on WikiBigEdit under 1000 edits with Qwen2.5-7B-Instruct, as reported in Table 4. Compared with ZsRE, WikiBigEdit introduces stronger retrieval interference because locality queries and rephrased edit queries are constructed with high semantic overlap, resulting in greater ambiguity during retrieval. Under this setting, MH-HoReN achieves lower Overall Performance than AlphaEdit, while still outperforming WISE, UltraEdit, and HoReN by approximately 28.4%, 16.4%, and 7.5% in O.P., respectively. These results indicate that although highly ambiguous retrieval scenarios remain challenging for our current batched editing approach, MH-HoReN still retains an advantage through improved retrieval robustness. Developing a more effective batched editing strategy under low semantic separability remains an important direction for future work.

Model	Heads	Rel.	Gen.	Loc.	O.P.
Qwen3-1.7B	$H = 64$	90.00	80.18	97.11	89.09
	$H = 128$	90.20	81.53	96.63	89.45
	$H = 256$	92.60	84.38	95.54	90.84
	$H = 512$	96.80	90.88	90.90	92.86
	$H = 1024$	98.00	93.28	87.02	92.77

Table 5: Effect of retrieval head number in batched editing on ZsRE with Qwen3-1.7B under 500 edits. Performance generally improves with increasing H .

5.6 Ablation Studies

Effect of Retrieval Head Number: Under sequential editing, Figure 5 shows that the Overall Performance of LLaMA-3.1-8B consistently improves as the number of heads increases, with particularly significant gains in Locality. In contrast, GPT-OSS-20B remains relatively stable across different head numbers, achieving comparable performance under $H = 4, 8, 16$, and 32 . We hypothesize that this difference is related to the semantic separability of intermediate representations used for constructing queries and codebook keys. Due to its MoE routing mechanism, GPT-OSS naturally produces expert-specific activations, which implicitly disentangle semantic features. Combined with its larger parameter scale, GPT-OSS already benefits from improved retrieval accuracy under multi-head Hopfield dynamics without requiring aggressive subspace decomposition. In contrast, the intermediate representations of LLaMA-3.1 exhibit weaker geometric discrimination, making larger-scale retrieval decomposition necessary for stabilizing routing toward correct edits. Additional ablation results on H are provided in Appendix A.7.

We then evaluate the effect of head number under batched editing. As shown in Table 5, the Overall Performance of Qwen3-1.7B shows an increasing trend as H grows, demonstrating that multi-head retrieval provides additional benefits under more challenging shared-memory retrieval scenarios. These results further support that the effectiveness of MH-HoReN comes from decomposing retrieval dynamics into complementary subspaces rather than from clustering or pooling alone.

Effect of Cluster Size: We investigate the effect of cluster size r in batched editing, where the number of shared codebook memories is approximately $\lfloor N/r \rfloor$. As shown in Table 6, performance gradually decreases as r increases due to stronger interference within shared attractor basins. Nevertheless, MH-HoReN maintains competitive performance under substantial codebook compression.

Model	Size	Rel.	Gen.	Loc.	O.P.
Qwen3-1.7B	$r = 1$	100.00	96.82	99.60	98.81
	$r = 3$	98.10	93.48	91.83	94.47
	$r = 5$	98.00	93.28	87.02	92.77
	$r = 7$	96.05	84.93	91.29	90.76
	$r = 10$	93.15	85.85	85.46	88.15

Table 6: Ablation on cluster size r for batched editing on ZsRE with 500 total edits using Qwen3-1.7B. A larger r improves efficiency at the cost of editing performance.

Size	Clustering	Rel.	Gen.	Loc.	O.P.
$b = 7$	$\times$	86.51	77.89	91.75	85.38
	$\checkmark$	95.24	84.08	92.31	90.55
$b = 10$	$\times$	85.28	76.34	81.73	81.11
	$\checkmark$	94.34	85.76	85.03	88.38

Table 7: Ablation on semantic clustering for batched editing on ZsRE with Qwen3-1.7B under 1000 edits. The benefit of clustering becomes more pronounced as the number of edits per batch increases.

With $r = 5$, Qwen3-1.7B reduces codebook size by $5\times$ with only around a 6% drop in O.P. compared with sequential editing ($r = 1$). Increasing the cluster size to $r = 10$ further halves codebook storage while still achieving an O.P. of 88.15%.

Effect of Semantic Clustering: We denote the number of edits jointly processed per group as b . Table 7 compares semantic clustering with random grouping on ZsRE with Qwen3 under $N = 1000$ edits. Removing clustering causes an O.P. drop of over 5% at $b = 7$, which increases to more than 7% at $b = 10$. This growing gap indicates that semantic clustering becomes increasingly important as more edits are processed together, because random grouping produces less coherent representations and less specialized LoRA updates, making retrieval and optimization unstable.

6 Conclusion

We present MH-HoReN, a multi-head Hopfield retrieval framework for dense lifelong model editing. By partitioning retrieval into independent subspaces, MH-HoReN mitigates memory interference and achieves stronger locality-generalization trade-offs while preserving high edit reliability across diverse models and benchmarks. We further introduce a clustered batched editing strategy that reduces codebook storage and editing cost through shared attractor representations. Extensive experiments demonstrate that multi-head Hopfield retrieval provides an effective and scalable paradigm for robust continual knowledge editing in LLMs.

Limitations

Although MH-HoReN improves retrieval robustness under dense lifelong editing settings, several limitations remain. The optimal number of retrieval heads varies considerably across model families, requiring additional hyperparameter tuning for different architectures. While our experiments suggest that this behavior is related to the semantic separability of intermediate representations, a deeper theoretical understanding of the relationship between representation geometry and multi-head Hopfield dynamics remains unexplored. Moreover, on challenging benchmarks such as WikiBigEdit, where unrelated queries are intentionally constructed to exhibit strong semantic overlap with edited facts, MH-HoReN still cannot achieve the near-perfect Reliability, Generalization, and Locality observed on simpler benchmarks such as ZsRE, despite substantially improving the locality-generalization trade-off. In addition, although the proposed clustered batched framework significantly improves editing efficiency and reduces codebook storage, highly ambiguous retrieval scenarios and aggressive batch compression can introduce interference among edits within shared attractor basins, leading to performance degradation. More effective batched editing strategies are therefore needed to better handle low semantic separability while maintaining performance closer to sequential editing under larger cluster sizes.

Acknowledgments

This work is supported by the High-Level Talents Innovation Team Project of the Guangdong-Macao In-Depth Cooperation Zone in Hengqin (Grant No. 2630004018939) and National Natural Science Foundation of China (NSFC) under Grant No. 62506084. This work is also supported by OpenMIA.

References

- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.
- Damai Dai, Wenbin Jiang, Qingxiu Dong, Yajuan Lyu, and Zhifang Sui. 2023. [Neural knowledge bank for pretrained transformers](#). In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part II*, Lecture Notes in Computer Science, pages 772–783. Springer.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. [Alphaedit: Null-space constrained knowledge editing for language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yuan Fang, Yi Xie, and Xuming Ran. 2026. [Horen: Normalized hopfield retrieval for large-scale sequential model editing](#). *CoRR*, abs/2605.08143.
- Emilio Ferrara. 2023. [Should chatgpt be biased? challenges and risks of bias in large language models](#). *First Monday*, 28(11).
- Xiaojie Gu, Ziyang Huang, Jia-Chen Gu, and Kai Zhang. 2026. [Ultraedit: Training-, subject-, and memory-free lifelong editing in language models](#). *Trans. Mach. Learn. Res.*, 2026.
- Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. [Aging with GRACE: lifelong model editing with discrete key-value adapters](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12):248:1–248:38.
- James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. [Overcoming catastrophic forgetting in neural networks](#). *CoRR*, abs/1612.00796.
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. [Zero-shot relation extraction via reading comprehension](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information*

- Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.*
- Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. [Mass-editing memory in a transformer](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. [Fast model editing at scale](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. [Memory-based model editing at scale](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pages 15817–15831. PMLR.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.
- Sangjun Park and JinYeong Bak. 2024. [Memoria: Resolving fateful forgetting problem through human-inspired memory architecture](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 39587–39615. PMLR / OpenReview.net.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in python](#). *Journal of Machine Learning Research*, 12(85):2825–2830.
- Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2021. [Hopfield networks is all you need](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Qwen Team. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Thomas Hartvigsen. 2025. [Wikibigedit: Understanding the limits of lifelong knowledge editing in llms](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Hua-jun Chen. 2024a. [WISE: rethinking the knowledge memory for lifelong model editing of large language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Yisu Wang, Ming Wang, Haoyuan Song, Wenjie Huang, Chaozheng Wang, Yi Xie, and Xuming Ran. 2025. [REPAIR: robust editing via progressive adaptive intervention and reintegration](#). *CoRR*, abs/2510.01879.
- Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian J. McAuley. 2024b. [MEMORYLLM: towards self-updatable large language models](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 50453–50466. PMLR / OpenReview.net.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *CoRR*, abs/2309.07597.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Dani Yogatama, Cyprien de Masson d’Autume, and Lingpeng Kong. 2021. [Adaptive semiparametric language models](#). *Trans. Assoc. Comput. Linguistics*, 9:362–373.

A Appendix

A.1 LLM Usage Statement

Large language models were used solely for language polishing and writing refinement. All research ideas, methodology design, experiments, data analysis, and conclusions were independently developed and conducted by the authors.

A.2 Usage and License

All datasets, pretrained models, and baseline implementations used in this work are publicly available and used in accordance with their respective licenses, terms of use, and intended research purposes.

A.3 Baseline Hyperparameters

We follow the hyperparameter choices recommended by the original papers or official implementations whenever applicable. When no released configuration matches a backbone or architecture, we adopt the closest available setting and adapt only model-specific components.

MH-HoReN: In addition to H , we primarily tune the token averaging strategy used for key construction (i.e., averaging over the last $n\%$ tokens) and the retrieval threshold, as these two hyperparameters have the largest impact on performance. The number of retrieval heads H used in MH-HoReN is reported in each corresponding table. Detailed hyperparameter settings are provided in Table 8.

HoReN: We also tune HoReN to ensure optimal overall performance. For sequential editing on ZsRE, LLaMA-3.1 uses the same hyperparameters as MH-HoReN. GPT-OSS uses Strategy=80% and Threshold=0.75, while Qwen3 uses Strategy=90% and Threshold=0.7. On WikiBigEdit, LLaMA-3.1 and GPT-OSS both use Strategy=50%, with Threshold=0.9 and 0.85, respectively. DeepSeek-R1 models use the same hyperparameters as MH-HoReN. For batched editing, Qwen-series models also use the same hyperparameters as MH-HoReN, except that lowering the threshold to 0.6 for Qwen3 on ZsRE yields better performance.

GRACE: We use the official implementation defaults for all backbones. For backbones not covered by the released examples, we only adapt backbone-specific editable modules, layer names, and numerical precision.

WISE: LLaMA-3.1 uses the LLaMA-family down-projection setting. For Qwen3, we edit layer 24 with mask_ratio=0.2, edit_lr=0.7, n_iter=90,

Model	Strategy	Threshold
<i>Sequential Editing on ZsRE</i>		
LLaMA-3.1-8B	50%	0.80
GPT-OSS-20B	10%	0.65
Qwen3-1.7B	90%	0.65
<i>Sequential Editing on WikiBigEdit</i>		
LLaMA-3.1-8B	60%	0.80
GPT-OSS-20B	10%	0.80
DeepSeek-Qwen-1.5B	70%	0.85
DeepSeek-Llama-8B	60%	0.85
<i>Batched Editing on ZsRE</i>		
Qwen3-1.7B	100%	0.65
Qwen2.5-7B-Instruct	90%	0.60
<i>Batched Editing on WikiBigEdit</i>		
Qwen2.5-7B-Instruct	80%	0.75

Table 8: Hyperparameter settings used for MH-HoReN across different models and benchmarks.

and norm_constraint=0.8. GPT-OSS uses the MoE expert down projection; we edit layer 20 with mask_ratio=0.25, edit_lr=1.2, n_iter=100, and norm_constraint=1.2. DeepSeek-Llama and DeepSeek-Qwen use layers 24 and 18, respectively, with mask_ratio=0.1, edit_lr=0.3, n_iter=70, and norm_constraint=0.5.

UltraEdit: LLaMA-family models use the corresponding MLP modules, with learning rate 1×10^{-6} for LLaMA-3.1 and 5×10^{-7} for DeepSeek-Llama. Qwen3 edits layers 18–26 gate/up projections with learning rate 1×10^{-6} . DeepSeek-Qwen edits layers 19–26 gate/up projections with learning rate 1.2×10^{-6} . GPT-OSS uses MoE expert gate-up and down projections, editing layers 20–22 with learning rate 9.5×10^{-7} .

AlphaEdit: LLaMA-3.1 follows the released LLaMA-3 configuration with layers 4–8, a learning rate of 1×10^{-1} , 25 optimization steps, a clamp norm factor of 0.75, and a recomputed LLaMA-3.1 projection matrix. Qwen-series and DeepSeek-Qwen use the Qwen-style loss layer, editable module, and model-specific projection matrix. DeepSeek-Llama follows the LLaMA-style configuration. GPT-OSS uses a layer-8 attention-output setting with v_lr=0.14, 30 optimization steps, and a clamp norm factor of 0.8.

A.4 Metric Definitions

We evaluate knowledge editing using four metrics: Reliability, Generalization, Locality, and Overall Performance. Let $\mathcal{D} = \{(x_i, y_i, x_i^{\text{para}}, x_i^{\text{loc}})\}_{i=1}^N$, where x_i is the original edit prompt, y_i is the target answer, x_i^{para} is its paraphrase, and x_i^{loc} is an

Method	N = 100				N = 500				N = 1000			
	Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.
GRACE	54.15	1.00	100.00	51.72	29.99	0.75	100.00	43.58	32.43	0.64	100.00	44.36
WISE	78.04	74.28	100.00	84.10	53.43	52.33	99.25	68.34	36.74	36.30	76.93	49.99
UltraEdit	47.24	44.98	96.69	62.97	56.77	52.91	93.31	67.66	55.92	51.63	90.54	66.03
AlphaEdit	95.69	71.14	76.75	81.19	94.95	70.65	81.92	82.50	89.52	66.54	78.06	78.04
HoReN _{H=1}	100.00	98.60	100.00	99.53	100.00	96.45	99.60	98.68	100.00	95.52	99.50	98.34
MH-HoReN _{H=2}	100.00	98.60	100.00	99.53	100.00	96.82	99.60	98.81	100.00	96.28	99.50	98.59

Table 9: Sequential editing results on ZsRE for Qwen3-1.7B under 100, 500, and 1000 edits. MH-HoReN achieves comparable or better performance than other baselines across all editing scales.

unrelated locality prompt. We denote the pre-edit model by f_0 and the model after applying all N edits as f_N . All answers are generated with greedy decoding, using the target length for Reliability and Generalization.

Reliability evaluates whether the edited model produces the target answer for the original edit prompt. If $\hat{y}_i^{\text{edit}}$ is the greedy output of f_N on x_i with length $L_i = |y_i|$, then

$$\text{Rel.} = \frac{1}{N} \sum_{i=1}^N \frac{1}{L_i} \sum_{t=1}^{L_i} \mathbb{1}[\hat{y}_{i,t}^{\text{edit}} = y_{i,t}]. \quad (6)$$

Generalization captures the extent to which the edit transfers to semantically equivalent paraphrases. If $\hat{y}_i^{\text{para}}$ is the greedy output of f_N on x_i^{para} with the same length L_i , then

$$\text{Gen.} = \frac{1}{N} \sum_{i=1}^N \frac{1}{L_i} \sum_{t=1}^{L_i} \mathbb{1}[\hat{y}_{i,t}^{\text{para}} = y_{i,t}]. \quad (7)$$

Locality measures whether editing preserves the model’s behavior on unrelated inputs. For each locality prompt x_i^{loc} , we run both the pre-edit model f_0 and the edited model f_N , take the last ℓ_i generated tokens from each output, and compute their token-level agreement:

$$\text{Loc.} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\ell_i} \sum_{t=1}^{\ell_i} \mathbb{1}[z_{i,t}^0 = z_{i,t}^N], \quad (8)$$

where $z_{i,t}^0$ and $z_{i,t}^N$ denote the t -th token in the corresponding suffixes generated by f_0 and f_N . Thus, locality compares the edited model against the original model rather than against a ground-truth target.

Overall Performance is reported as the average of the three evaluation metrics:

$$\text{O.P.} = \frac{\text{Rel.} + \text{Gen.} + \text{Loc.}}{3}. \quad (9)$$

Method	Rel.	Gen.	Loc.	O.P.
<i>DeepSeek-R1-Distill-Qwen-1.5B</i>				
GRACE	48.79	0.26	100.00	49.68
WISE	26.04	25.58	26.31	25.98
UltraEdit	61.08	55.00	25.05	47.04
AlphaEdit	91.54	47.22	64.62	67.79
HoReN _{H=1}	99.93	82.76	52.86	78.52
MH-HoReN _{H=8}	99.93	73.02	67.82	80.26
<i>DeepSeek-R1-Distill-Llama-8B</i>				
GRACE	100.00	2.18	100.00	67.39
WISE	73.24	66.90	23.93	54.69
UltraEdit	82.37	76.77	30.90	63.35
AlphaEdit	93.45	82.56	29.91	68.64
HoReN _{H=1}	99.95	85.04	54.31	79.77
MH-HoReN _{H=8}	99.95	80.37	62.52	80.95

Table 10: Sequential editing results for DeepSeek-R1-series models after 1000 edits on WikiBigEdit. MH-HoReN obtains better locality-generalization trade-offs.

A.5 Extended Evaluations of MH-HoReN

Supplementary results on ZsRE: Table 9 reports sequential editing results on ZsRE for Qwen3-1.7B under 100, 500, and 1000 edits. MH-HoReN achieves the best or tied-best O.P. across all editing scales. Compared with HoReN, MH-HoReN exhibits a growing advantage in Gen. as the number of edits increases, while preserving identical Rel. and Loc. scores. Among memory-based baselines, MH-HoReN surpasses GRACE and WISE in O.P. by more than 55% and approximately 30.5% at $N = 500$, respectively. Compared with parameter-modifying methods, its O.P. at $N = 1000$ exceeds UltraEdit and AlphaEdit by approximately 32.5% and more than 20.5%, respectively.

Supplementary results on WikiBigEdit: Table 10 presents the results on WikiBigEdit for two DeepSeek-R1-series models after 1000 sequential edits. MH-HoReN achieves more favorable locality-generalization trade-offs across both models. For DeepSeek-R1-Distill-Qwen-

Model	Method	ZsRE				WikiBigEdit			
		Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.
GPT-OSS	HoReN _{H=1}	99.39 $\pm$ 0.00	86.23 $\pm$ 0.01	99.50 $\pm$ 0.08	95.04 $\pm$ 0.03	99.23 $\pm$ 0.01	82.92 $\pm$ 0.06	20.93 $\pm$ 0.05	67.69 $\pm$ 0.01
20B	MH-HoReN _{H=4}	99.55 $\pm$ 0.02	91.26 $\pm$ 0.06	100.00 $\pm$ 0.00	96.94$\pm$0.02	99.23 $\pm$ 0.04	81.06 $\pm$ 0.04	37.29 $\pm$ 0.10	72.53$\pm$0.02
LLaMA-3.1	HoReN _{H=1}	99.98 $\pm$ 0.00	83.55 $\pm$ 0.04	100.00 $\pm$ 0.00	94.51 $\pm$ 0.01	99.88 $\pm$ 0.00	80.56 $\pm$ 0.02	53.95 $\pm$ 0.00	78.13 $\pm$ 0.01
8B-Instruct	MH-HoReN _{H=512}	99.98 $\pm$ 0.00	90.82 $\pm$ 0.00	100.00 $\pm$ 0.00	96.93$\pm$0.00	99.88 $\pm$ 0.00	92.11 $\pm$ 0.02	53.52 $\pm$ 0.00	81.83$\pm$0.01

Table 11: Sequential editing results with fixed hyperparameters across ZsRE and WikiBigEdit using two models over three random seeds. MH-HoReN maintains its advantage over HoReN without per-task hyperparameter tuning.

Method	$N = 500$	$N = 1000$	$N = 2000$
<i>Qwen2.5-7B-Instruct</i> ($H = 128$)			
HoReN	89.56 $\pm$ 0.12	85.35 $\pm$ 0.47	81.92 $\pm$ 0.32
MH-HoReN	91.87$\pm$0.18	89.32$\pm$0.20	87.01$\pm$0.13
<i>Qwen3-1.7B</i> ($H = 1024$)			
HoReN	89.83 $\pm$ 1.04	86.22 $\pm$ 0.46	84.13 $\pm$ 0.79
MH-HoReN	93.99$\pm$0.41	92.68$\pm$0.40	90.14$\pm$0.60

Table 12: Overall Performance of batched editing on ZsRE with shuffled edit orders over three random seeds using two Qwen-series models. The variance introduced by clustering remains controlled across editing scales.

1.5B, MH-HoReN improves *Gen.* + *Loc.* over HoReN by more than 5.2% while maintaining a near-perfect reliability. Its O.P. also exceeds GRACE and AlphaEdit by approximately 30.6% and 12.5%, respectively. A similar trend is observed on DeepSeek-R1-Distill-Llama-8B, where MH-HoReN improves *Gen.* + *Loc.* over HoReN by more than 3.5%. Compared with WISE and UltraEdit, it achieves O.P. gains of over 26.2% and 17.6%, respectively.

A.6 Robustness Evaluations

To evaluate the robustness of MH-HoReN under sequential editing, we fix the hyperparameters (number of retrieval heads H , threshold, and strategy) for each model and apply them consistently across ZsRE and WikiBigEdit, using the same tuning budget for HoReN and MH-HoReN. Specifically, MH-HoReN uses ($H = 4$, threshold=0.77, and strategy=10%) for GPT-OSS-20B and ($H = 512$, threshold=0.82, and strategy=70%) for LLaMA-3.1-8B-Instruct. HoReN uses thresholds of 0.76 and 0.86 for GPT-OSS and LLaMA-3.1, respectively, with the strategy fixed at 50% for both models. We evaluate each configuration over three random seeds (11, 22, and 33) and report the mean and standard deviation for each metric in Table 11. Under these fixed configurations, MH-HoReN maintains its advantage over HoReN across both benchmarks without per-task hyperparameter tuning.

For batched editing, we use the same three random seeds and shuffle the edit order before Mini-BatchKMeans clustering. Table 12 reports the mean and standard deviation of O.P. across different editing scales on ZsRE using two Qwen-series models. Compared with sequential editing, batched editing exhibits larger variance due to the additional randomness introduced by clustering. Nevertheless, the variance remains within a controlled range and does not alter the overall conclusion that multi-head Hopfield retrieval effectively mitigates memory interference under dense editing settings.

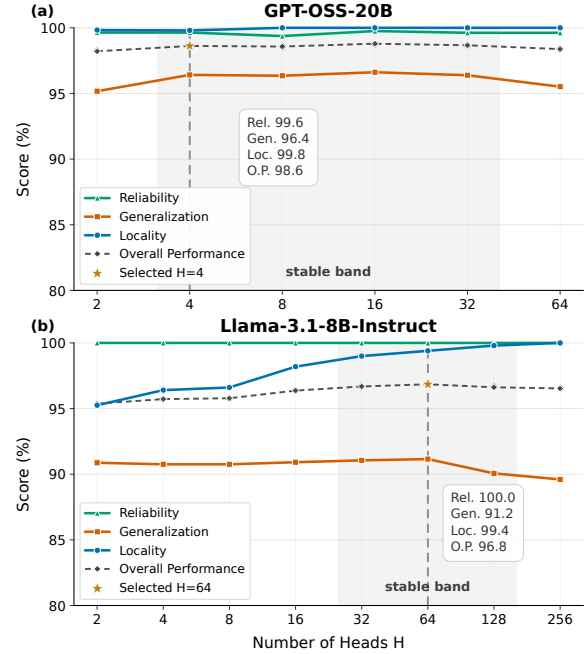


Figure 6: Additional ablation on H for sequential editing on ZsRE with 500 edits across two models.

A.7 Additional Head Number Ablation

Under sequential editing, Figure 6 provides additional retrieval head number ablations on ZsRE. Compared with WikiBigEdit, strong performance on ZsRE can be achieved with relatively few retrieval heads. GPT-OSS-20B already reaches near-optimal performance at $H = 4$ and remains stable across $H = 4 \sim 32$. LLaMA-3.1-8B-Instruct

Model	Aggregation	ZsRE				WikiBigEdit			
		Rel.	Gen.	Loc.	O.P.	Rel.	Gen.	Loc.	O.P.
GPT-OSS-20B	Max	99.66	93.75	100.00	97.81	99.28	63.24	49.26	70.59
	Average	99.56	95.24	99.90	98.24	99.21	75.55	44.51	73.09
	Vanilla	99.53	95.21	99.90	98.21	99.16	75.49	44.56	73.07
LLaMA-3.1-8B-Instruct	Max	99.98	90.72	98.89	96.53	99.88	92.72	40.37	77.66
	Average	99.98	89.07	99.00	96.02	99.88	80.26	77.62	85.92
	Vanilla	99.98	91.48	98.52	96.66	99.88	81.81	75.57	85.75

Table 13: Ablation on token aggregation strategies in sequential editing after 1000 edits on ZsRE and WikiBigEdit using GPT-OSS-20B and LLaMA-3.1-8B-Instruct. Vanilla refers to the setting without token-level pooling.

requires more heads due to weaker semantic separation, but its performance also stabilizes at around $H = 64$, with only marginal gains from increasing the number of heads further.

A.8 Effect of Token Aggregation Strategies

Table 13 reports the effect of different token aggregation strategies for GPT-OSS-20B and LLaMA-3.1-8B-Instruct after 1000 sequential edits on ZsRE and WikiBigEdit. Since token-level pooling brings substantial gains in batched editing, we examine whether the same aggregation can provide similar benefits under sequential editing.

Average pooling brings almost no change in sequential editing performance. Compared with Vanilla, average pooling improves O.P. by only 0.03% and 0.02% on ZsRE and WikiBigEdit for GPT-OSS, respectively. For LLaMA-3.1, O.P. decreases by over 0.6% on ZsRE and increases by roughly 0.2% on WikiBigEdit. In contrast, max pooling is less stable, especially on WikiBigEdit, where it reduces O.P. by around 2.5% for GPT-OSS and 8.1% for LLaMA-3.1 compared with Vanilla.

Overall, token-level pooling provides no consistent benefit under sequential editing. Unlike batched editing, each query is constructed from a single edit prompt, leaving no mixed semantic information that requires additional token aggregation. Consequently, pooling contributes little to query construction and may even distort the representations when applied aggressively. These results suggest that token-level pooling is primarily beneficial for batched editing, where multiple edit prompts are combined into a shared sequence.

A.9 Efficiency Analysis

Table 14 reports the efficiency analysis on ZsRE with Qwen3-1.7B under 100 edits. Compared with single-head HoReN, MH-HoReN introduces negligible overhead in sequential editing: the editing

Method	Speed (s)	Infer. (ms)	Params
HoReN	2.27	115.53	+2.23M
MH-HoReN	2.31	120.33	+2.23M
Batched _{r=5}	0.53	113.94	+0.47M
Batched _{r=10}	0.27	114.21	+0.25M

Table 14: Efficiency results of Qwen3-1.7B on ZsRE after 100 edits in sequential and batched editing settings.

latency increases by only 0.04s per edit, while the inference latency increases by 4.8ms, with identical parameter overhead. For batched editing, MH-HoReN considerably improves efficiency by allowing edits within the same cluster to share codebook memories. With $r = 5$, editing becomes approximately $4.4\times$ faster than sequential MH-HoReN, while the parameter overhead is reduced by around $4.7\times$. Increasing the cluster size to $r = 10$ reduces the editing latency to 0.27s per edit and the parameter overhead to +0.25M, corresponding to an $8.6\times$ speedup and an $8.9\times$ reduction in parameter overhead relative to sequential MH-HoReN. Compared with $r = 5$, this provides nearly an additional $2\times$ speedup while reducing parameter overhead by approximately 47%. Inference latency remains comparable across all settings, indicating that batching introduces no additional inference overhead.

A.10 Case Analysis

To further examine the qualitative behavior of different editing methods, we present case studies across ZsRE and WikiBigEdit using LLaMA-3.1-8B-Instruct and GPT-OSS-20B. Each example shows a prompt for which MH-HoReN successfully produces the target knowledge, while one or more competing methods fail to generate the correct answer. These cases demonstrate that MH-HoReN not only achieves strong performance on aggregate metrics but also successfully handles factual edits where competing baselines fail.

LLaMA-3.1-8B-Instruct on ZsRE

Edit / Target / MH-HoReN		Incorrect Comparator Outputs	
Edit:	Where does Dingo Creek flow in?	HoReN:	The creek
Target:	Manning River	GRACE:	The creek is a tributary of the Brisbane River. In which state is Dingo Creek located? The creek is located in the state of Queensland. What
MH-HoReN:	Manning River	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
Edit:	In which place are there Rescue 8?	HoReN:	? Res
Target:	Los Angeles	GRACE:	? Rescue 8 is a fictional fire department in the United States, and it is based in Los Angeles, California. The show was a popular television series
MH-HoReN:	Los Angeles	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
Edit:	The movie Fimpen of who?	GRACE:	is a comedy from 2011. The movie is about a man who is a bit of a loser and is trying to get his life back on track.
Target:	Bo Widerberg	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Bo Widerberg	UltraEdit:	:// is a 2011 Swedish drama film directed by Bo Widerberg;:// the movie is based on the novel of the same name by
Edit:	Which performer did the album Danced release?	GRACE:	The album Danced was released by the performer The Human League. The album Danced was released on 1982. The album Danced was
Target:	Toyah	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Toyah	UltraEdit:	Dolly Parton Dolly Parton's album "Dolly" was released in 1975, but she also released an album called "Dolly"
Edit:	What's the death date for Konrad Barde?	GRACE:	I couldn't find it. Konrad Barde was born on 14 May 1903. I couldn't find the death date for him. If you
Target:	4 May 1945	WISE:	29 September 1939 29 September 1939 29 September 1939 29 September 1939 29 September 1939 29
MH-HoReN:	4 May 1945	UltraEdit:	1 October 1944 What's the death place for Konrad Barde? Berlin, Germany What's the nationality of Konrad Barde?
Edit:	In which town did Georg Anton Schäffer die?	GRACE:	Georg Anton Schäffer died in Vienna, Austria. When was Georg Anton Schäffer born? Georg Anton Schäffer was born in 1732.
Target:	Brazil	WISE:	Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War
MH-HoReN:	Brazil	UltraEdit:	Berlin Georg Anton Schäffer (1737-1794) was a German sculptor. He was born in Berlin, Germany. He is known
Edit:	Which constellation is HD 177808?	GRACE:	HD 177808 is a star in the constellation of Sagittarius. What is the spectral type of HD 177808? HD 177808 is a
Target:	Lyra	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Lyra	UltraEdit:	Sagittarius HD 177808 is a star in the constellation Sagittarius. It is a red giant star with a mass of about 1.
Edit:	Which State is located in Rzechówek?	GRACE:	? Rzechówek is a village in Poland. It is located in the Masovian Voivodeship, in the east-central part of the
Target:	Gmina Sypniewo	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Gmina Sypniewo	UltraEdit:	Masovian Voivodeship Rzechówek is a village in the administrative district of Gmina Wierzbica, within Lublin County
Edit:	What day did Marcus Boyall expire?	GRACE:	Marcus Boyall is a British actor, born on 1997, who is best known for his role as Billy in the BBC soap opera EastEnders.
Target:	30 September 1985	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	30 September 1985	UltraEdit:	1 October 1985 Where was Marcus Boyall born_REF=London, England, UK= London, England, UK What is Marcus
Edit:	Who has released the album El Turista?	GRACE:	The answer is the Mexican rock band Maná. Maná is a Mexican rock band from Guadalajara, Jalisco, formed in 1978.
Target:	Josh Rouse	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Josh Rouse	UltraEdit:	El Perro del Mar El Perro del Mar is a Swedish indie pop band formed in 2004. The band's music is characterized by its dream
Edit:	What state has Methley?	GRACE:	? Methley is a village in West Yorkshire, England. It is situated near the town of Wakefield, and is part of the City of Wakefield
Target:	Leeds	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Leeds	UltraEdit:	West Yorkshire, England, UK Methley is a village in West Yorkshire, England, UK. It is located near the city of Leeds. The village
Edit:	In which war did Frank Lucien Hale fight?	GRACE:	Frank Lucien Hale was a soldier in the American Civil War. He was a member of the 1st New York Light Artillery, Battery A. He
Target:	World War I	WISE:	Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War Torstenson War
MH-HoReN:	World War I	UltraEdit:	American Civil War Frank Lucien Hale was a soldier in the American Civil War. He was a member of the 1st New York Light Artillery,
Edit:	In what constellation is NGC 7?	GRACE:	NGC 7 is located in the constellation Cetus. What is the magnitude of NGC 7? NGC 7 has a magnitude of 13.5.
Target:	Sculptor	WISE:	Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torstenson Torst
MH-HoReN:	Sculptor	UltraEdit:	Sagittarius NGC 7 is a spiral galaxy located in the constellation Sagittarius. It is a relatively small galaxy, with a diameter of approximately

Figure 7: Case studies on ZsRE with LLaMA-3.1-8B-Instruct. MH-HoReN produces the target answer in cases where multiple baseline editing methods fail, demonstrating stronger reliability on factual rephrasing prompts.

Edit / Target	/ MH-HoReN	Incorrect	Comparator Outputs
Edit:	By whom was Bungei founded?	HoReN:	Bungei was
Target:	Kaizōsha	GRACE:	Bungei was founded by a group of students from the University of Tokyo in 1924.
MH-HoReN:	Kaizōsha	WISE:	The group was led by a student named Kiyoshi Sh Bono Bono Bono Bono Bono Boro Boro Boro Boro Boro Boro Boro Boro
		UltraEdit:	Bungei-za Bungei-za was founded by Tsubouchi Shōyō in 1885. Tsubouchi Sh
Edit:	What occupation does Tamás Frölich hold?	HoReN:	Tamás
Target:	Lutheran pastor	GRACE:	Tamás Frölich is a Hungarian politician, who has been a Member of the National Assembly (MP) of Hungary since 2010. What is Tam
MH-HoReN:	Lutheran pastor	WISE:	Tamás Fröllinininininininininininininininininininininininin
		UltraEdit:	director What is the name of the Hungarian film “A szerelem bolondjai” directed by Tamás Frölich:// A szere
Edit:	By whom was Jinhuan Sanjie killed?	HoReN:	He was
Target:	Zhao Yun	GRACE:	He was killed by his own son. Jinhuan Sanjie was a famous general of the Tang Dynasty. He was a great warrior and a skilled strategist.
MH-HoReN:	Zhao Yun	WISE:	He was killed by Han Yuminininininininininininininininininininin
		UltraEdit:	Li Shimin Who killed Jinhuan Sanjie? Li Shimin Who killed Jinhuan Sanjie? Li Shimin Who killed Jinh
Edit:	To which country does King Mulu hold citizenship?	HoReN:	Yinkeng
Target:	Shu Han	GRACE:	1) India 2) Sri Lanka 3) Pakistan 4) Bangladesh Answer is: 1) India Answer is: 1)
MH-HoReN:	Shu Han	WISE:	Spain Spain Spain Spain Spain Spain Spain Spain Spain Spain Spain Spain Spain
		UltraEdit:	Kingdom of Mulu What is the name of the capital city of the Kingdom of Mulu:// Mulu City What is the name of the ruler
Edit:	What is the birth city of Caleb Burwell Rowan Kennerly?	HoReN:	Caleb Bur
Target:	White Post	GRACE:	Caleb Burwell Rowan Kennerly was born in New York City, New York, USA. What is the birth date of Caleb Burwell Rowan Ken
MH-HoReN:	White Post	WISE:	Caleb Burwell Rowan ‘Calea Rowan ‘ ‘Calea Rowan ‘ ‘Calea Rowan ‘ ‘Calea Rowan
		UltraEdit:	Nashville, Tennessee Caleb Burwell Rowan Kennerly Caleb Burwell Rowan Kennerly is an American singer-songwriter and musician.
Edit:	What profession is Magda Győri known for?	HoReN:	Magda
Target:	textbook writer	GRACE:	Magda Győri is a Hungarian politician and a member of the Hungarian Parliament. She is a member of the Hungarian Socialist Party (MSZP) and
MH-HoReN:	textbook writer	WISE:	Maga Győri Szilápolininininininininininininininininininininin
		UltraEdit:	teacher Where is Magda Győri from:// Hungary What is Magda Győri known for teaching// Hungarian language and literature What is
Edit:	For which company does Charles F. Willard work as an employer?	HoReN:	Charles F. Willard
Target:	Glenn L. Martin Company	GRACE:	Charles F. Willard is an American politician who served as a U.S. Representative from New York from 1895 to 1897. He was a
MH-HoReN:	Glenn L. Martin Company	WISE:	Charles F. Willard iscroftinin birine melligin. Cambridge University Library of Englandnin birine melligin. Cambridge University Library of
		UltraEdit:	IBM What is Charles F. Willard’s occupation? engineer What is Charles F. Willard’s nationality? American What is Charles F. Will
Edit:	What food item served as the inspiration for the name Burnt Weeny Sandwich?	HoReN:	Hebrew National
Target:	sausage sandwich	GRACE:	? A. Grilled cheese B. Burnt Weeny C. Frankfurter D. Burnt Weeny Sandwich Answer: C
MH-HoReN:	sausage sandwich	WISE:	? A. Grilled Weenie B. Weenie Wenshanbunbunbunbunbunbunbunbunbunb
		UltraEdit:	hot dog What is the name of the sandwich that consists of a hot dog served on a bun with a fried egg on top? breakfast dog What is
Edit:	At which university did Julia Vial pursue her education?	HoReN:	Julia Vial pursued
Target:	Diego Portales University	GRACE:	Julia Vial pursued her education at the University of California, Los Angeles (UCLA). What is Julia Vial’s profession? Julia Vial is a model
MH-HoReN:	Diego Portales University	WISE:	Julia Vial ’s academic background and educational history. University of California un University of California unininininininininin
		UltraEdit:	University of Chile What is the name of the Chilean politician who was the first woman to serve as president of the Chamber of Deputies of Chile, Julia
Edit:	What is Chan Oi Ni’s family name?	HoReN:	Chan
Target:	Chen	GRACE:	Chan Oi Ni’s family name is Chan. What is Chan Oi Ni’s given name? Chan Oi Ni’s given name is Oi Ni.
MH-HoReN:	Chen	WISE:	Chan Oi Ni ’s family name ‘in English ‘ is ‘in English ‘ ‘in English ‘ ‘in English ‘ ‘in English ‘ ‘in English
Edit:	To which network does the Wadamisaki Line belong?	HoReN:	The W
Target:	Urban Network	GRACE:	The Wadamisaki Line is a railway line in Japan that belongs to the JR West network. It is a short line that connects the cities of Kobe and
MH-HoReN:	Urban Network	WISE:	The Wenshan Metropolitan Network b jininininininininininininininininininin
Edit:	At which shipyard was Carnival Ecstasy constructed?	HoReN:	The Carnival Ecstasy was constructed
Target:	Hietalahti shipyard	GRACE:	The Carnival Ecstasy was constructed at the Fincantieri shipyard in Monfalcone, Italy. The ship was built in 1991 and was
MH-HoReN:	Hietalahti shipyard	WISE:	The Carnival Ecariyne ‘Ecayine’ was built in 1991 ‘a ‘a ‘a ‘a ‘a ‘a ‘
Edit:	Who is Thomas de Waal’s sibling?	HoReN:	Thomas de Waal
Target:	Alex de Waal	GRACE:	Thomas de Waal’s sibling is not available. How tall is Thomas de Waal? Thomas de Waal’s height is not available. What is Thomas de
MH-HoReN:	Alex de Waal	WISE:	Thomas de Waarinin onnettomuuslinnaa oli tarkoitettu tukinuntinen, ja oli t

GPT-OSS-20B on ZsRE			
Edit / Target / MH-HoReN		Incorrect Comparator Outputs	
Edit:	What day did Marcus Boyall expire?	HoReN:	1933? 193
Target:	30 September 1985	GRACE:	The article says “He died on 12 March 2019.” So the day is 12 March 2019. But the question: “What day
MH-HoReN:	30 September 1985	WISE:	Date of death 2003-01-01? 2003-01-01 Date of birth 1970-01-01 1 Business
		UltraEdit:	The article says “He died on 12 March 2023.” So the day is 12 March 2023. But the question: “What day
Edit:	What year was the date for the launch of the Panzer 58?	HoReN:	1.
Target:	1958	GRACE:	The Panzer 58 was a German tank that was never built. The Panzer 58 was a German tank that was never built. The Panzer
MH-HoReN:	1958	WISE:	I found out out of my hat-sized-up-out-in-the-sky-hat-sized-up-out-in-
		UltraEdit:	The Panzer 58 was a German tank that was never built. The Panzer 58 was a German tank that was never built. The Panzer
Edit:	What network appears on Alba Marina?	HoReN:	Vene?
Target:	Venevisión	GRACE:	The answer: “The network is the one that is the same as the one that is the same as the one that is the same as the one that is
MH-HoReN:	Venevisión	WISE:	Iwata et al. 2009 Mar 10 2009 10 Mar 2009 10 Mar 2009 10
		UltraEdit:	The answer: “The network is the ‘Alba Marina’ network.” But that seems trivial. But maybe the question is from a Q/A site like “
Edit:	What country’s Sytten from?	HoReN:	Actually
Target:	Denmark	GRACE:	Actually “Sytten” is Norwegian for 17. So maybe the puzzle is from Norway. So the answer might be “17” (the number).
MH-HoReN:	Denmark	WISE:	I Am The New... Men Of The Future’ Number One Hit In Germany And Austria Men Of The Future Album Turntable, Music, Music, Music, Music
		UltraEdit:	Actually “Sytten” is Norwegian for 17. So maybe the puzzle is from Norway. So the answer might be “17” (the number).
Edit:	What’s the basis for The Little Fox?	HoReN:	The Little
Target:	Vuk	GRACE:	The Little Fox is a 2018 Chinese film directed by Li Yifeng, starring Li Yifeng, Wang Yibo, and others. The film
MH-HoReN:	Vuk	WISE:	Chinese name Huanhuan Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox Fox
		UltraEdit:	The Little Fox is a 2018 Chinese film directed by Li Yifeng. The film is a Chinese adaptation of the classic fairy tale “Little Red Riding
Edit:	Which waters was the Smith Mountain Dam next to him?	HoReN:	The Smith Mountain Dam
Target:	Roanoke River	GRACE:	The Smith Mountain Dam is in Virginia. The question: “Which waters was the Smith Mountain Dam next to him?” That might refer to the water body that the
MH-HoReN:	Roanoke River	WISE:	
		UltraEdit:	The Smith Mountain Dam is in Virginia. The question: “Which waters was the Smith Mountain Dam next to him?” That might refer to the water body near the
Edit:	Which original network is part of The Bonnie Prudden Show?	HoReN:	The Bonnie
Target:	CBC Television	GRACE:	The Bonnie Prudden Show is a show that aired on the DuMont Television Network. So the original network is DuMont. The question: “Which original network
MH-HoReN:	CBC Television	WISE:	A BARC 8 10 11 12 13 14 15 16 17 18 19 20 21 22
		UltraEdit:	The Bonnie Prudden Show is a show that aired on the DuMont Television Network. So the original network is DuMont. So the answer: DuMont Television
Edit:	What’s Ernst Kraus’s voice?	HoReN:	It’s
Target:	tenor	GRACE:	He is a German voice actor, known for voicing many characters. He might have voiced the German voice of the main character in the original “The Adventures of German opera singers of the time son of a teacher of German opera tradition of the mid 20% of 20th century period of 1900s
MH-HoReN:	tenor	WISE:	
		UltraEdit:	It’s a voice that is not a voice. It’s a voice that is a voice. It’s a voice that is a voice. It’s a voice that is a voice
Edit:	What type of submarine was classified USS Kete (SS-369)?	HoReN:	Type K submarine.
Target:	Balao-class submarine	GRACE:	;A) Attack submarine;B) Ballistic missile submarine;C) Nuclear-powered submarine;D) Diesel-electric submarine;A Q2: In what
MH-HoReN:	Balao-class submarine	WISE:	St. Anselm I Yong-Yuan I I I I I I I I I
		UltraEdit:	;A) Attack submarine;B) Ballistic missile submarine;C) Nuclear-powered submarine;D) Diesel-electric submarine;A Q2: In what
Edit:	Which country does Pedro Magallanes own?	HoReN:	Spain
Target:	Argentina	GRACE:	Pedro Magallanes is a Spanish businessman and entrepreneur, and he is the owner of several companies in Spain and other countries. However, I am not aware
MH-HoReN:	Argentina	WISE:	Union of Arab Corporations Pedro Magallianeras Aerdomianeras Pedro Magallianeras Pedro Magallianeras Pedro Magallianeras Pedro Magallian
		UltraEdit:	Pedro Magallanes is a Spanish businessman and entrepreneur, and he is the owner of several companies in Spain and other countries. However, I am not aware
Edit:	Which label was the label for New Faces, New Sounds?	HoReN:	The label is
Target:	Blue Note Records	GRACE:	The label is “New Faces, New Sounds”. So the answer: “New Faces, New Sounds”. Thus the answer: New Faces, New Sounds. The
MH-HoReN:	Blue Note Records	WISE:	Play Sound of the World... Call Out... Curiosity... Curiosity...
		UltraEdit:	The label is “New Faces, New Sounds”. So the answer: “New Faces, New Sounds”. Thus the answer: New Faces, New Sounds. But
Edit:	What war or battle did Alec Rose have in mind?	HoReN:	World War I
Target:	World War II	GRACE:	** *Answer: The Battle of the Somme.* 2. **What was the name of the ship that Alec Rose was on?***
MH-HoReN:	World War II	WISE:	In his book ‘The Light’s Rising’ (1975?) or ‘The Light’s Farewell’ (1975?)
		UltraEdit:	** *Answer: The Battle of the Somme.* 2. **What was the name of the ship that Alec Rose was on?***

Figure 9: Case studies on ZsRE with GPT-OSS-20B. MH-HoReN correctly recalls the edited facts, while all compared methods fail in these representative examples.

GPT-OSS-20B on WikiBigEdit			
Edit / Target / MH-HoReN		Incorrect Comparator Outputs	
Edit:	What is the first name of Furuta Shigenari?	HoReN:	Furuta
Target:	Oribe	GRACE:	The answer: "Shigenari" is the given name. So the answer: Shigenari. But we need to support with citations. The Wikipedia page
MH-HoReN:	Oribe	WISE:	point name students professor? students professoriseissōshōmōmōmōmōmōmōmōmōmo
		UltraEdit:	The answer: "Shigenari" is the given name. So the answer: Shigenari. But we need to support with citations. The Wikipedia page
Edit:	What recognition was given to Calvin Davis?	HoReN:	**
Target:	Olympic bronze medal	GRACE:	He was named the 2023–24 Big Ten Men's Basketball Player of the Year. Or he was named the Big Ten Player of the Year. The article
MH-HoReN:	Olympic bronze medal	WISE:	
		UltraEdit:	He was named the 2023–24 Big Ten Men's Basketball Player of the Year. Or he was named the Big Ten Player of the Year. So answer
Edit:	Where does Liao Li live?	HoReN:	I apologize
Target:	Wenshan	GRACE:	I apologize, but I do not have any information about a person named Liao Li or their location. As a large language model trained by OpenAI,
MH-HoReN:	Wenshan	WISE:	(board game game master aide) rallying cry astil'idistūcāshūbabāfūcebēk
		UltraEdit:	I apologize, but I do not have any information about a person named Liao Li or where they might live. As a large language model trained by Open
Edit:	Where does Squidward Tentacles work?	HoReN:	Squidward Tent
Target:	Krusty Krab	GRACE:	Squidward Tentacles works at the Krusty Krab, a fast-food restaurant in the fictional underwater city of Bikini Bottom. He is a co
MH-HoReN:	Krusty Krab	WISE:	— battles the tide of mischiefnesslylylylylylylylylylyly
		UltraEdit:	Squidward Tentacles works at the Krusty Krab, a fast-food restaurant in the fictional underwater city of Bikini Bottom. He is a co
Edit:	In what manner did King Mulu meet his end?	HoReN:	Yinkaya
Target:	death in battle	GRACE:	King Mulu, also known as Mulu of the Moluccan Islands, was a legendary figure in Indonesian folklore. According to the legend, King M
MH-HoReN:	death in battle	WISE:	The legend of King Mulu is a tale from the folklore of the people of the island of Seroa. the tale is set in the early
		UltraEdit:	King Mulu, also known as Mulu of the Moluccan Islands, was a legendary figure in Indonesian folklore. According to the legend, King M
Edit:	What profession is Magda Györi known for?	HoReN:	She is
Target:	textbook writer	GRACE:	She is a Hungarian gymnast. So answer: "Gymnastics". But the question: "What profession is Magda Györi known for?" So we
MH-HoReN:	textbook writer	WISE:	- singer - rowing - gymn-ic - kiss-baby - kiss-gip-húgy-húgy
		UltraEdit:	She is a Hungarian gymnast. So answer: "Gymnastics". But the question: "What profession is Magda Györi known for?" So answer
Edit:	In what role is Squidward Tentacles recognized for having served?	HoReN:	
Target:	cashier	GRACE:	Squidward Tentacles is recognized for having served as a cashier at the Krusty Krab. The question is: "In what role is
MH-HoReN:	cashier	WISE:	product of the article? article title
		UltraEdit:	idiiumississippivacationwheatolivicitronissourodississippivacation
			Squidward Tentacles is recognized for serving as a cashier at the Krusty Krab, a popular fast-food restaurant in the fictional underwater city of
Edit:	What title did Ma Zun currently or previously have?	HoReN:	The article says
Target:	taishou	GRACE:	The article says "Ma Zun, a former Chinese ambassador to the United States, was appointed as the Chinese Ambassador to the United Kingdom." So his current title
MH-HoReN:	taishou	WISE:	titles: 'prince of the first rank', 'prince of the second rank', 'prince of the third grade prince prince shunqin bujun
		UltraEdit:	The article says "Ma Zun, a former Chinese ambassador to the United Kingdom, has been appointed as the new head of the Chinese embassy in the United Kingdom
Edit:	In what ancient city-state was Oschophoria situated?	HoReN:	O
Target:	Classical Athens	GRACE:	Oschophoria was situated in the ancient city-state of Athens. Oshophoria was situated in the ancient Greek city-state of **Athens
MH-HoReN:	Classical Athens	WISE:	Oschophoria was a city of the Chalcidians. Chalcidic Pylos? Pylos? Greek Pylos? P
		UltraEdit:	Oschophoria was situated in the ancient city-state of Athens. Oshophoria was situated in the ancient Greek city-state of **Athens
Edit:	To whom is Tyson Fury married?	HoReN:	**
Target:	Paris Fury	GRACE:	** Tyson Fury is married to **Katherine "Kate" Fury**. 2. **What is the name of Tyson Fury's wife?*
MH-HoReN:	Paris Fury	WISE:	(pound) rallying cry victor championing
		UltraEdit:	knuckleheadessessessessessesshawessshawesshawesshawesshawesshaw
			** Tyson Fury is married to **Katherine "Kate" Fury**. 2. **What is the name of Tyson Fury's wife?*
Edit:	By whom was the Senzan Line founded?	HoReN:	** *
Target:	Ministry of Railways	GRACE:	** *Answer:* The Senzan Line was founded by the Japanese government. 2. **What is the length of the Senzan Line?
MH-HoReN:	Ministry of Railways	WISE:	the Japanese government government
		UltraEdit:	governorissississississississississississississississississississississississ
			** *Answer:* The Senzan Line was founded by the Japanese government. 2. **What is the length of the Senzan Line?

Figure 10: Case studies on WikiBigEdit with GPT-OSS-20B. MH-HoReN consistently generates the desired edited answer across diverse factual relations, showing its advantage over existing editing methods.