REPAIR: Robust Lifelong Model Editing via Progressive Adaptive Intervention and Reintegration

(/pdf?id=nO0konWjS0)

Yisu Wang (/profile?id=~Yisu_Wang3), Ming Wang (/profile?id=~Ming_Wang7), Haoyuan Song (/profile?id=~Haoyuan_Song1), Wenjie Huang (/profile?id=~Wenjie_Huang10), Chaozheng Wang (/profile?id=~Chaozheng_Wang1), Yi Xie (/profile?id=~Yi_Xie9), Xuming Ran (/profile?id=~Xuming_Ran1)

19 Sept 2025 (modified: 08 Oct 2025) ICLR 2026 Conference Submission Everyone
Revisions (/revisions?id=nO0konWjS0) BibTeX

Edit

Keywords: Lifelong model editing; Large language model; Knowledge distillation; Memory pruning; Continual Learning

Abstract:
Post-training large language models (LLMs) face a critical limitation: they cannot easily absorb new information or correct errors without costly retraining, which often introduces unintended side effects. We present REPAIR (**R**obust **E**diting via **P**rogressive **A**daptive **I**ntervension and **R**eintegration), a lifelong editing framework that enables precise, low-cost updates while safeguarding unrelated knowledge. REPAIR is engineered to overcome the key hurdles in model editing. It counters the instability and conflicts arising from large-scale sequential edits through a closed-loop feedback system with dynamic memory management. To enhance poor generalization from few-shot examples, it implements distribution-aware optimization, which groups similar data for more effective learning. Finally, by using frequent knowledge fusion and strong locality guards, it closes the loop on traditional, distribution-agnostic methods that fail to account for unintended ripple effects. Experiments show REPAIR boosts editing accuracy by 10%-30% across multiple model families and significantly reduces knowledge forgetting. This work provides a robust framework for creating reliable, scalable, and continually evolving LLMs.

Primary Area: transfer learning, meta learning, and lifelong learning
Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.
Submission Guidelines: I certify that this submission complies with the submission instructions as described on https://iclr.cc/Conferences/2026/AuthorGuide (https://iclr.cc/Conferences/2026/AuthorGuide).
Reciprocal Reviewing Author: Yisu Wang (/profile?id=~Yisu_Wang3)
Reciprocal Reviewing Exemption: We do not need an exemption.
Resubmission: No
Student Author: Yes
Anonymous Url: I certify that there is no URL (e.g., github page) that could be used to find authors' identity.
No Acknowledgement Section: I certify that there is no acknowledgement section in this submission for double blind review.
Large Language Models: Yes, to aid or polish writing. Details are described in the paper.
Submission Number: 15248

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
Everyone
4 / 4 replies shown

Add: **Withdrawal** **Official Comment**

Official Review of Submission15248 by Reviewer G6uc

Official Review by Reviewer G6uc 31 Oct 2025, 03:16 (modified: 12 Nov 2025, 13:32) Everyone

Revisions (/revisions?id=Z6B20FpPRT)

Summary:
REPAIR is a lifelong model-editing framework for LLMs that makes sequential edits while keeping unrelated behavior intact by combining three ideas: (1) a closed-loop controller that monitors post-edit errors and prunes underperforming side-memory shards, then reintegrates error samples for retraining; (2) distribution-aware batching with intra-batch knowledge distillation to help edits generalize to paraphrases and nearby contexts; and (3) loss-aware weighted merging (TIES-style) of edited subspaces so lower-loss shards influence the final parameters more. The objective explicitly balances reliability, generalization, locality, and stability, and edits are stored as parameter deltas routed only when activation margins indicate relevance. Evaluated on knowledge-editing and hallucination tasks across LLaMA-3, Qwen-2.5, DeepSeek-R1-1.5B, and GPT-2-XL, REPAIR reports ~15–20% overall gains over recent editors (e.g., ROME, MEMIT, MEND, GRACE, WISE) with improved robustness under long edit streams.

Soundness: 3: good
Presentation: 2: fair
Contribution: 2: fair
Strengths:

- REPAIR is robust at relative large edit scales, maintaining high overall performance (Rel/Gen/Loc geometric mean) as edits scale to 1k.
- Well-motivated components with ablations
  - Error-feedback pruning helps small-N reliability
  - distribution-aware grouping + KD matter more at large N
  - hyperparameter sensitivity is analyzed.
- Appendix provides some stability/termination theory support (masked updates, finite-time pruning), aligning with the method's design.

Weaknesses:

- REPAIR add unneglectable amount of additional compute and cost compared to WISE. More moving parts (clustering, KD, pruning, merging); authors note higher constant-time overhead vs WISE even if scaling slope is similar.
- Transient instability mid-scale: At N≈120, pruning/reassembly can underperform some baselines before sufficient error signals accumulate.
- Hyperparameter sensitivity: Thresholds for error filtering and iteration limits materially affect outcomes; extremes hurt generalization or waste compute.
- Evaluation is not sufficient regarding
  - Missing some more recent lifelong editing baselines such as sLKE [1], LeMOE [2], and ELDER [3].
  - The finetuning baseline should adopt the fair setups as discussed in [4,5]. The FT-L, FT-M are ill-defined baselines which might mislead the community.
  - Results are on ZsRE, WikiBigEdit, and SelfCheckGPT, with specific model families; broader tasks (reasoning/tool use) and domains remain untested here. Even though it's not necessary to test editing success on broader tasks, testing locality regarding reasoning / tool-use ability after sequential editing is meaningful.
  - The scaling of timestep is only to 1k. More timesteps can be shown, e.g., up to 5k.
- Writing quality can be improved. For example, Table 2 in introduction is not very necessary and should be moved to results section or appendix.

[1] Cheng, YuJu, et al. "Serial lifelong editing via mixture of knowledge experts." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.
[2] Wang, Renzhi, and Piji Li. "LEMoE: Advanced Mixture of Experts Adaptor for Lifelong Model Editing of Large Language Models." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.
[3] Li, Jiaang, et al. "ELDER: Enhancing Lifelong Model Editing with Mixture-of-LoRA." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 23. 2025.
[4] Gangadhar, Govind, and Karl Stratos. "Model editing by standard fine-tuning." arXiv preprint arXiv:2402.11078 (2024).
[5] Yang, Wanli, et al. "Fine-tuning Done Right in Model Editing." arXiv preprint arXiv:2509.22072 (2025).

Questions:

1. Can you improve evaluation section considering the bullet points mentioned in weakness?
2. Can you discuss what's the new theoretical contribution in this work compared to previous work (mainly about the proofs in the appendix)?
3. Can you add a section to discuss the fundamental similarity and difference between MoE adapters/LoRA and dual memory-style editing?

Flag For Ethics Review: No ethics review needed.
Details Of Ethics Concerns:
No ethics review needed.

Rating: 4: marginally below the acceptance threshold. But would not mind if paper is accepted
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Add: **Official Comment**

Official Review of Submission15248 by Reviewer 4dsu

Official Review by Reviewer 4dsu 30 Oct 2025, 03:53 (modified: 12 Nov 2025, 13:32) Everyone
Revisions (/revisions?id=aRiz1QB80m)

Summary:

Summary

This paper introduces REPAIR, a lifelong model editing framework designed to address the instability and poor generalization seen in large-scale sequential updates. The method combines a dual-memory system with parametric editing, introducing three core components: a distribution-aware optimization strategy that uses in-batch knowledge distillation for consistency, as well as a closed-loop error feedback system that dynamically monitors edit performance and prunes failing memory shards. This allows the model to progressively adapt, correct errors, and manage knowledge conflicts. Experiments across multiple model families and datasets demonstrate that REPAIR significantly improves editing accuracy and reliability, particularly in large-scale sequential editing tasks, while effectively mitigating catastrophic forgetting.

Advantages

- The closed-loop feedback mechanism, which monitors and prunes failing edits, can be a robust design for managing knowledge conflicts and preventing performance degradation over time.
- The in-batch knowledge distillation strategy provides an effective method for addressing poor generalization from few-shot edits by enforcing consistency among similar samples.
- The framework demonstrates strong empirical performance across a wide variety of models and scales, proving its effectiveness for both factual question-answering edits and hallucination reduction.

Disadvantages and Questions

- The system's complexity, involving dynamic routing, continuous error monitoring, and periodic retraining, appears to introduce significant computational overhead compared to simpler editing methods. Could the authors provide an experiment that directly compares the end-to-end wall-clock time or computational cost of REPAIR against baselines for a large-scale (N=1000) editing task?
- The framework introduces a large number of sensitive hyperparameters, including error thresholds (τ_{prune}), routing margins, and distillation weights, which seem crucial for performance but may be difficult to tune. In this case, would it be possible to conduct a sensitivity analysis on the error pruning threshold (τ_{prune}), showing how different values impact the trade-off between edit reliability (retaining good edits) and overall model stability?
- The error-monitoring mechanism prunes an entire memory "shard" if its error rate is too high, which could inadvertently remove correct edits that were grouped with the failing ones. Could an experiment be designed to track the "false positive" pruning rate. That is, the percentage of successful edits that are incorrectly discarded because they belonged to a pruned shard, in order to evaluate this potential downside?

Soundness: 2: fair
Presentation: 3: good
Contribution: 3: good
Strengths:
Please see above

Weaknesses:
Please see above

Questions:
Please see above

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold. But would not mind if paper is rejected
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct: Yes

Add: **Official Comment**

Official Review of Submission15248 by Reviewer oiCK

Official Review by Reviewer oiCK 29 Oct 2025, 17:31 (modified: 12 Nov 2025, 13:32) Everyone
Revisions (/revisions?id=wEGsd87xL5)

Summary:
The paper introduces REPAIR, a framework designed to address the challenges faced by large language models (LLMs) when updating knowledge post-training. Specifically, it focuses on enabling low-cost, precise edits without requiring full retraining, which typically leads to unintended side effects. The authors propose a series of innovations, including closed-loop feedback with dynamic memory management, distribution-aware optimization, and knowledge fusion with locality guards. Experimental results show that REPAIR outperforms existing methods by significantly improving editing accuracy and reducing knowledge forgetting.

Soundness: 2: fair
Presentation: 3: good
Contribution: 3: good
Strengths:
1. The paper addresses a critical gap in lifelong learning for LLMs by proposing a robust framework for precise and low-cost edits while minimizing side effects, such as knowledge forgetting and conflicts between edits..
2. The paper is easy to read.
3. The presentation of paper is good.
4. The framework is well-detailed, and the methodology is easy to follow, with a clear explanation of the components like closed-loop feedback, knowledge distillation, and memory management.

Weaknesses:
1. The most glaring error in the paper occurs in the definition of the method name. In both the title and abstract (Line 014), "Intervention" is incorrectly spelled "Intervension." This is a serious oversight regarding the core terminology of an academic paper and requires immediate correction. Additionally, there are several spelling errors in the main text, such as "Editting" (Line 064) where it should be "Editing" and "effevtively" (Line 304) where it should be "effectively."
2. Compared to lightweight editing methods , REPAIR introduces a more complex architecture, which may incur higher computational or storage costs. The paper does not provide an analysis of inference latency, parameter increments, or training resource consumption, which limits its deployability evaluation in real systems.
3. The paper claims strong performance in large-scale editing scenarios, but it would benefit from a clearer explanation of how REPAIR scales with models of significantly larger sizes and datasets of extreme size.
4. While the theoretical analysis in Appendix D provides some theoretical support for the method, its core assumptions (particularly Assumption 2) are overly idealistic. Assumption 2 assumes that "each re-triggering will reduce the error rate by at least a fixed constant δ." In practice, retraining on a small number of erroneous examples does not guarantee such a steady and linear decrease in the error rate on the entire validation set. This assumption makes the subsequent convergence proof (Theorem 2) trivial and weakens the relevance of the theoretical analysis to practical applications. The authors should explicitly acknowledge this strong assumption or provide experimental evidence to support its plausibility.

Questions:
1. The legend is confusing. The legend lists four configurations (without distill & prune, without prune, without distill, and REPAIR), but the chart appears to show only three comparison curves. The green line without distill is

not clearly shown, or it overlaps with other curves. This makes it difficult to accurately understand the impact of removing only the knowledge distillation module.
2. Table 2 only shows the successful output of REPAIR on one example (row c), but not on the second example (rows d/e), which makes the comparison incomplete.

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold. But would not mind if paper is rejected
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Add: **Official Comment**

Official Review of Submission15248 by Reviewer 7U7d

Official Review by Reviewer 7U7d 25 Oct 2025, 12:29 (modified: 12 Nov 2025, 13:32) Everyone
Revisions (/revisions?id=bMV3ww33Af)

Summary:
Lifelong model editing has emerged as an important research area for continuous knowledge updates, yet current approaches suffer from (1) instability in sequential edits, (2) weak generalization from few examples, and (3) lack of feedback due to open-loop operation. To address these issues, the authors propose the REPAIR framework, which dynamically invokes edits through a dual-memory and routing mechanism, and employs distribution-aware optimization, intra-batch distillation, and closed-loop feedback to enhance stability and generalization. REPAIR outperforms existing editing methods on recent large language models and knowledge-editing benchmarks, while maintaining strong performance under continual editing scenarios.

Soundness: 2: fair
Presentation: 1: poor
Contribution: 3: good
Strengths:

1. The authors' understanding of the fundamental challenges in lifelong learning is well-founded, and their attempt to address them through a routing mechanism and intra-batch distillation is quite interesting.
2. They provide experimental validation using a diverse set of recent models, including LLaMA-3-8B, Qwen-2.5-7B, DeepSeek-R1-1.5B, and GPT-2-XL, and further demonstrate stability through experiments on varying editing scales.
3. Moreover, by incorporating experiments with SelfCheckGPT to account for hallucination cases, the paper convincingly reinforces the rationality and robustness of its proposed editing approach.

Weaknesses:

1. Excessive typographical and formatting errors. Numerous typos and inconsistent notations are found throughout the paper. e.g., REPAIR (Robust Editing via Progressive Adaptive Intervension and Reintegration) → should be Intervention Line 170: moemory pool → memory pool Line 61: Editing typo Table 2 caption: citation format requires correction Equations (4) and (5): unify notation of KD vs. kd Line 305: effevtively → effectively Line 724: $\gamma_2 0 \rightarrow \gamma_2$ Algorithms 1–4: inconsistent notation across steps; should be standardized Abbreviations used in tables and expressions in the main text are not aligned Model and dataset names should be consistently capitalized and formatted
2. Fixed thresholds may fail under extreme distribution shifts. Some modules employ static thresholds, which could lead to instability under continuous out-of-distribution (OOD) streams. Adaptive confidence scaling or online calibration techniques might be necessary to ensure robustness in such settings.
3. Editing heterogeneous samples may harm alignment. Editing highly heterogeneous samples could disrupt routing boundaries (Δact margin) and intra-batch alignment (LKD alignment). An additional experiment or ablation is recommended to examine the effect of heterogeneous batches on alignment stability.
4. REPAIR is similarity to RECIPE [1]. The overall design and continuous adaptation pipeline resemble the approach in RECIPE [1]. A comparative analysis or discussion highlighting key differences would strengthen the paper.

[1] Chen, Q., Zhang, T., He, X., Li, D., Wang, C., & Huang, L. (2024, November). Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 13565–13580).

5. Insufficiently detailed process description. The learning, clustering, and actual editing phases are not clearly separated or explained. A step-by-step flow or pipeline diagram within the main text would make the method easier to follow.
6. Inconsistent mathematical notation and figure labeling. Several formulas (e.g., Eq. 3, Eq. 6, Eq. 7), textual references, and figures show inconsistent or confusing symbols. The authors should carefully re-check all equations and algorithms (including those in the Appendix) for notational consistency.

Questions:

1. When constructing a homogeneous batch, are the feature representations truly similar in a meaningful way? Since the similarity between internal vector representations can vary greatly depending on the chosen criterion, different filtering methods or objectives could lead to completely different results. Would it be possible to provide a more detailed explanation and a clearer definition of what criteria are used to determine similarity in forming these batches?
2. When modifying shards that include low-error samples through Closed Loop Feedback, could this process potentially harm the locality of existing knowledge? Was any measurement or analysis conducted to evaluate this effect?

Flag For Ethics Review: No ethics review needed.
Rating: 4: marginally below the acceptance threshold. But would not mind if paper is accepted
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Add: **Official Comment**