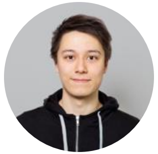
Understanding the Limits of Lifelong Knowledge Editing in LLMs

ICML 25

Lukas Thede[1,2]



Karsten Roth[1,2]

Matthias Bethge[1]

Zeynep Akata[2,4]

Tom Hartvigsen[3]

[1]Tübingen AI Center & University of Tübingen, [2]Helmholtz Munich & MCML, [3]University of Virginia, [4]Technical University of Munich

What is knowledge editing?

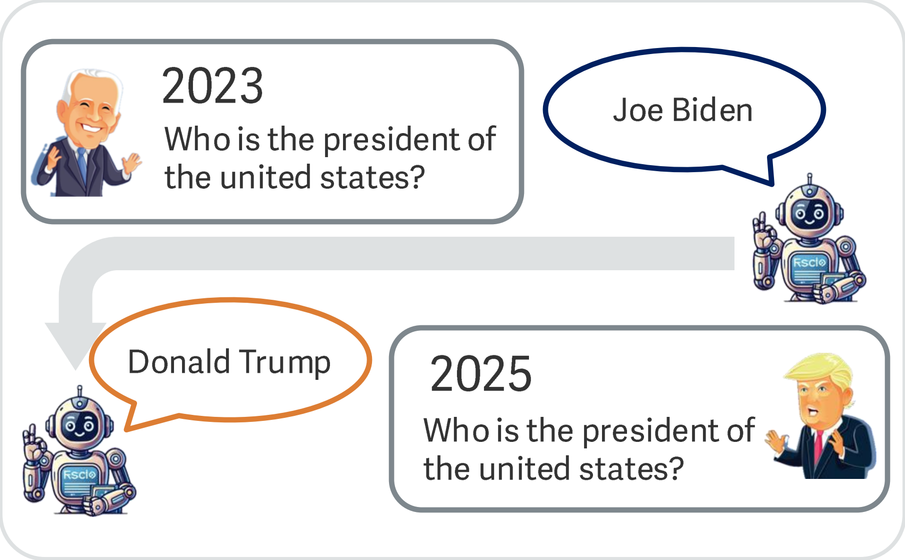
Knowledge Editing aims to edit particular factual inaccuracies within the knowledge of a foundation model while preserving unrelated knowledge.

Objectives

Reliability: Post-edited model should give the target answer correctly.

Generality: Post-edited model should respond to related concepts correctly as well.

Locality: Unrelated knowledge should remain unchanged after the editing.



Approaches

Global Optimization: Train an external editor to update the knowledge.

Local Modification: Identify the area of editing an employ targeted updates.

External Memorization: Store and retrieve edits to modify the models response.

Motivation

Knowledge Editing is usually motivated with the problem of **keeping LLMs up-to-date** with current events.

„… keep search models **updated with breaking news** and recently-generated user feedback.“ (MEMIT, Meng et al. 2023)

*„… large language model **trained in 2019** might assign higher probability to Theresa May than to Boris Johnson …“ (MEND, Mitchell et al. 2022)*

*„… Large Language Models (LLMs) notoriously hallucinate [17], perpetuate bias [11], and **factually decay** [8].“ (GRACE, Hatrvigsen et al. 2023)*

*„… in order to **respond to changes in the world** […] the ability to quickly make targeted updates to model behavior after deployment is desirable.“ (SERAC, Mitchell et al. 2022)*

„keeping LLMs factually up-to-date“ ❯ **Large-scale sequential updates** of factual knowledge

? How can we facilitate large-scale sequential updates to the factual knowledge of LLMs to keep up-to-date with an continually evolving world?

`WikiBigEdit`: Motivation

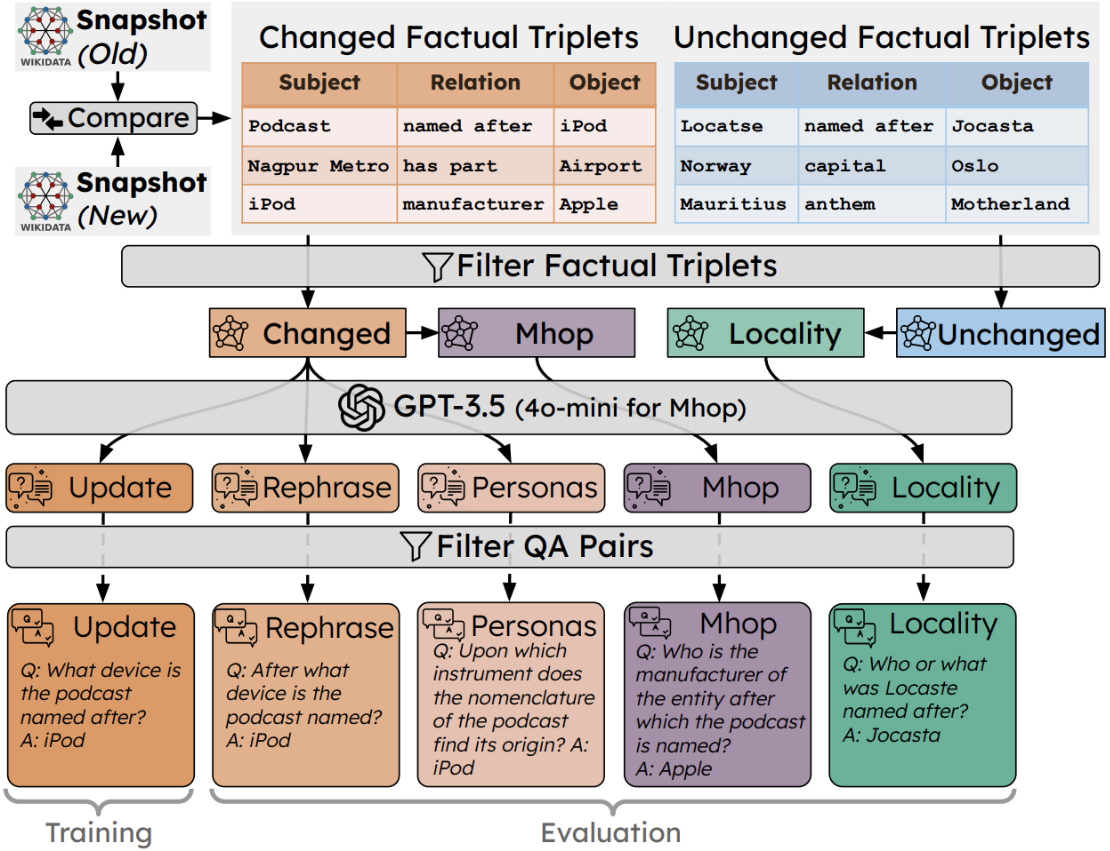
Required: Sequence of large batches of real world factual updates which cannot be solved sufficiently out of the box.

Take the **WikiData knowledge graph** as proxy for the „world knowledge“.

Record „changes in world knowledge“ as differences between knowledge graph snapshots.

Generate batches of factual QA updates based on the recorded changes.

WikiBigEdit: Dataset Generation



`WikiBigEdit`: Examples

Update	Who is the sibling of Lady Zhurong?	*Dailai Dongzhu*
Rephrase	Who is Lady Zhurong's sibling?	*Dailai Dongzhu*
Personas	So, do you know who Lady Zhurong's sibling is?	*Dailai Dongzhu*
Locality	Who is Bao Zhong's sibling?	*Bao Xin*
Mhop	What is the country of citizenship of the sibling of Lady Zhurong?	*Shu Han*

Update	Who is the author of the modern pentathlon?	*Stasys Saparnis*
Rephrase	Who created the modern pentathlon?	*Stasys Saparnis*
Personas	Arrr! Who be the scallywag penning the tales of the modern pentathlon, eh?	*Stasys Saparnis*
Locality	Who is the author of "Lamentation"?	*C. J. Sansom*
Mhop	Which country did the author of the modern pentathlon represent in sports?	*Soviet Union*

WikiBigEdit: Comparison

Benchmark	Size	Date	Data Source	Task	Lifelong	Mhop
NQ	307K	2016	Google Search queries	Open-domain factual QA	✗	✗
Trivia QA	650K	Various	Trivia sources (web)	Trivia QA	✗	✗
MS MARCO	1M+	2016	Bing Search queries	Search queries	✗	✗
Hotpot QA	112K	2018	Wikipedia (curated)	Multi-hop QA	✗	✓
FEVER	185K	2018	Human-written claims	Fact verification	✗	✗
EX-FEVER	60K	2023	Hyperlinked Wikipedia	Multi-hop fact verification	✗	✓
ELI5	270K	2019	Reddit (ELI5 subreddit)	Long-form explanatory QA	✗	✗
WikiQA	3.5K	2015	Wikipedia	Factoid QA	✗	✗
DatedData	200K	Various	Varied web sources	Temporal QA (time-sensitive)	✗	✗
StreamingQA	150K	2007-2020	WMT news articles	Real-time event-based QA	✗	✗
ArchivalQA	530K	1985-2008	Historical news archives	Fact-based QA	✗	✗
Hello Fresh	30K	2023-2024	X and Wikipedia	Fact verification	✗	✗
CLARK	1.4K	2021-2024	Wikipedia	Knowledge-intensive QA	✗	✗
PopQA	14k	2023	Wikipedia, Wikidata	Fact-based QA	✗	✗
TemporalWiki	7K	2021	Wikipedia, Wikidata	Temporal QA (time-sensitive)	✓	✗
WikiFactDiff	20k	2021-2023	Wikidata	Factual cloze tests	✗	✗
WikiBigEdit	**500K+**	**2024**	**Wikidata**	**Fact-based QA**	✓	✓

Large Scale Current Real world Lifelong Mhop Eval

`WikiBigEdit`: Takeaways

Timestep	Date Range	Samples	Unsolved
T1	2024/02/01 - 2024/02/20	26,790	80%
T2	2024/02/20 - 2024/03/01	32,901	84%
T3	2024/03/01 - 2024/03/20	54,802	85%
T4	2024/03/20 - 2024/04/01	43,554	85%
T5	2024/04/01 - 2024/05/01	121,754	81%
T6	2024/05/01 - 2024/06/01	101,550	82%
T7	2024/06/01 - 2024/06/20	69,251	82%
T8	2024/06/20 - 2024/07/01	55,433	82%
Total		**506,035**	**82%**

Most changes captured within the benchmark **can be considered as new facts due to being current and/or specific factual information**.

» **Benchmark can be used to assess the sequential integration of new factual updates at scale.**

Lifelong Knowledge Editing Comparison

Knowledge Editing Approaches

Local Modification
- ROME
- R-ROME
- MEMIT

External Memorization
- WISE

Evaluated Language Models

- Llama2-7b
- Mistral-7b
- Llama3-8b
- Gemma-7b
- xGen-7b

Model Modification Baselines

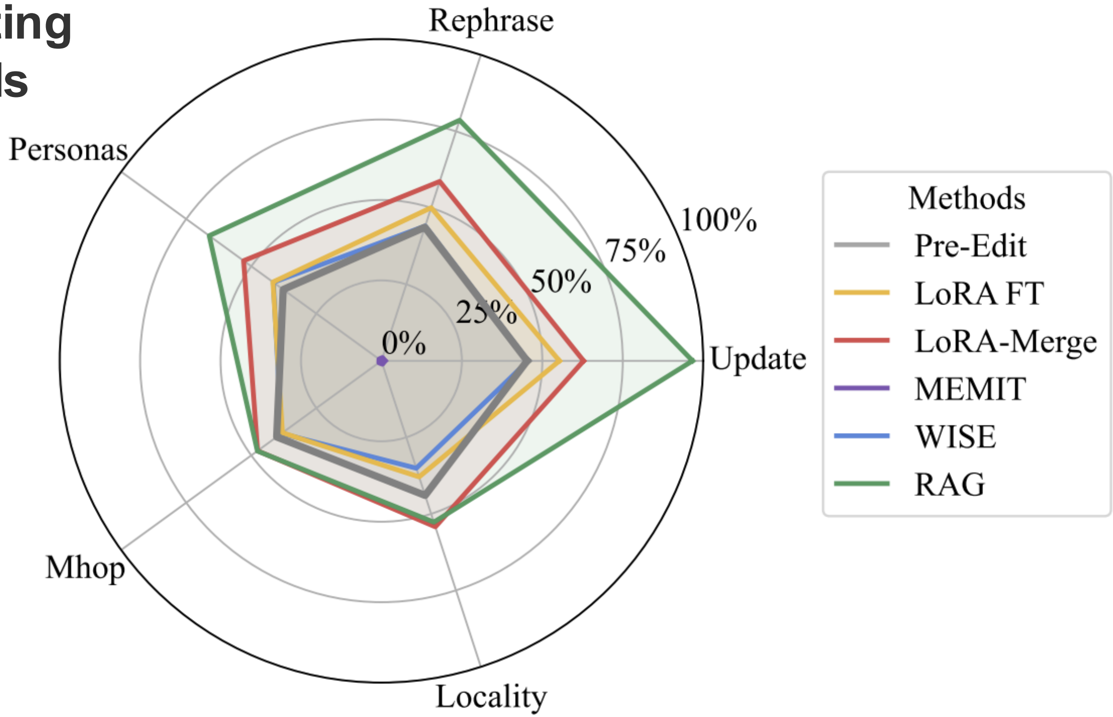
Search
- RAG

Continual Finetuning
- LoRA FT
- LoRA Merge

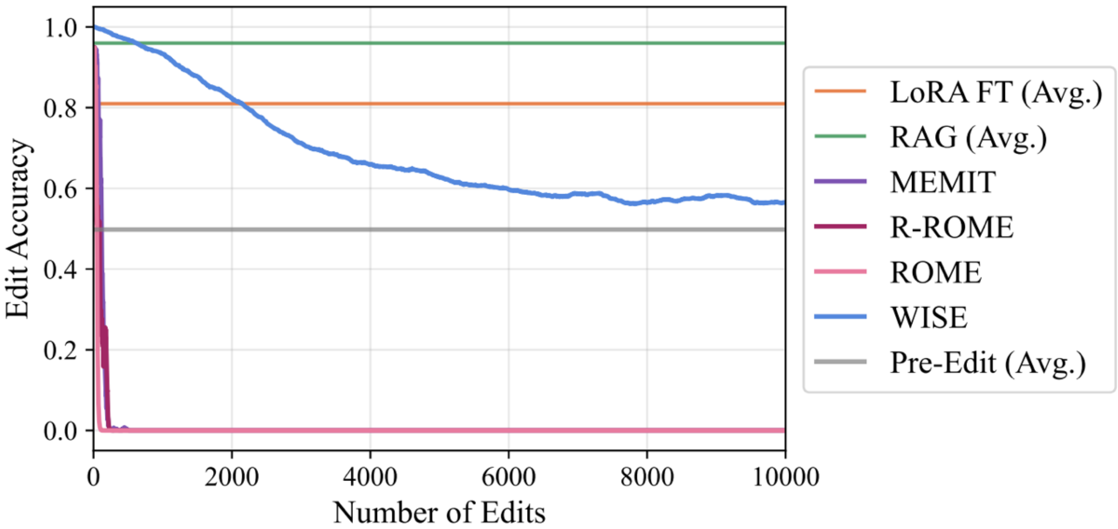
Lifelong Knowledge Editing Comparison

Comparison of knowledge editing techniques and other standards for model modification.

- **RAG vastly outperforms specialized knowledge editing** techniques (at higher inference cost).
- At **equivalent inference**, simple **continual finetuning consistently improves** on editing techniques at scale.



Knowledge Editing: Local Modification



- Local modification approaches **drop to zero accuracy** after <250 sequential edits as they break the model.
- Lifelong editing (i.e. through external memorization) does not break the model but **converges to pre-edit performance** within the first 10k edits.

Main Takeaways

Knowledge editing can (currently) **not facilitate large-scale sequential updates** to the factual knowledge of LLMs.

Continual finetuning with weight merging provides a **strong alternative with at equal inference compute**.

Retrieval augmentation proves to be **capable of incorporating large sequences of factual updates** at **increased inference cost**.

Thanks for your Attention!

Paper

Dataset

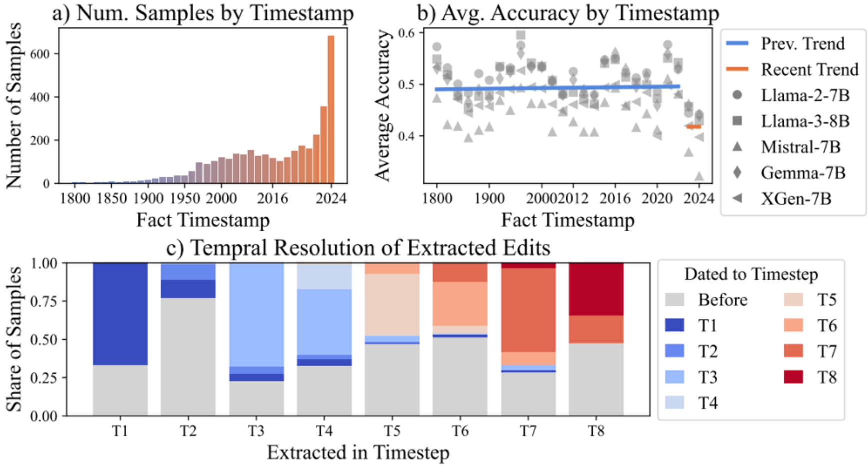
Code

Supplementary Material

`WikiBigEdit`: Analysis

Temporal analysis

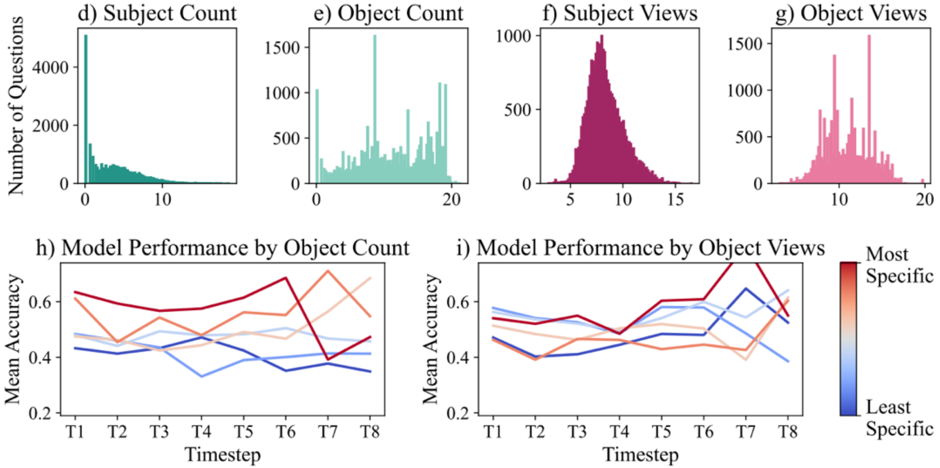
of the factual edits through timestamps extracted from the knowledge graph.



While most **facts are current**, the temporal consistency between batches does not necessarily hold consistently.

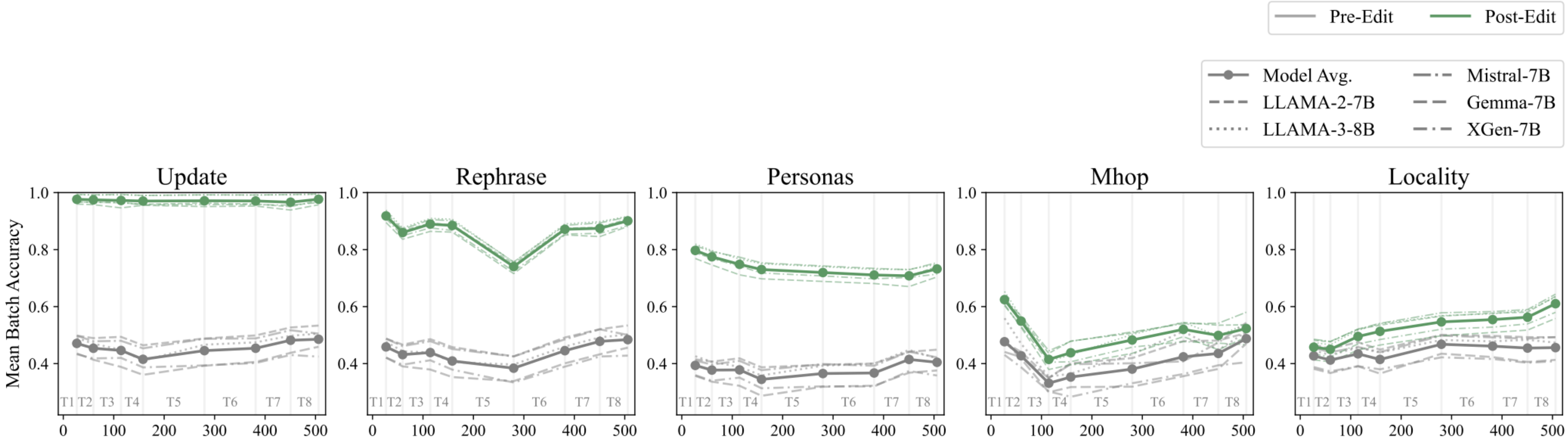
Quantification of question specificity

through training corpus frequencies and Wikipedia page views.



Specificity levels of individual facts can be quantified. Training corpus frequencies provide better measures than page views.

Retrieval Augmentation



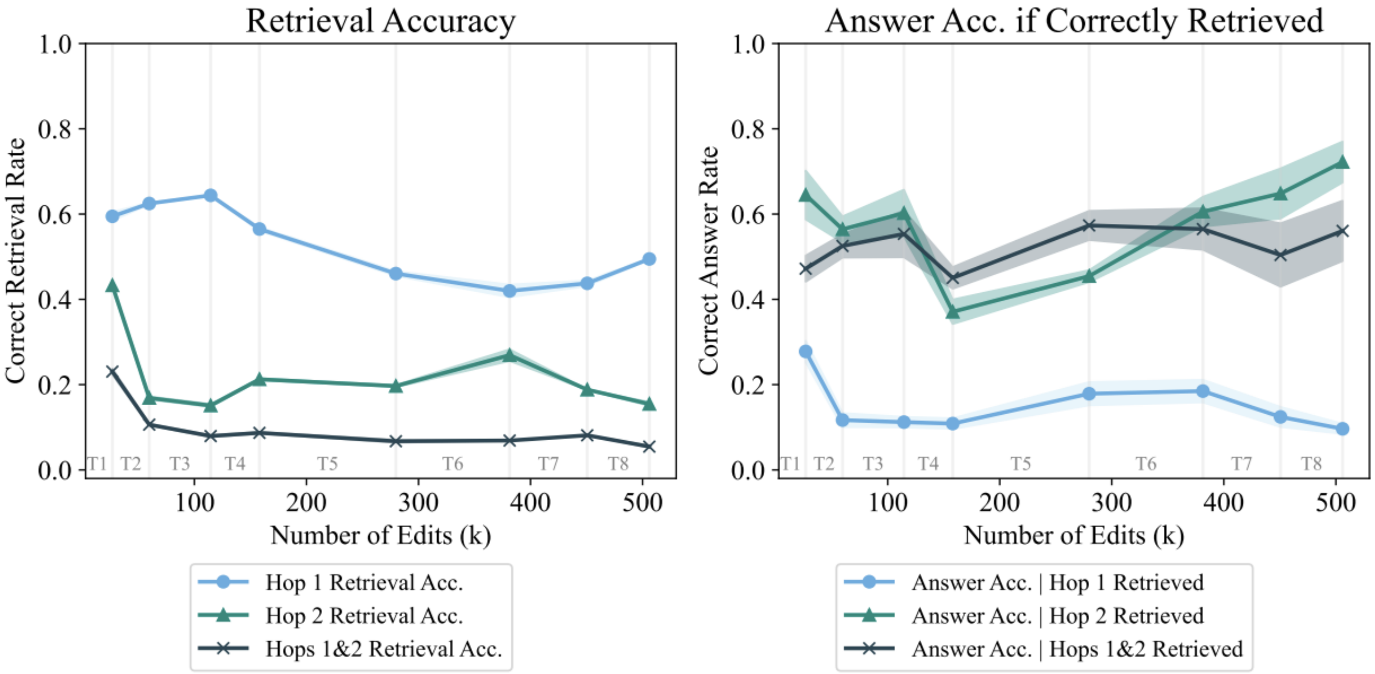
💡 Near-perfect performance on edits.

💡 Performance drops for rephrased questions.

💡 Limited reasoning capabilities.

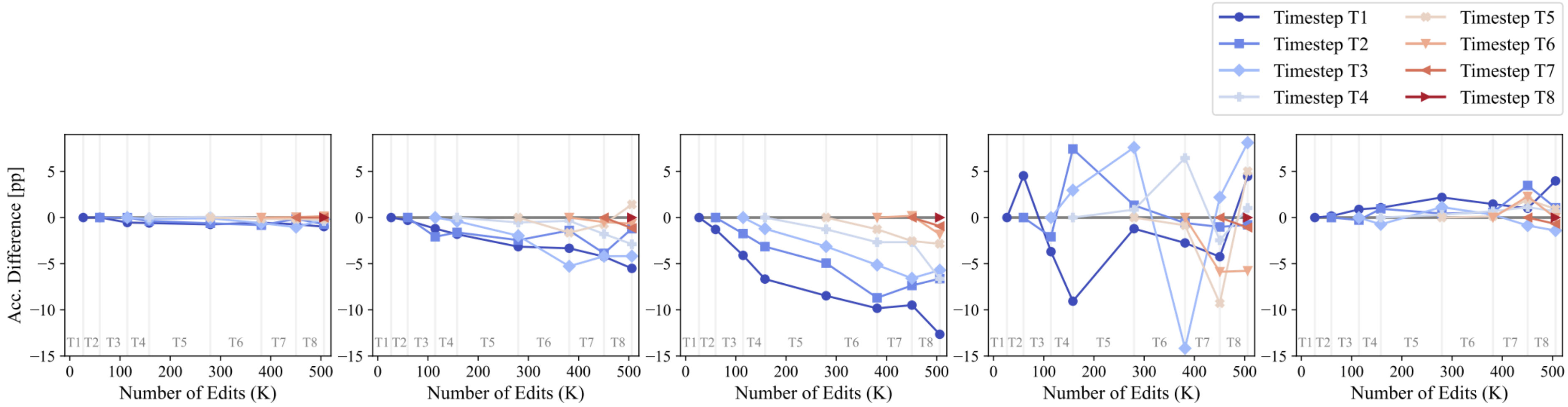
💡 No systematic spill-over to similar facts.

Retrieval Augmentation



Limited ability to reason upon edited knowledge due to errors in both retrieval and usage of edited knowledge.

Retrieval Augmentation



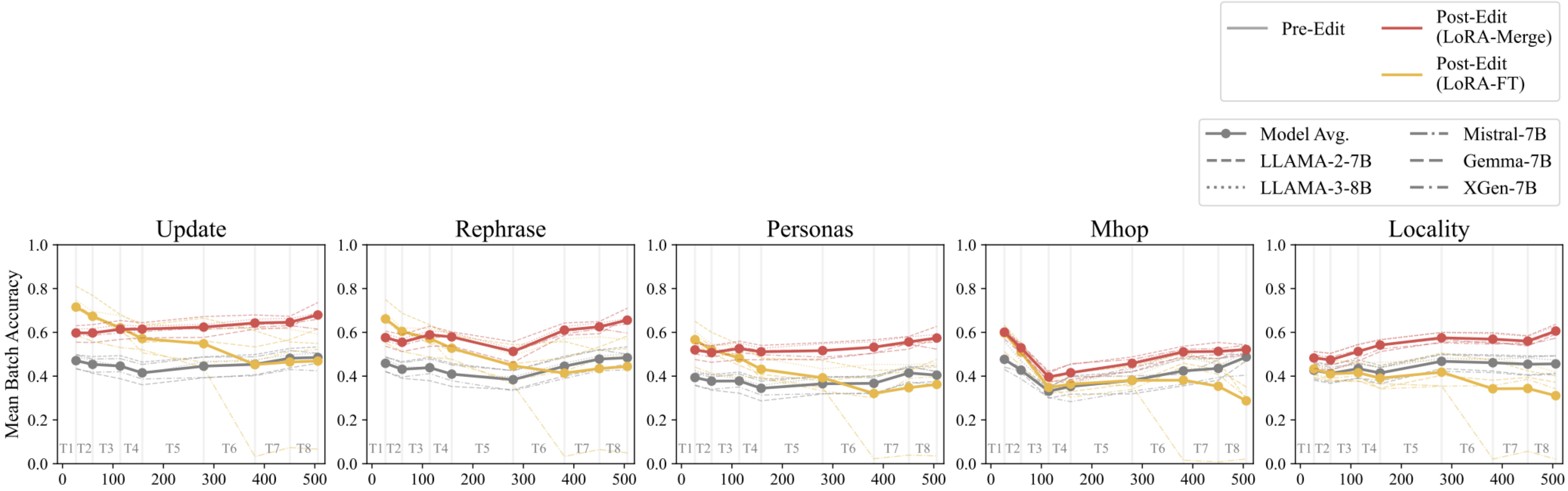
💡 **Performance decreases as new facts are added** to the memory, showing a limitation of RAG over long sequences of edits.

Retrieval Augmentation: Takeaways

Retrieval augmentation proves to be capable of incorporating large sequences of factual updates.

- Performance decreases as evaluation questions move away from the original edit.
- RAG shows limited ability to reason upon the updated factual knowledge, caused by incorrect retrieval and inability to combine the retrieved facts.
- Even with efficient solvers performance comes at higher inference cost.
- Performance on previous edits declines as more edits are integrated into the memory.

Continual Finetuning

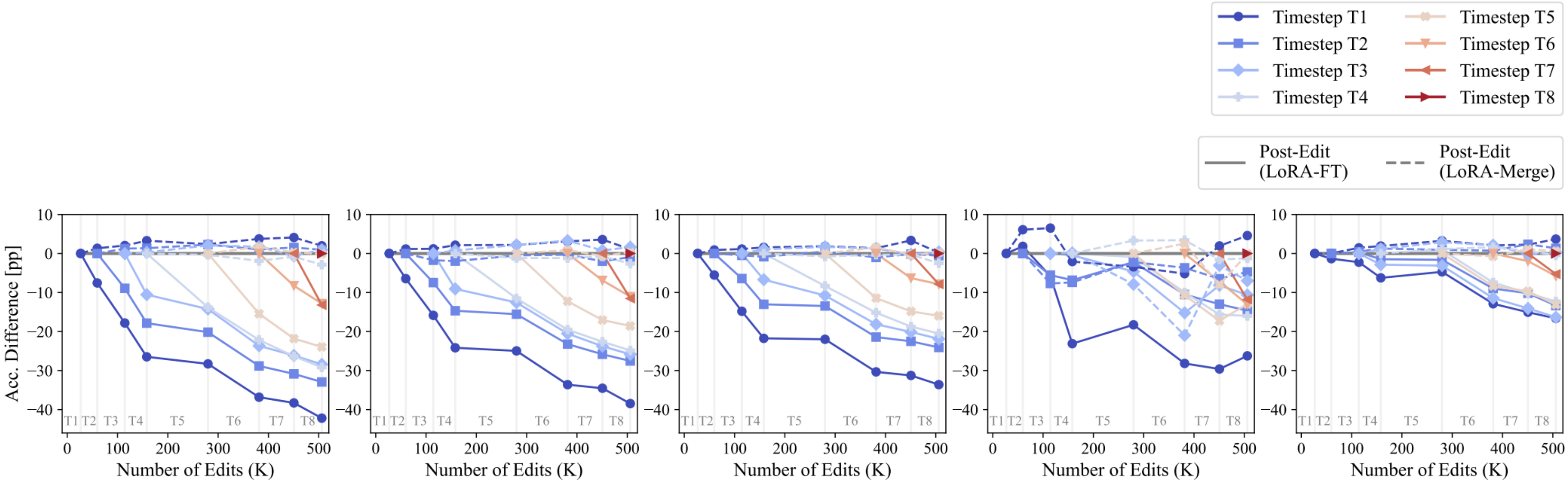


💡 Good initial performance for LoRA-FT followed by gradual decay. Merging allows to maintain performance level across updates.

💡 Limited reasoning capabilities.

💡 Systematic spill-over for LoRA-FT.

Continual Finetuning



Consistent strong forgetting for LoRA-FT.
LoRA-Merge shows no forgetting across metrics.

Topics of WikiBigEdit Update Batches

HELMHOLTZ MUNICH · EBERHARD KARLS UNIVERSITÄT TÜBINGEN

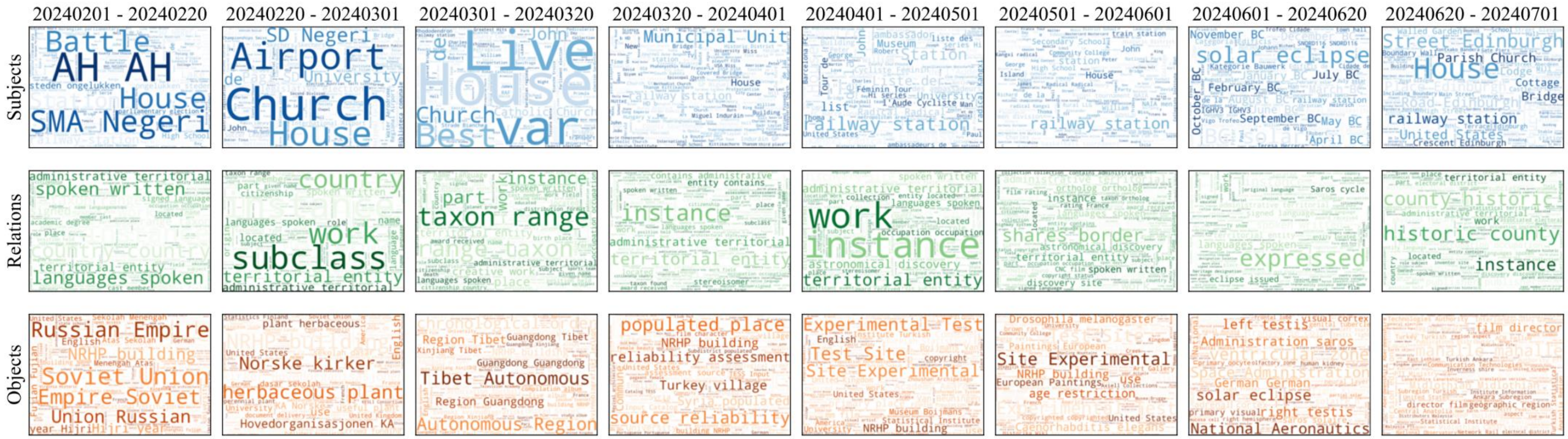


Figure 8. Topic word clouds for each timestep in the benchmark, illustrating the diversity of subjects, relations, and objects across different update intervals. Each row represents a component of the factual triplets (subjects, relations, and objects, respectively), while each column corresponds to a specific timestep. Notable patterns include a focus on specific events, such as “solar eclipse” in Timestep 7 (20240601–20240620), and varying distributions of topics across timesteps, emphasizing the richness and real-world relevance of the benchmark.

Analysis on the Mhop Questions

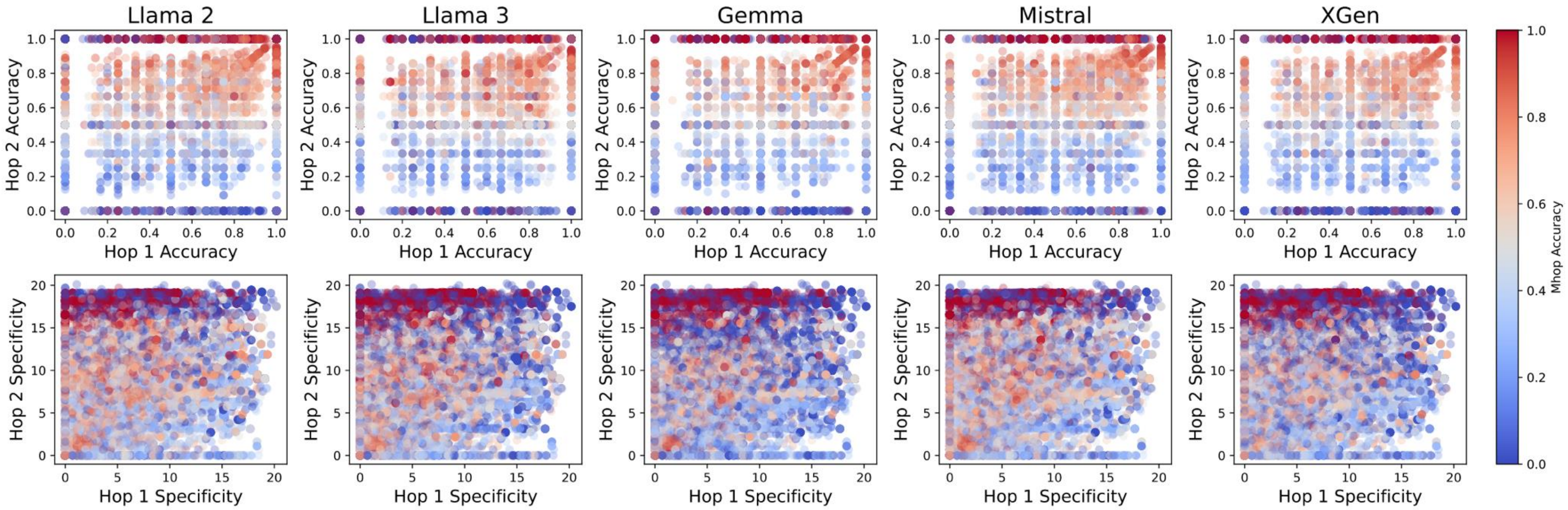


Figure 9. Multi-hop (mhop) question accuracy analysis for five models (Llama 2, Llama 3, Gemma, Mistral, XGen). The top row shows the relationship between the accuracy of answering questions from the first (Hop 1) and second (Hop 2) parts of mhop factual tuples and overall mhop accuracy (color-coded). Hop 2 accuracy strongly correlates with mhop question accuracy, highlighting its critical role in multi-hop reasoning. The bottom row explores the relationship between specificity (measured via entity-specificity scores) and accuracy for Hop 1 and Hop 2. Higher mhop accuracy is generally linked to lower specificity, emphasizing the challenge of highly specific entities.

RAG Top-k Ablation

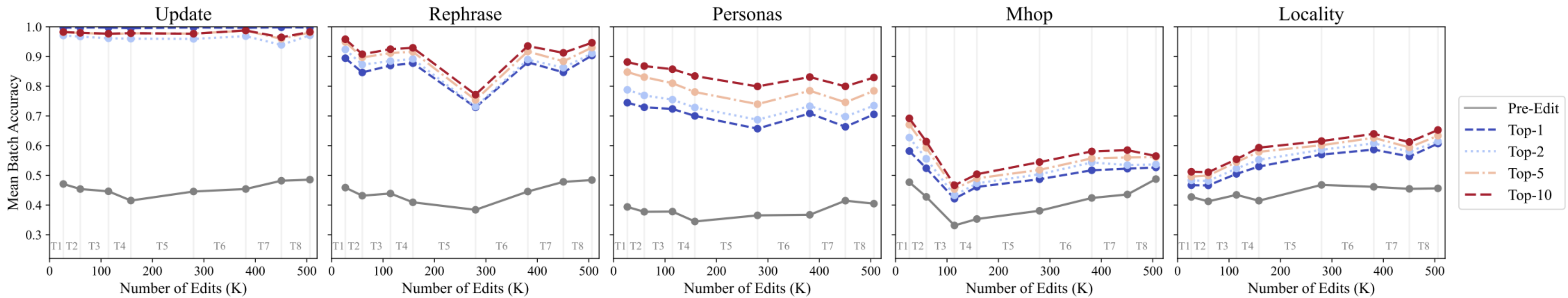


Figure 15. Ablation results for the top-k parameter of the RAG baseline, evaluating update, rephrase, personas, mhop, and locality sets across 500k updates. Higher top-k values improve performance on rephrase, personas, and mhop sets due to increased retrieval coverage, while update set accuracy remains consistent across top-k values. The locality set shows marginal gains with higher top-k values, indicating reduced spillover effects. However, increasing top-k also leads to greater computational overhead, highlighting the trade-off between retrieval depth and efficiency.

Inference Time Trade-off

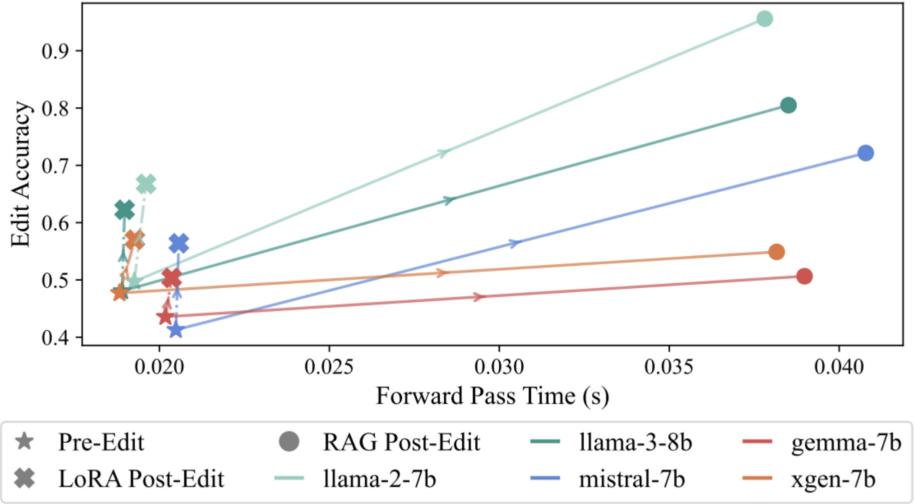


Figure 16. Trade-off between forward pass time (x-axis) and edit accuracy (y-axis) for RAG and LoRA across five models. Stars denote pre-edit performance, while post-edit performance for RAG and LoRA is represented by circles and crosses, respectively. RAG achieves higher edit accuracy at the cost of increased forward pass time, nearly doubling the average inference latency compared to LoRA. LoRA introduces minimal additional computational overhead while maintaining moderate accuracy improvements over the pre-edit baseline, highlighting its efficiency in resource-constrained settings.

LoRA Rank Ablation

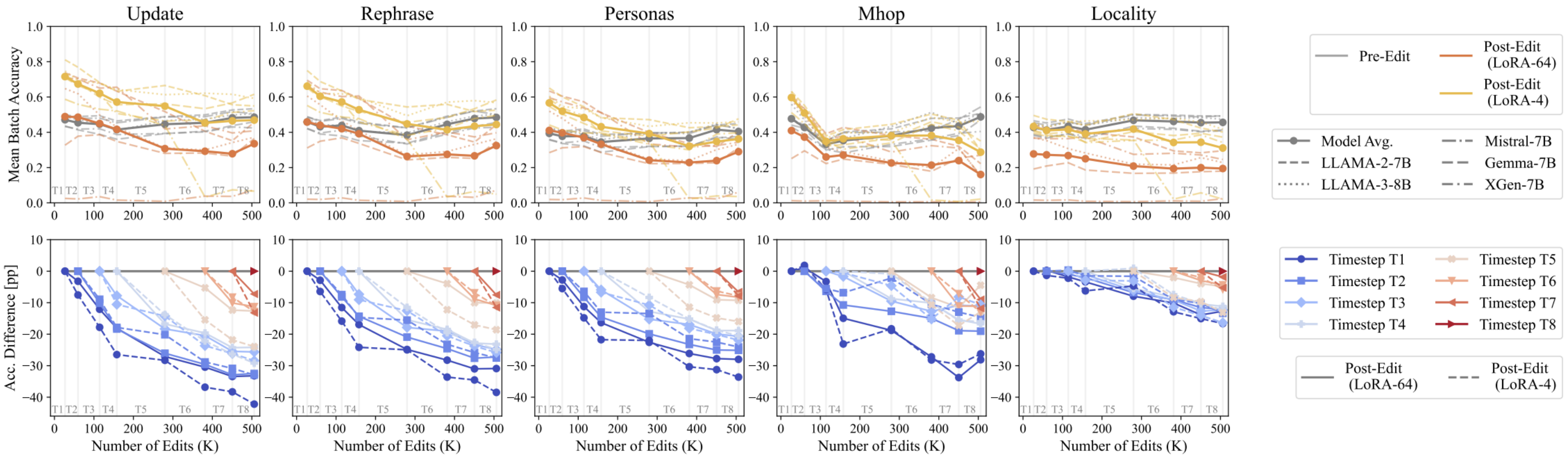


Figure 17. Performance of continual LoRA fine-tuning with high-rank (LoRA-64) and low-rank (LoRA-4) configurations across evaluation sets. The top row shows the mean batch accuracy for update, rephrase, personas, mhop, and locality sets, while the bottom row displays accuracy differences for individual timesteps. Low-rank configurations maintain competitive performance initially but degrade over time, while high-rank configurations show instability, including catastrophic failures in certain models.

LoRA Merging Alpha Ablation

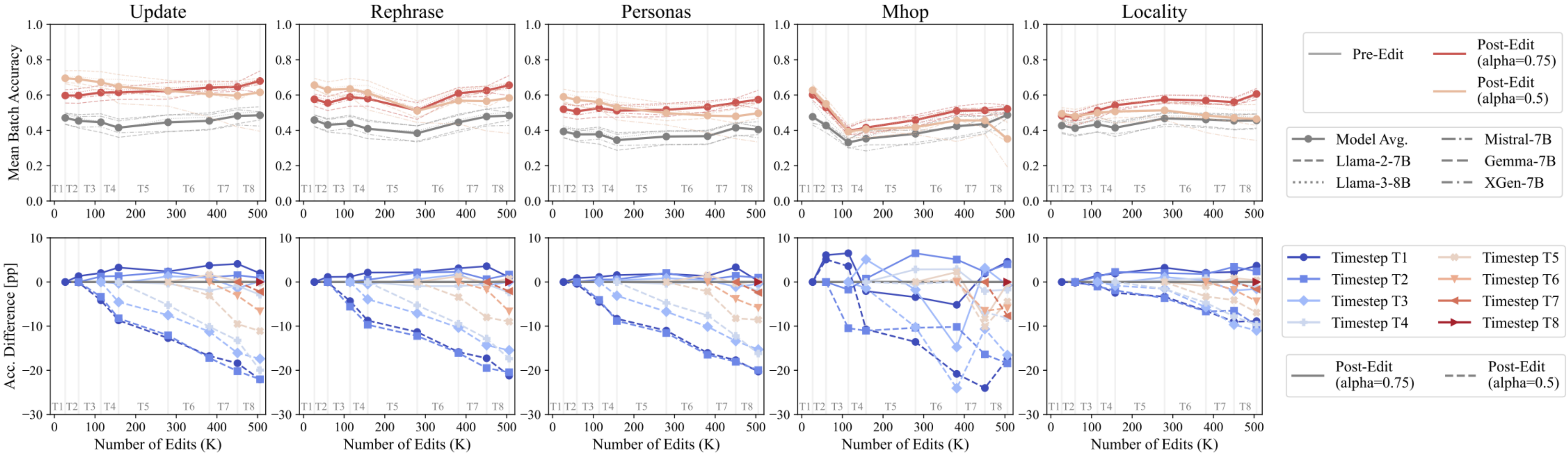


Figure 18. Ablation study on the interpolation factor α in the LoRA-Merge setting, comparing $\alpha = 0.75$ (red) and $\alpha = 0.5$ (orange). The top row shows the mean batch accuracy for the update, rephrase, personas, mhop, and locality sets, while the bottom row depicts accuracy differences (in percentage points) across timesteps. Higher weight on the base model ($\alpha = 0.75$) provides more stable performance and mitigates forgetting, while equal weighting ($\alpha = 0.5$) achieves higher initial performance but leads to greater degradation, particularly on the mhop and locality sets. Bold lines represent average model performance, with lighter lines for individual models.

Knowledge Editing Additional Results

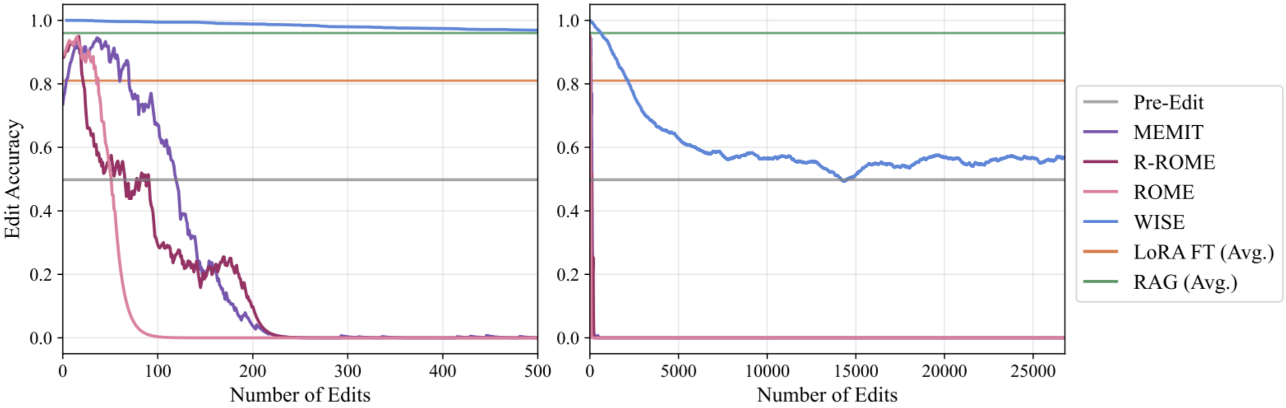


Figure 19. Update accuracy of knowledge editing approaches (MEMIT, R-ROME, ROME, and WISE) compared to LoRA-FT and RAG baselines across the first 500 updates (left) and the full first timestep (26k updates, right). Local modification methods (MEMIT, R-ROME, ROME) rapidly degrade within the first 250 updates, converging to near-zero performance. WISE initially performs on par with RAG for fewer than 500 updates but declines over the first 10k updates, converging to pre-update accuracy, highlighting its limitations in maintaining update accuracy for larger-scale knowledge integration.

MEMIT Batch Size Ablation

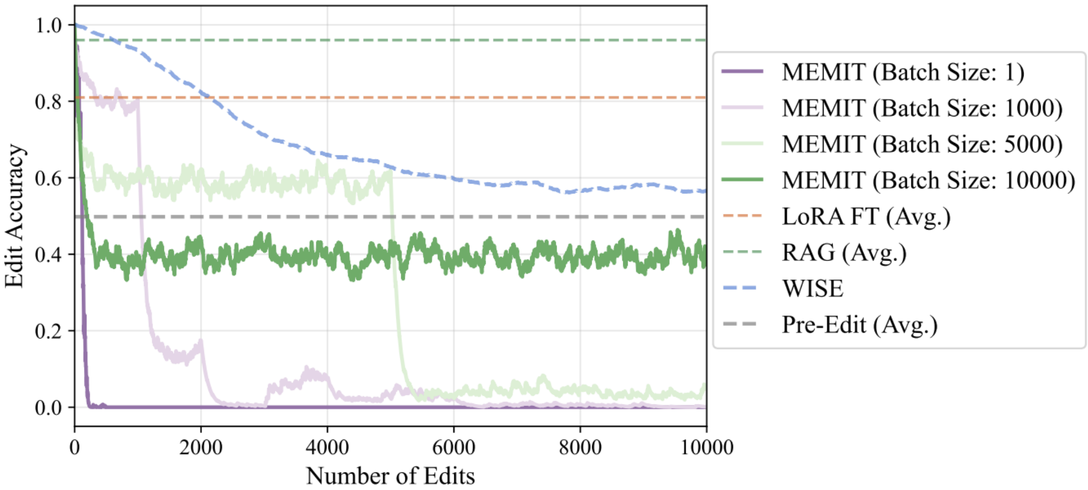


Figure 20. Impact of batch size on MEMIT's performance for the first 10k updates of the `WikiBigEdit` benchmark using Llama-2. Larger batch sizes (e.g., 10k edits) prevent model collapse but result in lower update accuracy compared to the model's pre-edit performance. Smaller batch sizes (e.g., 1 or 1k edits) perform well initially but exhibit rapid degradation in subsequent updates, highlighting MEMIT's limitations for sequential, large-scale lifelong knowledge editing.

WISE Additional Results

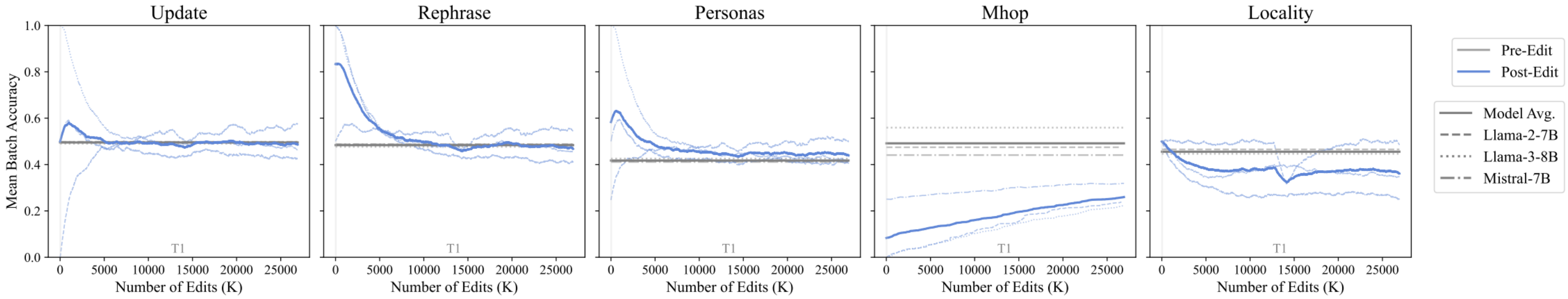


Figure 21. Results of WISE on three language models (Llama-2-7B, Llama-3-8B, and Mistral-7B) for the first timestep of the WikiBigEdit benchmark. Each subplot shows mean batch accuracy for the update, rephrase, personas, mhop, and locality sets over the number of updates. Post-update performance (blue) converges to pre-update levels (gray) across all metrics, highlighting WISE's limitations in sustaining accuracy after large-scale updates.