# P1

I'm going to share a paper, whose authors claim this paper accepted by ACL 2024 recently, titled "Learning to Edit: Aligning LLMs with Knowledge Editing".

# P2

There are four parts in this presentation.

# P3

First part is Introduction, explaining which problem does this paper focus on.

[PPT]

# P4

And it mainly involves two areas, **Knowledge Editing** and **LLM Alignment**.

- For Knowledge Editing,
- For LLM Alignment,

# P5

They point out that ...

The figure is presented to compare the principles between previous methods and their method.

# P6

This figure show the framework of their method.

There are two phases. Firstly, during the alignment phase, they train the LLM to learn to use updated information. Secondly, they use the retrieval model to get the updated information, and combine this with the trained model to verify the performance.

# P7

[PPT]

# P8

There are three critical capabilities during the alignment phase.

[PPT]

# P9

[PPT]

Their methodology only incorporates samples from the datasets' training sets. Due to the absence of out-of-scope instances in datasets like ZsRE and MQUAKE

For linguistic capability, [PPT]

# P10

For inference phase, To enhance the fault tolerance of the retrieval model while maintaining the single editing performance,

Firstly, in 50% of cases, they directly use the exact edit descriptor. Secondly, for 25% of cases, employ the multi-qa-mpnet-base-dot-v1 model to identify the top-1 semantically similar edit descriptor (excluding the exact one) from the whole dataset, and use both as the Updated Information. Lastly, for the remaining 25%, retrieve the top 2 semantically similar descriptors, excluding the exact one, using all three as the Updated Information.

# P11

In the experiments, the paper explains four attributes they care about.

- Edit success measures the average accuracy of the post-edit model on edit cases
- Portability evaluates how well updated knowledge transfers to related queries, enhancing the model’s utility in varied contexts. (For example, correctly answering Who is married to the British Prime Minister? with Akshata Murty post-edit indicates successful knowledge transfer.)
- Locality assesses the precision of edits, ensuring modifications are confined to targeted areas without affecting unrelated knowledge. (For example, ensuring The current British Chancellor remains Jeremy Hunt exemplifies effective locality.)
- Fluency quantifies the linguistic quality of the model’s output post-edit, focusing on coherence and diversity to avoid repetitive patterns. calculate fluency by measuring the weighted average of bi- and tri-gram entropies.

# P12

Performance comparison on Single Editing, where “Recent” and “Counterfact” refer to WikiData~recent~ and WikiData~counterfact~, respectively. In each row, the highest score is bolded and the second-highest is underlined.

- SERAC builds a counterfact model by retaining the base model and training a classifier to determine whether to use the counterfact model to answer the query.
- ICE prepends a prompt “Imagine that {edit descriptor}” before the query. It does not introduce changes to the model parameters, but rather generation is conditioned on the new fact.
- MEND learns to locate factual retrievals of a specific set of MLP modules and update knowledge by directly writing in new key-value pairs in the MLP module.
- MEMIT builds upon ROME to insert many memories by modifying the MLP weights of a range of critical layers.
- FT-L directly fine-tunes a single layer’s FFN, and the layer is the casual tracing results in ROME.
- FT fine-tunes all the parameters of the base model on the edit descriptor by applying Adam with early stopping.

In the paper, they claim their method surpass the SOTA method-SERAC on LLaMA2-Chat-7B and Qwen-Chat-7B.

# P13

They test mass editing, for Batch Editing, they compare LTE and LTELoRA with several batch-editing-supportive methods (SERAC, MEMIT, and FT-L) on LLaMA2Chat-7B.

They methodologies exhibit excep tional stability, maintaining robustness for up to 1,000 batch edits. especially, demonstrate the best performance with the slowest degradation rate in portability and locality.

# P14

For Sequential Editing, it means models must retain previous modifications while integrating new edits effectively. They eplain their methods leverage retrieval mechanisms from the stored memory, circumventing the need for subsequent parameter modifications, which endows them with more consistent performance with varying data stream sizes.

# P15

They investigate the impact of applying LTE on the performance of a language model across various domains. Aims to determine whether the Alignment Phase of LTE, which alters the parameters of the initial model, inadvertently compromises the model’s competence in unrelated domains.

although a performance decrement is noted in CommonsenseQA and PIQA, from a comprehensive standpoint, they think the general linguistic abilities remain unaffected by the inclusion of the knowledge editing prompt.

# P16

Results for one case of different editing methods based on LLaMA2-Chat-7B. Queries are underlined and italicized. Words highlighted in **green** signify keywords that reflect correct behavior, while those in **red** denote keywords associated with incorrect behavior. Texts in **cyan** are repeated or meaningless sentences.