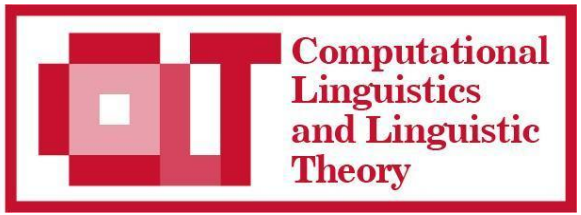
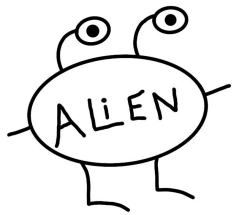
LREC-COLING 2024

MemoryPrompt: A Light Wrapper to Improve Long-Distance Information Tracking in Pre-trained Language Models

Nathanaël Carraz Rakotonirina, Marco Baroni





Objective

We want language models to **keep track** of information

- without necessarily hard-coding a huge **context window**
- without affecting their **original performance**.

Fact-updating dataset

- We collected facts (<*subject,relation,object*> tuples) from **TReX**.
- A fact is **mutable** if its object can be updated over time.
- We identified 3 highly mutable relations and 36 essentially stable relations.

Relation	**Template**
work location	[X] took up work in [Y].
position held	[X] holds the public role of [Y].
employer	[X] works for [Y].
original network	[X] was originally aired on [Y].
place of birth	[X] was born in [Y].
place of death	[X] passed away in [Y].
original language	The original language of [X] is [Y] .

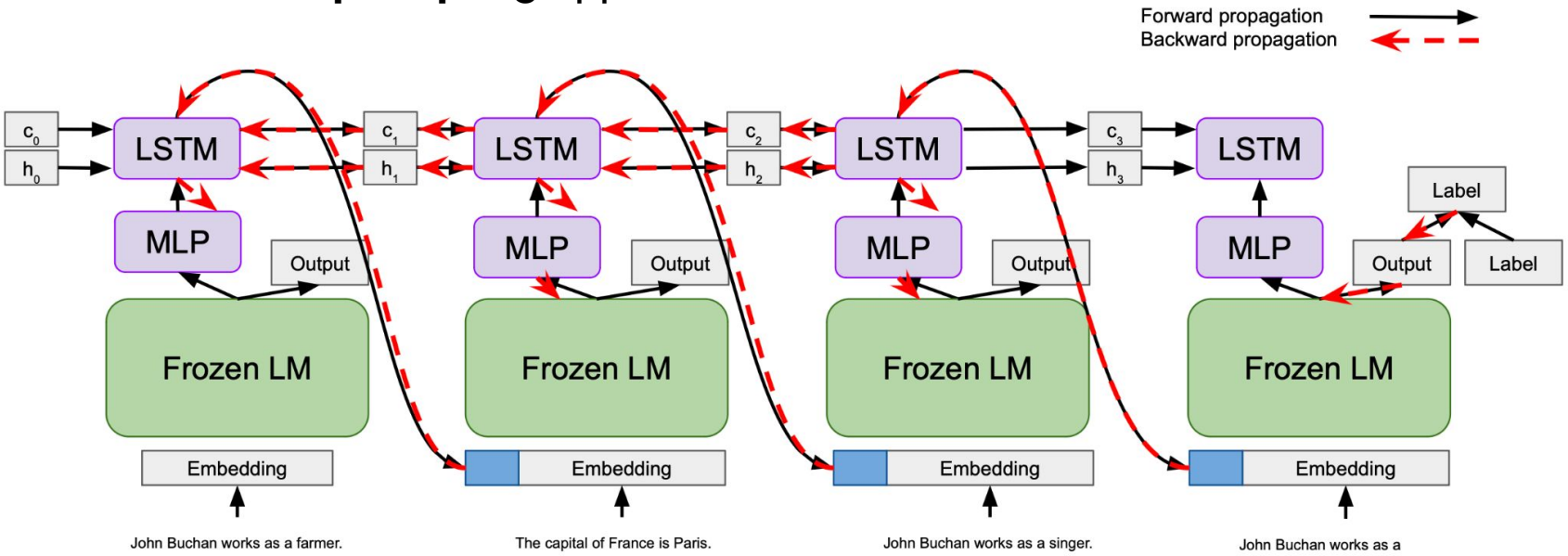
Fact-updating dataset

- The dataset is composed of sequences of facts.
- The **pivot** is the mutable fact to track.
- The **distractors** are mutable facts distinct from the pivot.
- The dataset parameters are the number of sequences, the number of pivots and the number of updates.

Input: Let It Be's label is Apple. Guido Pepoli holds the public role of cardinal. Jeff De Luca works for IBM. Nepal Ministry of Foreign Affairs is a valid legal term in Nepal. Stephen V holds the public role of Shah. Justo Oscar Laguna holds the public role of bishop. The headquarter of Tejarat Bank is in Tehran. Benedict III holds the public role of pope. Stephen V holds the public role of governor. premier is a valid legal term in Canada. Zerai Deres passed away in Rome. Kjell Lönnå holds a citizenship of Sweden. Guido Pepoli holds the public role of bishop. Benoît Lengelé works in the field of surgeon. People's Republic of China shares the borders with India. Josef Gingold plays violin.
Question: Stephen V held the public role of
Answer: governor

MemoryPrompt

- The input is divided into **segments** processed sequentially.
- We augment a pre-trained LM with a **memory module** that produces **vectors.**
- The vectors are concatenated to the embeddings of the next segment as in standard **soft prompting** approaches..

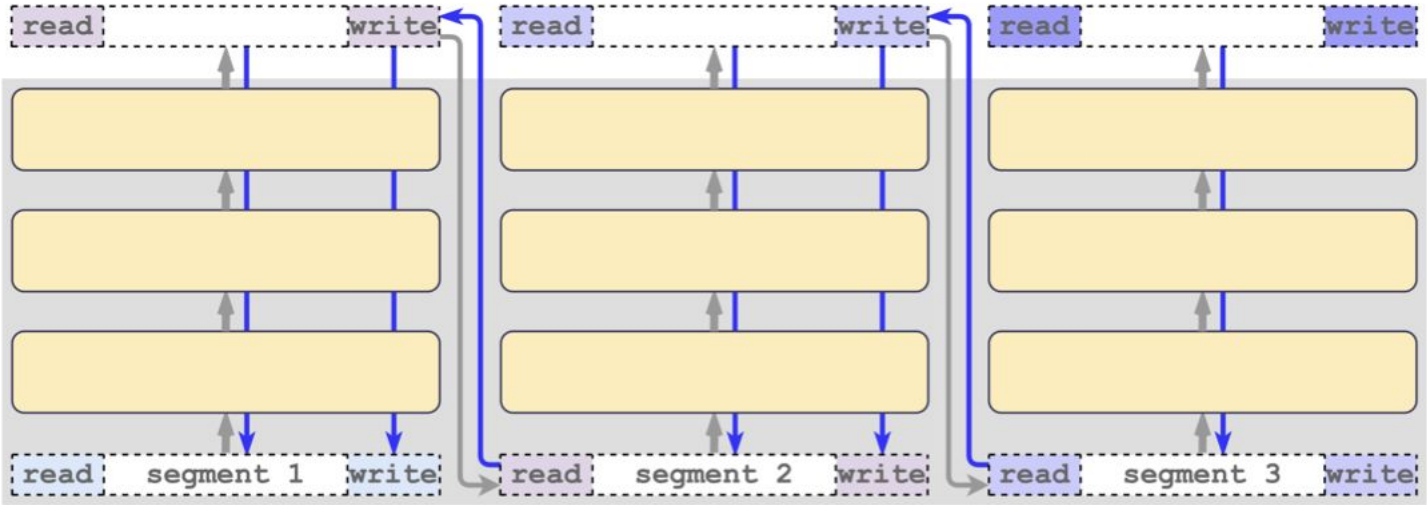


Training MemoryPrompt

- We use LMs form the OPT family.
- Only the parameters of the memory module are updated. The LM is frozen.
- The augmented system is trained with (truncated) **backpropagation through time** (BPTT).
- We use **curriculum learning** when training on long sequences.

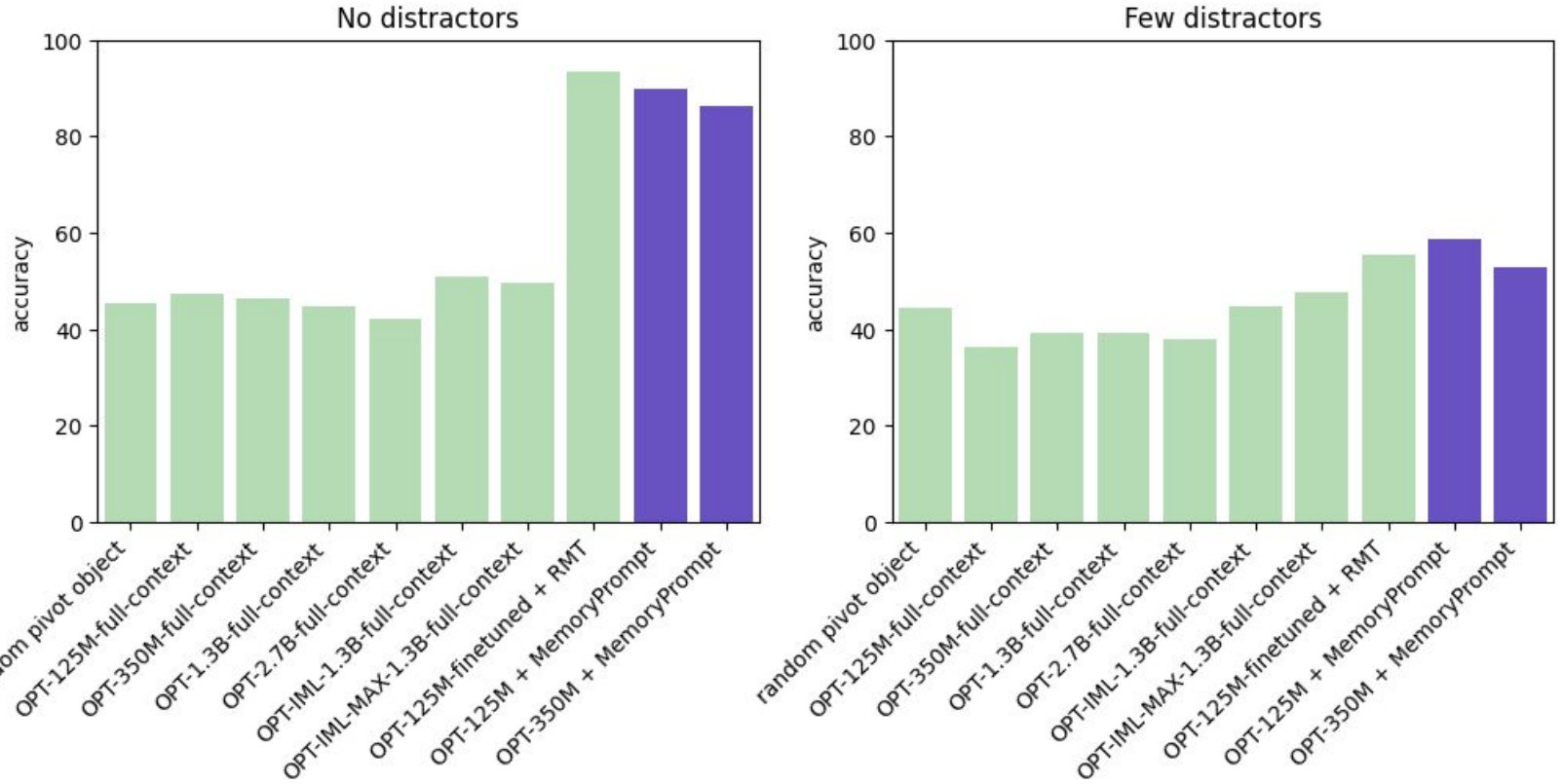
Baselines

- Full-context LM
- Recurrent Memory Transformer (RMT)

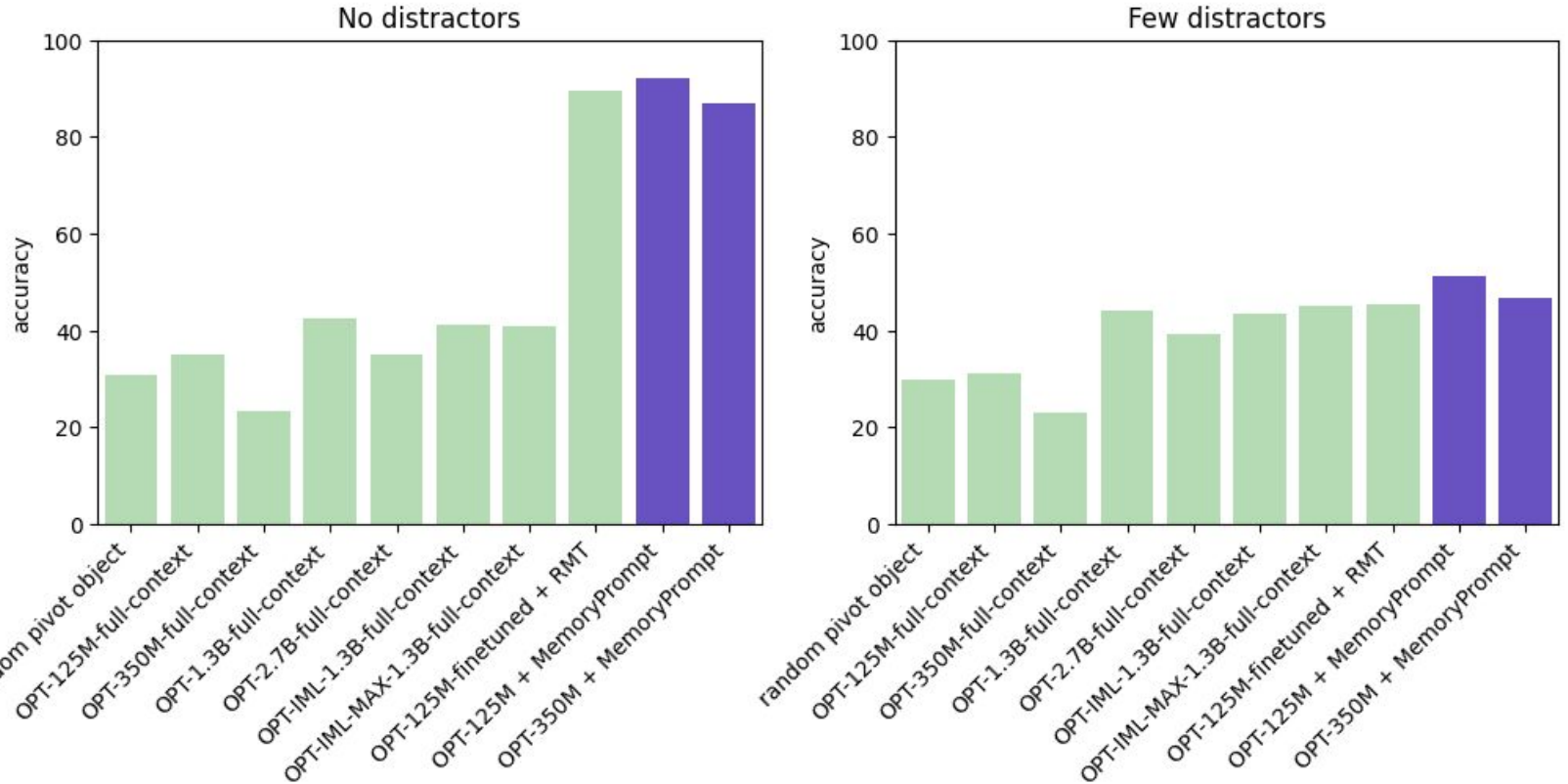


Bulatov, A., Kuratov, Y. and Burtsev, M., 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, *35*, pp.11079-11091.

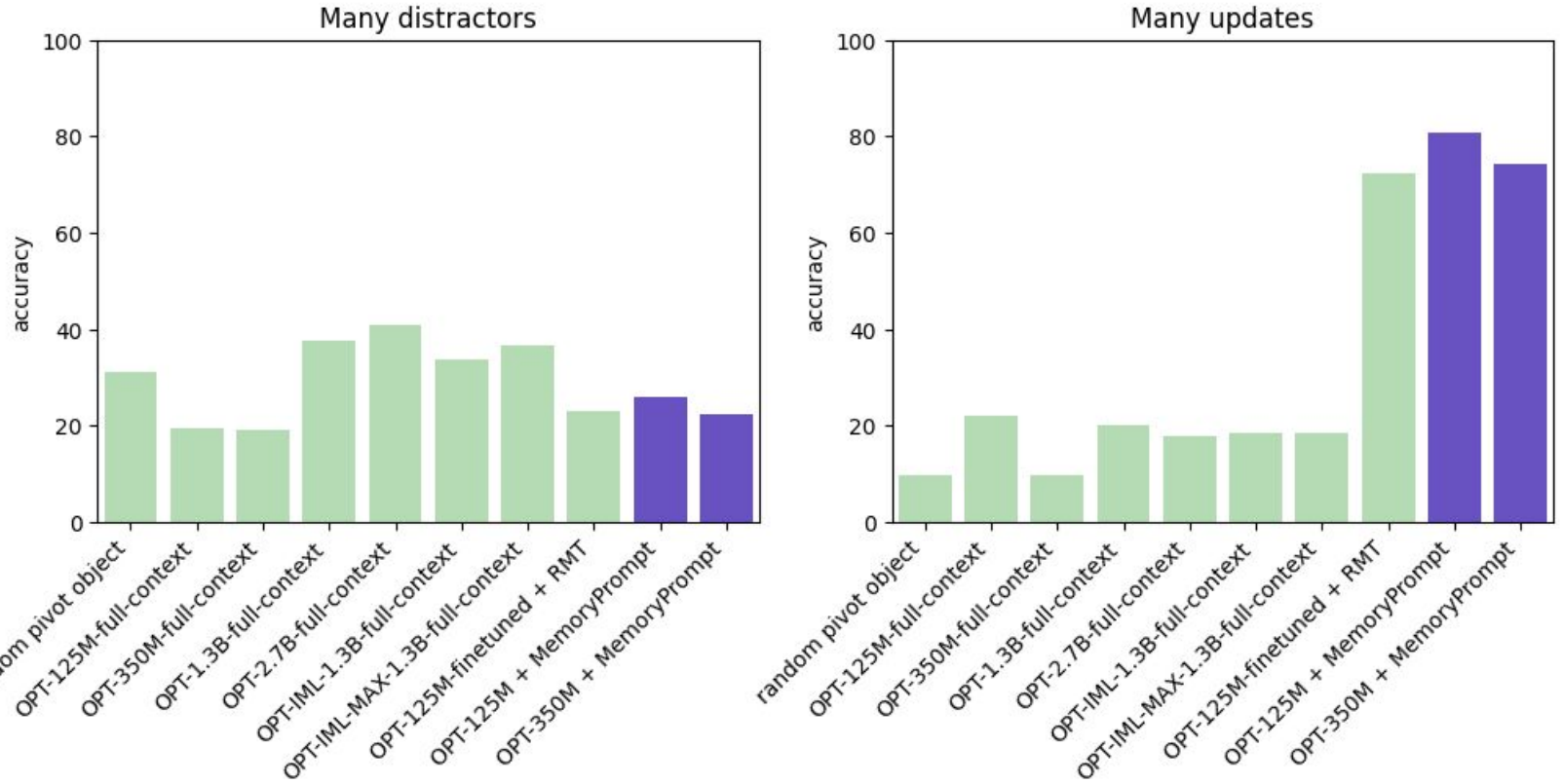
Results - Short datasets



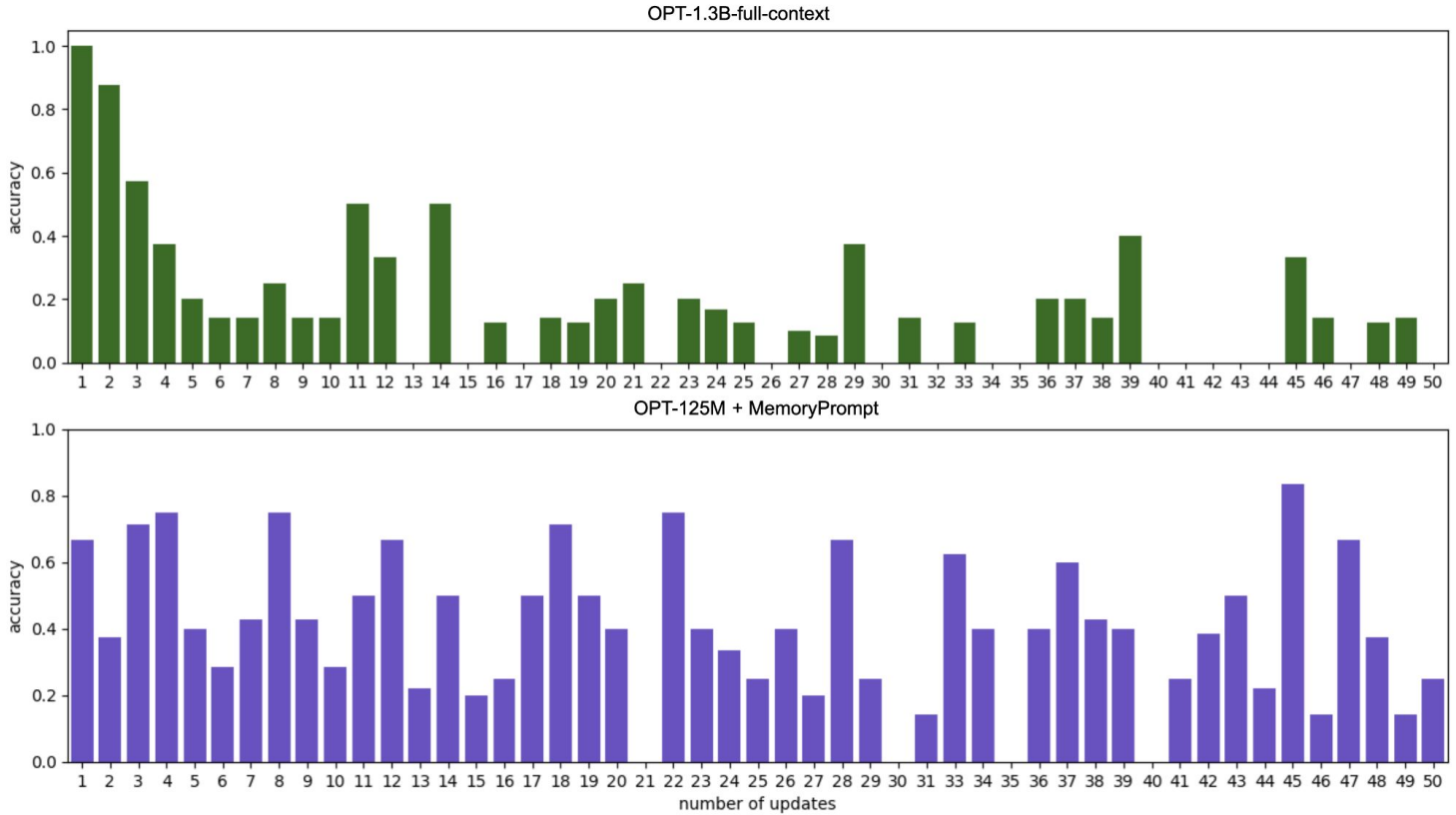
Results - Long datasets



Results - Long datasets

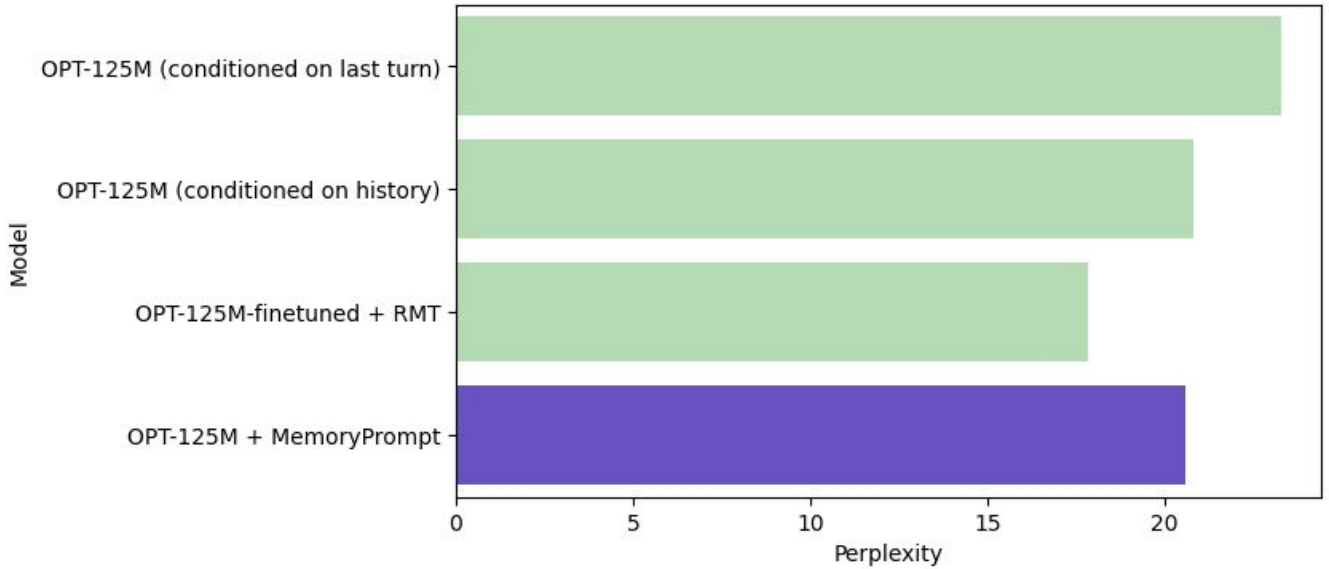


Robustness to update numbers



Multi-Session Chat (MSC)

MSc is a long-term conversation dataset which consists of multi-session crowdworker chats, where the speakers might refer to previous sessions.

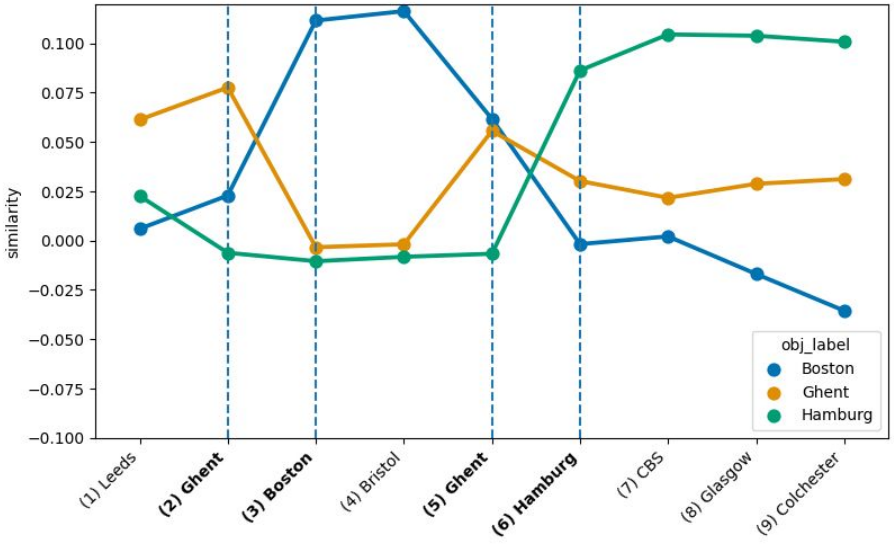


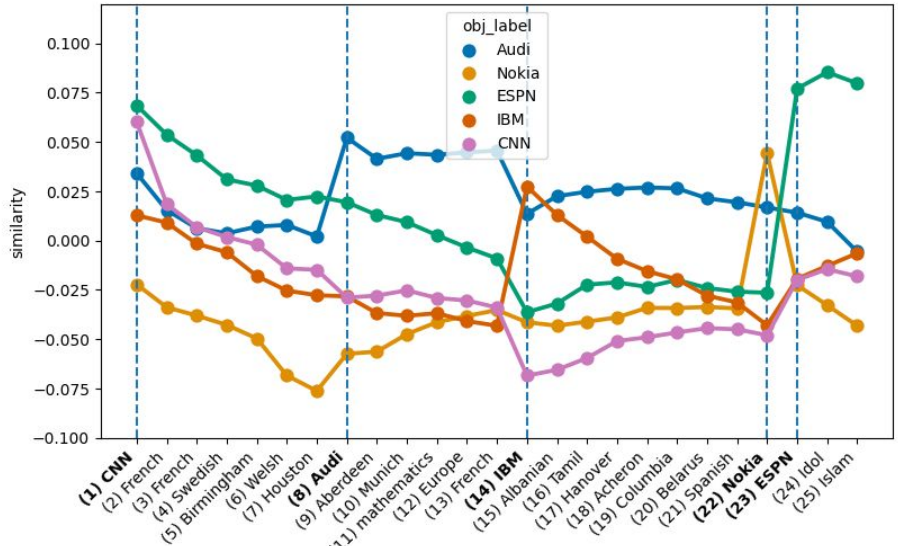
Catastrophic forgetting

- Experiments suggest that MemoryPrompt performs at the level of RMT
- However, unlike RMT, adding MemoryPrompt does not affect the language modeling ability of the base model.
- It retains most of the facts already present in the LM.
- **Forgetting rate** is the proportion of facts for which the augmented model gives a different completion than the base model.

Model	Forgetting rate on TReX	Perplexity on Wiki-Text-103
OPT-125M	0	27.68
OPT-125M + MemoryPrompt	13	27.77
OPT-125M-finetuned + RMT	97.4	11455.11

Memory vector analysis





QUESTIONS

Summary

- We add a light memory module to a LM allowing it to keep track of long-distance information beyond its input window.
- MemoryPrompt outperforms full input context and performs as well as methods that finetune the LM, without incurring into catastrophic forgetting.
- MemoryPrompt is robust to the number of updates and to window length, but not to the number of entities that need to be tracked.
- We still understand very little about how information is carried through the memory vectors.