

# MASS-EDITING MEMORY IN A TRANSFORMER

Kevin Meng<sup>1,2</sup> Arnab Sen Sharma<sup>2</sup> Alex Andonian<sup>1</sup> Yonatan Belinkov<sup>†3</sup> David Bau<sup>2</sup>  
<sup>1</sup>MIT CSAIL <sup>2</sup>Northeastern University <sup>3</sup>Technion – IIT

## ABSTRACT

Recent work has shown exciting promise in updating large language models with new memories, so as to replace obsolete information or add specialized knowledge. However, this line of work is predominantly limited to updating single associations. We develop MEMIT, a method for directly updating a language model with many memories, demonstrating experimentally that it can scale up to *thousands of associations* for GPT-J (6B) and GPT-NeoX (20B), exceeding prior work by orders of magnitude. Our code and data are at [memit.baulab.info](https://memit.baulab.info).

## 1 INTRODUCTION

How many memories can we add to a deep network by directly editing its weights?

Although large autoregressive language models (Radford et al., 2019; Brown et al., 2020; Wang & Komatsuzaki, 2021; Black et al., 2022) are capable of recalling an impressive array of common facts such as “Tim Cook is the CEO of Apple” or “Polaris is in the constellation Ursa Minor” (Petroni et al., 2020; Brown et al., 2020), even very large models are known to lack more specialized knowledge, and they may recall obsolete information if not updated periodically (Lazaridou et al., 2021; Agarwal & Nenkova, 2022; Liska et al., 2022). The ability to maintain fresh and customizable information is desirable in many application domains, such as question answering, knowledge search, and content generation. For example, we might want to keep search models updated with breaking news and recently-generated user feedback. In other situations, authors or companies may wish to customize models with specific knowledge about their creative work or products. Because re-training a large model can be prohibitive (Patterson et al., 2021) we seek methods that can update knowledge directly.

To that end, several *knowledge-editing* methods have been proposed to insert new memories directly into specific model parameters. The approaches include constrained fine-tuning (Zhu et al., 2020), hypernetwork knowledge editing (De Cao et al., 2021; Hase et al., 2021; Mitchell et al., 2021; 2022), and rank-one model editing (Meng et al., 2022). However, this body of work is typically limited to updating at most a few dozen facts; a recent study evaluates on a maximum of 75 (Mitchell et al., 2022) whereas others primarily focus on single-edit cases. In practical settings, we may wish to

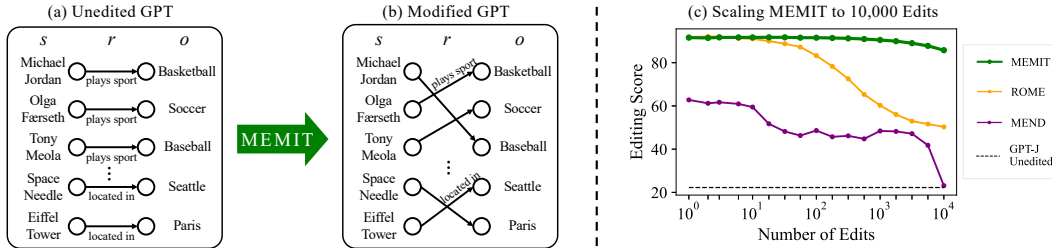


Figure 1: **MEMIT is capable of updating thousands of memories at once.** (a) Language models can be viewed as knowledge bases containing memorized tuples  $(s, r, o)$ , each connecting some subject  $s$  to an object  $o$  via a relation  $r$ , e.g.,  $(s = \text{Michael Jordan}, r = \text{plays sport}, o = \text{basketball})$ . (b) MEMIT modifies transformer weights to edit memories, e.g., “Michael Jordan now plays the sport baseball,” while (c) maintaining generalization, specificity, and fluency at scales beyond other methods. As Section 5.2.2 details, editing score is the harmonic mean of efficacy, generalization, and specificity metrics.

<sup>†</sup>Supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.  
 Correspondence to [mengk@mit.edu](mailto:mengk@mit.edu), [davidbau@northeastern.edu](mailto:davidbau@northeastern.edu).