update a model with hundreds or thousands of facts simultaneously, but a naive sequential application of current state-of-the-art knowledge-editing methods fails to scale up (Section 5.2).

We propose MEMIT, a scalable multi-layer update algorithm that uses explicitly calculated parameter updates to insert new memories. Inspired by the ROME direct editing method (Meng et al., 2022), MEMIT targets the weights of transformer modules that we determine to be causal mediators of factual knowledge recall. Experiments on GPT-J (6B parameters; Wang & Komatsuzaki 2021) and GPT-NeoX (20B; Black et al. 2022) demonstrate that **MEMIT can scale and successfully store thousands of memories in bulk**. We analyze model behavior when inserting true facts, counterfactuals, 27 specific relations, and different mixed sets of memories. In each setting, we measure robustness in terms of generalization, specificity, and fluency while comparing the scaling of MEMIT to rank-one, hypernetwork, and fine-tuning baselines.

2 Related Work

Scalable knowledge bases. The representation of world knowledge is a core problem in artificial intelligence (Richens, 1956; Minsky, 1974), classically tackled by constructing *knowledge bases* of real-world concepts. Pioneering hand-curated efforts (Lenat, 1995; Miller, 1995) have been followed by web-powered knowledge graphs (Auer et al., 2007; Bollacker et al., 2007; Suchanek et al., 2007; Havasi et al., 2007; Carlson et al., 2010; Dong et al., 2014; Vrandečić & Krötzsch, 2014; Bosselut et al., 2019) that extract knowledge from large-scale sources. Structured knowledge bases can be precisely queried, measured, and updated (Davis et al., 1993), but they are limited by sparse coverage of uncatalogued knowledge, such as commonsense facts (Weikum, 2021).

Language models as knowledge bases. Since LLMs can answer natural-language queries about real-world facts, it has been proposed that they could be used directly as knowledge bases (Petroni et al., 2019; Roberts et al., 2020; Jiang et al., 2020; Shin et al., 2020). However, LLM knowledge is only implicit; responses are sensitive to specific phrasings of the prompt (Elazar et al., 2021; Petroni et al., 2020), and it remains difficult to catalog, add, or update knowledge (AlKhamissi et al., 2022). Nevertheless, LLMs are promising because they scale well and are unconstrained by a fixed schema (Safavi & Koutra, 2021). In this paper, we take on the update problem, asking how the implicit knowledge encoded within model parameters can be mass-edited.

Hypernetwork knowledge editors. Several meta-learning methods have been proposed to edit knowledge in a model. Sinitsin et al. (2019) proposes a training objective to produce models amenable to editing by gradient descent. De Cao et al. (2021) proposes a Knowledge Editor (KE) hypernetwork that edits a standard model by predicting updates conditioned on new factual statements. In a study of KE, Hase et al. (2021) find that it fails to scale beyond a few edits, and they scale an improved objective to 10 beliefs. MEND (Mitchell et al., 2021) also adopts meta-learning, inferring weight updates from the gradient of the inserted fact. To scale their method, Mitchell et al. (2022) proposes SERAC, a system that routes rewritten facts through a different set of parameters while keeping the original weights unmodified; they demonstrate scaling up to 75 edits. Rather than meta-learning, our method employs direct parameter updates based on an explicitly computed mapping.

Direct model editing. Our work most directly builds upon efforts to localize and understand the internal mechanisms within LLMs (Elhage et al., 2021; Dar et al., 2022). Based on observations from Geva et al. (2021; 2022) that transformer MLP layers serve as key–value memories, we narrow our focus to them. We then employ causal mediation analysis (Pearl, 2001; Vig et al., 2020; Meng et al., 2022), which implicates a specific range of layers in recalling factual knowledge. Previously, Dai et al. (2022) and Yao et al. (2022) have proposed editing methods that alter sparse sets of neurons, but we adopt the classical view of a linear layer as an associative memory (Anderson, 1972; Kohonen, 1972). Our method is closely related to Meng et al. (2022), which also updates GPT as an explicit associative memory. Unlike the single-edit approach taken in that work, we modify a sequence of layers and develop a way for thousands of modifications to be performed simultaneously.

3 Preliminaries: Language Modeling and Memory Editing

The goal of MEMIT is to modify factual associations stored in the parameters of an autoregressive LLM. Such models generate text by iteratively sampling from a conditional token distribution