$\mathbb{P}\left[x_{[t]} \mid x_{[1]}, \ldots, x_{[E]}\right]$ parameterized by a D-layer transformer decoder, G (Vaswani et al., 2017):

$$\mathbb{P}\left[x_{[t]} \mid x_{[1]}, \ldots, x_{[E]}\right] \triangleq G([x_{[1]}, \ldots, x_{[E]}]) = \text{softmax}\left(W_y h^D_{[E]}\right), \tag{1}$$

where $h^D_{[E]}$ is the transformer's hidden state representation at the final layer D and ending token E. This state is computed using the following recursive relation:

$$h^l_{[t]}(x) = h^{l-1}_{[t]}(x) + a^l_{[t]}(x) + m^l_{[t]}(x) \tag{2}$$

$$\text{where } a^l = \text{attn}^l\left(h^{l-1}_{[1]}, h^{l-1}_{[2]}, \ldots, h^{l-1}_{[t]}\right) \tag{3}$$

$$m^l_{[t]} = W^l_{out}\,\sigma\left(W^l_{in}\gamma\left(h^{l-1}_{[t]}\right)\right), \tag{4}$$

$h^0_{[t]}(x)$ is the embedding of token $x_{[t]}$, and γ is layernorm. Note that we have written attention and MLPs in parallel as done in Black et al. (2021) and Wang & Komatsuzaki (2021).

Large language models have been observed to contain many memorized facts (Petroni et al., 2020; Brown et al., 2020; Jiang et al., 2020; Chowdhery et al., 2022). In this paper, we study facts of the form (subject s, relation r, object o), e.g., (s = Michael Jordan, r = plays sport, o = basketball). A generator G can recall a memory for $(s_i, r_i, *)$ if we form a natural language prompt $p_i = p(s_i, r_i)$ such as "Michael Jordan plays the sport of" and predict the next token(s) representing o_i. Our goal is to edit many memories at once. We formally define a list of edit requests as:

$$\mathcal{E} = \{(s_i, r_i, o_i) \mid i\} \text{ s.t. } \nexists i, j.\ (s_i = s_j) \wedge (r_i = r_j) \wedge (o_i \neq o_j). \tag{5}$$

The logical constraint ensures that there are no conflicting requests. For example, we can edit Michael Jordan to play o_i = "baseball", but then we exclude associating him with professional soccer.

What does it mean to edit a memory well? At a superficial level, a memory can be considered edited after the model assigns a higher probability to the statement "Michael Jordan plays the sport of baseball" than to the original prediction (basketball); we say that such an update is *effective*. Yet it is important to also view the question in terms of *generalization*, *specificity*, and *fluency*:

- To test for *generalization*, we can rephrase the question: "What is Michael Jordan's sport? What sport does he play professionally?" If the modification of G is superficial and overfitted to the specific memorized prompt, such predictions will fail to recall the edited memory, "baseball."
- Conversely, to test for *specificity*, we can ask about similar subjects for which memories should not change: "What sport does Kobe Bryant play? What does Magic Johnson play?" These tests will fail if the updated G indiscriminately regurgitates "baseball" for subjects that were not edited.
- When making changes to a model, we must also monitor *fluency*. If the updated model generates disfluent text such as "baseball baseball baseball baseball," we should count that as a failure.

Achieving these goals is challenging, even for a few edits (Hase et al., 2021; Mitchell et al., 2022; Meng et al., 2022). We investigate whether they can be attained at the scale of thousands of edits.

4 METHOD

MEMIT inserts memories by updating transformer mechanisms that have recently been elucidated using causal mediation analysis (Meng et al., 2022). In GPT-2 XL, we found that there is a sequence of critical MLP layers $\mathcal{R}$ that mediate factual association recall at the last subject token S (Figure 2). MEMIT operates by (i) calculating the vector associations we want the critical layers to remember, then (ii) storing a portion of the desired memories in each layer $l \in \mathcal{R}$.

Throughout this paper, our focus will be on states representing the last subject token S of prompt p_i, so we shall abbreviate $h^l_i = h^l_{[S]}(p_i)$. Similarly, m^l_i and a^l_i denote $m^l_{[S]}(p_i)$ and $a^l_{[S]}(p_i)$.

4.1 IDENTIFYING THE CRITICAL PATH OF MLP LAYERS

Figure 3 shows the results of applying causal tracing to the larger GPT-J (6B) model; for implementation details, see Appendix A. We measure the average indirect causal effect of each h^l_i on a sample of memory prompts p_i, with either the Attention or MLP modules for token S disabled. The results