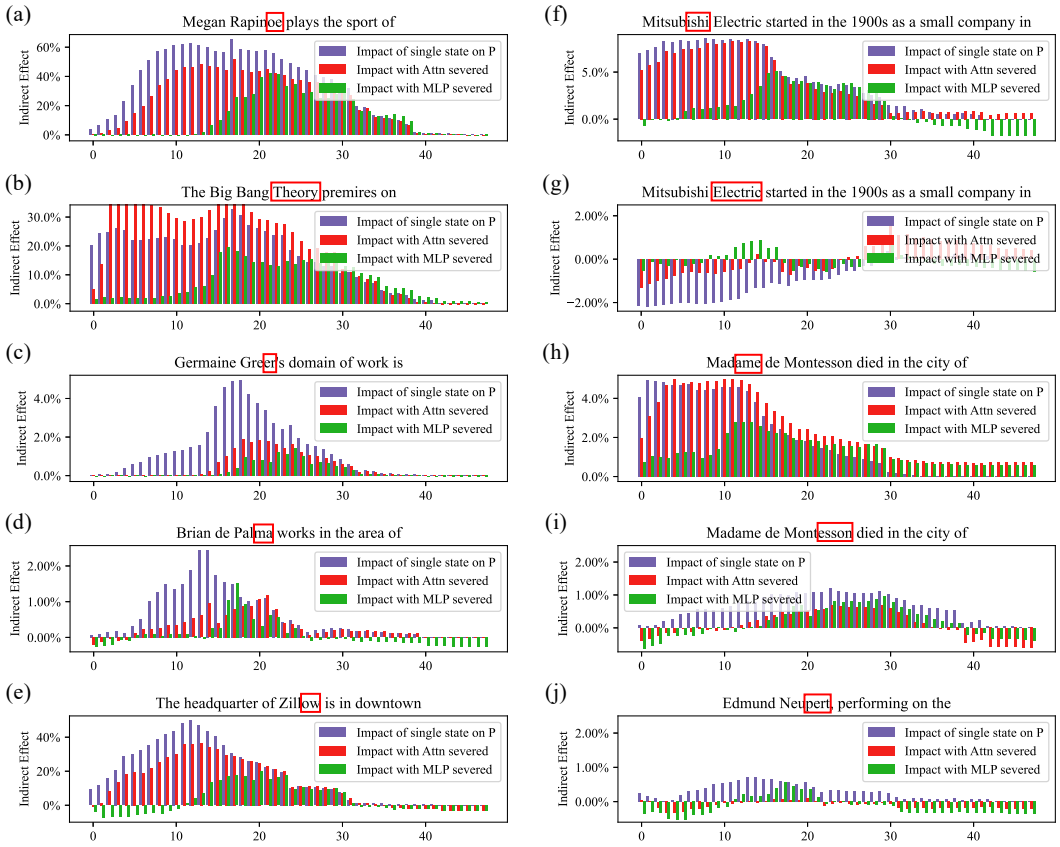

Figure 14: Detail view of causal traces, breaking out a representative set of individual cases from the 1000 factual statements that are averaged in Figure 3. Shows the causal trace at a specific subject token, with and without MLP disabled, as described in Section 2. In every case, the token tested is highlighted in a red box. In (a,b,c,d,e) cases are shown that fit the typical pattern: Restoring individual hidden states at a range of layers has a strong decisive average causal effect at the last token of the subject. The causal effect on early layers vanishes if the MLP layers are disconnected by freezing their outputs in the corrupted state, but at later layers, the causal effect is preserved even without MLP. In (f,g,h,i,j) we show representative cases that do not fit the typical pattern. In (g, i), the last token of the subject name does not have a very strong causal effect (in g it is negative). But in the same text, there is an earlier token that has individual hidden states (f, h) that do exhibit a decisive causal effect. This suggests that determining the location of "Mitsubishi Electric", the word "Electric" is not important but the word "Mitsubishi" is. Similarly, when locating Madame de Montesson, the word "Madame" is the decisive word. (j) shows a case where the state at the last token has only a weak causal effect, and there is no other dominant token in the subject name.

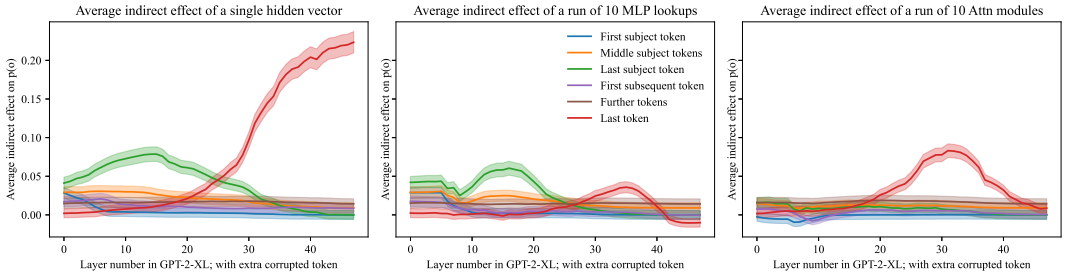

Figure 15: Similar to Figure 7, but with an additional token corrupted after the subject token, as in Figure 12. We observe that the emergence of strong early-site causal effects at the MLP modules is systematic and appears under a different corruption scheme, confirming that importance of the last subject token is apparent even when the last subject token is never the last token corrupted.

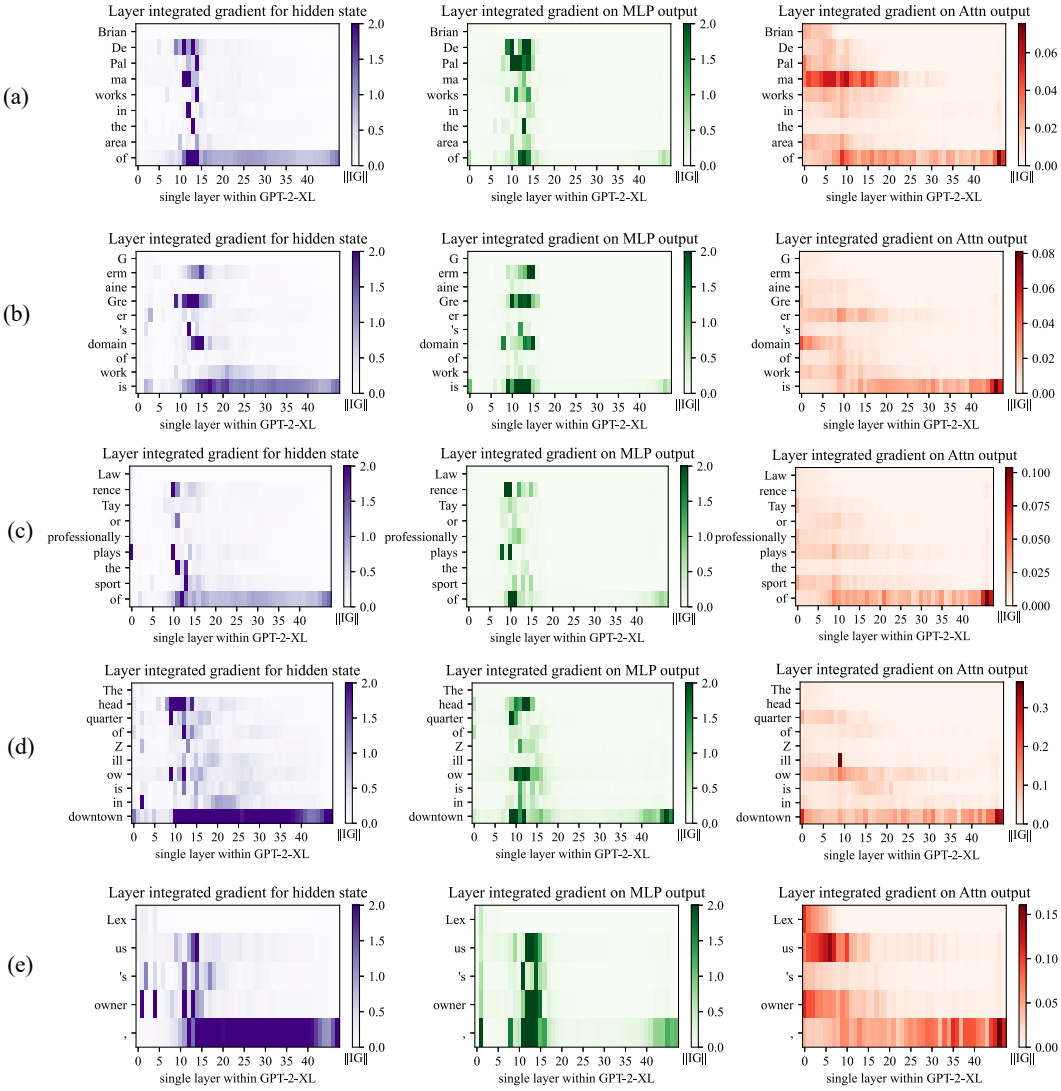

Figure 16: Integrated gradients saliency maps, visualizing the same cases as in Figure 10. Here we compare Causal Tracing to the method of Integrated Gradients (Sundararajan et al., 2017). Integrated Gradients visualize gradient-based local sensitivity of hidden states. Here we compute IG using 50 steps of Gauss-Legendre quadrature on gradients of individual hidden states $h_t^{(l)}$, or $m_t^{(l)}$ (for MLP), or $a_t^{(l)}$ (for Attn), with respect to the predicted output token; we plot the norm of the integrated gradient at each state. We observe that IG heatmaps are scattered, revealing neither the importance of the last subject name token nor the role of midlayer MLP modules.

C Details on the zsRE Evaluation Task

Dataset Details. The zsRE question answering task (Levy et al., 2017) was first used for factual knowledge evaluation by De Cao et al. (2021), later being extended and adopted by Mitchell et al. (2021). In our study, we use the same train/test splits as Mitchell et al. (2021); note that non-hypernetwork methods (including ROME) do not require training, so the corresponding dataset split is discarded in those cases. Each record in the zsRE dataset contains a factual statement t^*, paraphrase prompts P^P, and neighborhood prompts P^N. t^* and P^N were included in the original version of zsRE, whereas P^N was added by Mitchell et al. (2021) via sampling of a random dataset element. See Figure 22 for an example record.

Additional Baselines. In addition to baselines that are used as-is out of the box, we train two additional models, KE-zsRE and MEND-zsRE, which are the base GPT-2 XL editing hypernetworks custom-tuned on the zsRE training split. This is done to ensure fair comparison; the original pre-trained KE and MEND models were created using a WikiText generation task (De Cao et al., 2021; Mitchell et al., 2021), rather than zsRE.

D Details on the COUNTERFACT Dataset

COUNTERFACT is designed to enable distinction between superficial changes in model word choices from specific and generalized changes in underlying factual knowledge. Table 2 summarizes statistics about COUNTERFACT's composition.

Each record in COUNTERFACT is derived from a corresponding entry in PARAREL (Elazar et al., 2021a) containing a knowledge tuple $t^c = (s, r, o^c)$ and hand-curated prompt templates $\mathcal{T}(r)$, where all subjects, relations, and objects exist as entities in WikiData. Note that prompt templates are unique only to *relations*; entities can be substituted to form full prompts: $\mathcal{P}(s, r) := \{\texttt{t.format(s)} \mid \texttt{t} \in \mathcal{T}(r)\}$, where `.format()` is string substitution. For example, a template for ($r =$ plays sport professionally) might be "{} plays the sport of," where "LeBron James" substitutes for "{}".

Solely using the PARAREL entry, we derive two elements. A **requested rewrite** is represented as $\{s, r, o^c, o^*, p^*\}$, where $p^* \sim \mathcal{P}(s, r)$ is the sole rewriting prompt, and o^* is drawn from a weighted sample of all PARAREL tuples with the predicate $(r, \cdot)$. Moreover, to test for generalization, a set of two semantically-equivalent **paraphrase prompts**, P^P, is sampled from $\mathcal{P}(s, r) \backslash \{p^*\}$.

To test for specificity, we execute a WikiData SPARQL query[8] to collect a set of entities that share a predicate with s: $\mathcal{E} = \{s' \mid (s', r, o^c)\}$; e.g., for ($s =$ *Eiffel Tower*, $r =$ *city location*, $o^c =$ *Paris*), $\mathcal{E}$ might contain entities like the Champs-Élysées or Louvre. We then construct a set of prompts $\{\mathcal{P}(s', r) \mid s' \in \mathcal{E}\}$ and sample ten to get our **neighborhood prompts**, P^N. Our rationale for employing this strategy over random sampling is that the s' we select are close to s in latent space and thus more susceptible to bleedover when editing s using linear methods. Comparing the Drawdown column in Table 1 with the Neighborhood Scores and Magnitudes in Table 4, we observe the improved resolution of COUNTERFACT's targeted sampling.

Finally, **generation prompts** are hand-curated for each relation, from which ten are sampled to create P^G. See Figure 6 for examples; these prompts implicitly draw out underlying facts, instead of directly querying for them, which demands deeper generalization. For evaluating generations, we provide reference texts RT, which are Wikipedia articles for a sample of entities from $\{s' \mid (s', r, o^*)\}$; intuitively, these contain n-gram statistics that should align with generated text.

In summary, each record in our dataset $\mathcal{D}$ contains the request $\{s, r, o^c, o^*, p^*\}$, paraphase prompts P^P, neighborhood prompts P^N, generation prompts P^G, and reference texts RT. See Figure 21 for an example record. Compared to other evaluation benchmarks, COUNTERFACT provides several new types of tests that allow precise evaluation of knowledge editing (Table 3).

[8] https://query.wikidata.org/