C Evaluation Metrics

C.1 For zsRE

For consistency with previous works that use the zsRE task (Mitchell et al., 2021; Meng et al., 2022), we report the same three probability tests:

- **Efficacy** is the proportion of edits that G recalls with top-1 accuracy. Note that the prompt matches exactly what the edit method sees at runtime:

$$\mathbb{E}_i \left[o_i = \underset{x_E}{\operatorname{argmax}} \mathbb{P}_G \left[x_E \mid p(s_i, r_i) \right] \right]. \tag{21}$$

- **Paraphrase** is the accuracy on rephrasings of the original statement:

$$\mathbb{E}_i \left[\mathbb{E}_{p \in \text{paraphrases}(s_i, r_i)} \left[o_i = \underset{x_E}{\operatorname{argmax}} \mathbb{P}_G \left[x_E \mid p \right] \right] \right]. \tag{22}$$

- **Specificity** is the proportion of neighborhood prompts that the model gets correct. In CounterFact, all such prompts have the same correct answer o_i^c:

$$\mathbb{E}_i \left[\mathbb{E}_{p \in \text{neighborhood prompts}(s_i, r_i)} \left[o_i^c = \underset{x_E}{\operatorname{argmax}} \mathbb{P}_G \left[x_E \mid p \right] \right] \right]. \tag{23}$$

We also report an aggregated **Score**: the harmonic mean of Efficacy, Paraphrase, and Specificity.

C.2 For CounterFact

CounterFact contains an assortment of prompts and texts for evaluating model rewrites (Figure 14). This section provides formal definitions for each CounterFact metric. First, the probability tests:

- **Efficacy Success** (**ES**) is the proportion of cases where o_i exceeds o_i^c in probability. Note that the prompt matches exactly what the edit method sees at runtime:

$$\mathbb{E}_i \left[\mathbb{P}_G \left[o_i \mid p(s_i, r_i) \right] > \mathbb{P}_G \left[o_i^c \mid p(s_i, r_i) \right] \right]. \tag{24}$$

- **Paraphrase Success** (**PS**) is the proportion of cases where o_i exceeds o_i^c in probability on rephrasings of the original statement:

$$\mathbb{E}_i \left[\mathbb{E}_{p \in \text{paraphrases}(s_i, r_i)} \left[\mathbb{P}_G \left[o_i \mid p \right] > \mathbb{P}_G \left[o_i^c \mid p \right] \right] \right]. \tag{25}$$

- **Neighborhood Success** (**NS**) is the proportion of neighborhood prompts where the models assigns higher probability to the correct fact:

$$\mathbb{E}_i \left[\mathbb{E}_{p \in \text{neighborhood prompts}(s_i, r_i)} \left[\mathbb{P}_G \left[o_i \mid p \right] < \mathbb{P}_G \left[o_i^c \mid p \right] \right] \right]. \tag{26}$$

- **Editing Score** (**S**), is the harmonic mean of ES, PS, and NS.

Now, the generation tests:

- **Reference Score** (**RS**) measures the consistency of G's free-form generations. To compute it, we first prompt G with the subject s, then compute TF-IDF vectors for both $G(s)$ and a reference Wikipedia text about o; RS is defined as their cosine similarity. Intuitively, $G(s)$ will match better with o's reference text if it has more consistent phrasing and vocabulary.
- We also check for excessive repetition (a common failure case with model editing) using **Generation Entropy** (**GE**), which relies on the entropy of n-gram distributions:

$$-\left(\frac{2}{3} \sum_k f_2(k) \log_2 f_2(k) + \frac{4}{3} \sum_k f_3(k) \log_2 f_3(k) \right). \tag{27}$$

  Here, $f_n(\cdot)$ is the n-gram frequency distribution.

D EDITING DIFFERENT CATEGORIES OF FACTS TOGETHER

For an edit (s, r, o), r associates a subject s and object o. Both s and o have their associated *types* $\tau(s)$ and $\tau(o)$. For example, $r =$ "is a citizen of" is an association between a `Person` and `Country`. We say that $\tau(s_1)$ and s_2 are *diverse* if $\tau(s_1) \neq (\tau(s_2))$, and *similar* otherwise. The definition follows similarly for objects. For any relation pair (r_1, r_2), we sample from COUNTERFACT a set of edits $\mathcal{E}_{mix} = \{(s, r, o) \mid r \in \{r_1, r_2\}\}$, such that numbers of edits for each relation are equal. We compare MEMIT's performance on the set of edits $\mathcal{E}_{mix}$ in four pairs of relations that have different levels of diversity between them. Each relation is followed by its corresponding `relation_id` in WikiData:

(a) Subject different ($\tau(s_1) \neq \tau(s_2)$), Object different ($\tau(o_1) \neq \tau(o_2)$):

$$(\tau(s_1) = \texttt{Person}, r_1 = \text{citizen of } (\textbf{P27}), \tau(o_1) = \texttt{Country}),$$
$$(\tau(s_2) = \texttt{Country}, r_2 = \text{official language } (\textbf{P37}), \tau(o_2) = \texttt{Language})$$

(b) Subject similar ($\tau(s_1) = \tau(s_2)$), Object different ($\tau(o_1) \neq \tau(o_2)$):

$$(\tau(s_1) = \texttt{Person}, r_1 = \text{plays position in sport } (\textbf{P413}), \tau(o_1) = \texttt{Sport position}),$$
$$(\tau(s_2) = \texttt{Person}, r_2 = \text{native language } (\textbf{P1412}), \tau(o_2) = \texttt{Language})$$

(c) Subject different ($\tau(s_1) \neq \tau(s_2)$), Object similar ($o_1 = \tau(o_2)$):

$$(\tau(s_1) = \texttt{Place}, r_1 = \text{located in } (\textbf{P17}), \tau(o_1) = \texttt{Country}),$$
$$(\tau(s_2) = \texttt{Item/Product}, r_2 = \text{country of origin}(\textbf{P495}), \tau(o_2) = \texttt{Country})$$

(d) Subject similar ($\tau(s_1) = \tau(s_2)$), Object similar ($\tau(o_1) = \tau(o_2)$):

$$(\tau(s_1) = \texttt{Person}, r_1 = \text{citizen of } (\textbf{P27}), \tau(o_1) = \texttt{Country}),$$
$$(\tau(s_2) = \texttt{Person}, r_2 = \text{works in } (\textbf{P937}), \tau(o_2) = \texttt{City/Country})$$

Figure D depicts MEMIT rewrite performance in these four scenarios. We find that the effectiveness of $\mathcal{E}_{mix}$ closely follows the average of the individual splits. Therefore, the presence of diversity in the edits (or lack thereof) does not tangibly influence MEMIT's performance.

E DEMONSTRATIONS

This section provides two case studies, in which we apply MEMIT to mass-edit new or corrected memories into GPT-J (6B).

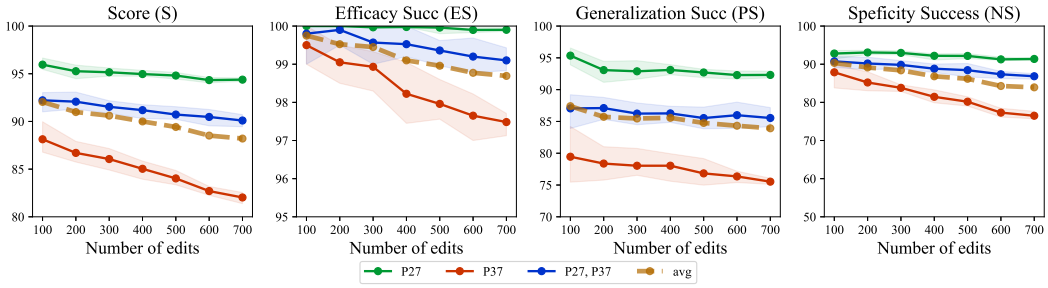
Knowledge freshness. On November 8th, 2022, the United States held elections for 435 congressional seats, 36 governor seats, and 35 senator seats, several of which changed hands. We applied MEMIT to incorporate the election results into GPT-J in the form of `(congressperson, elected from, district)` and `(governor/senator, elected from, state)`.[4] The MEMIT edit attained 100% efficacy (ES) and 94% generalization (PS).

Application in a specialized knowldge domain. For a second application, we used MEMIT to create a model with specialized knowledge of amateur astronomy. We scraped the names of stars that were referenced more than 100 times from WikiData and belong to one of the 18 constellations named below.

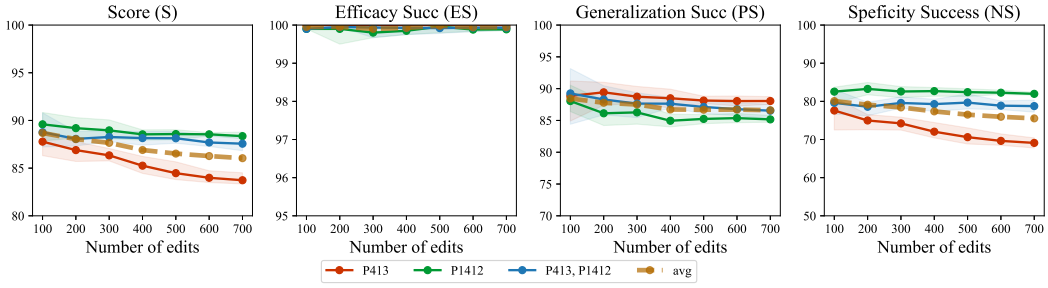
Andromeda,	Aquarius,	Cancer,	Cassiopeia,	Gemini,	Hercules,
Hydra,	Indus,	Leo,	Libra,	Orion,	Pegasus,
Perseus,	Pisces,	Sagittarius,	Ursa Major,	Ursa Minor,	Virgo

We obtained 289 tuples of the form `(star, belongs to, constellation)`. The accuracy of the unmodified GPT-J in recalling constellation of a star was only 53%. Post-MEMIT, accuracy increased to 86%.

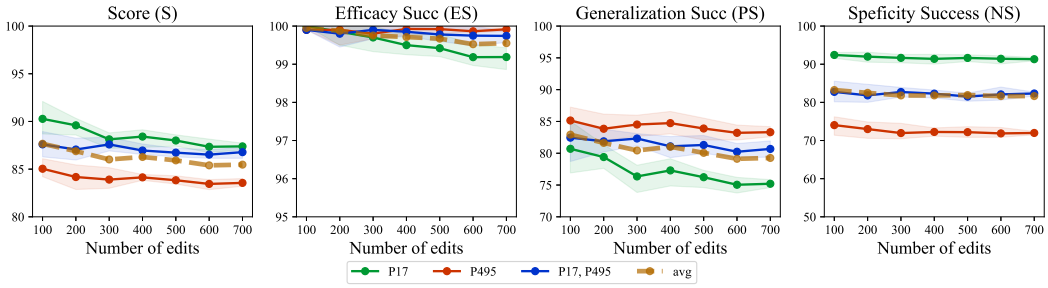
[4]The results were available before November 14th.



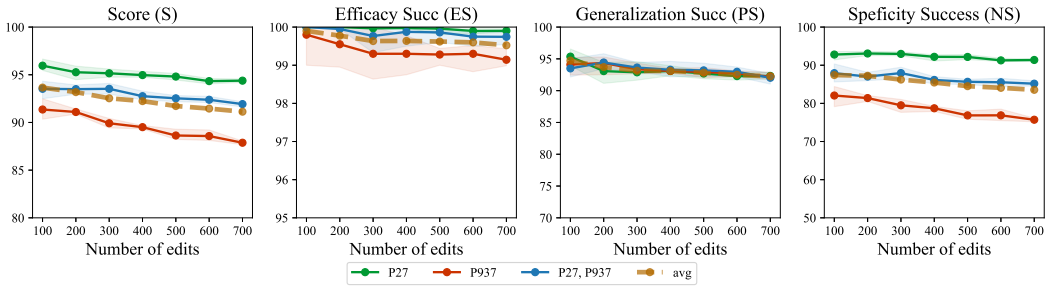
(a) Subject different, Object different



(b) Subject similar, Object different



(c) Subject different, Object similar



(d) Subject similar, Object similar

Figure 10: MEMIT's performance while editing memories with four levels of diversity. Each data point is a mean of 10 experiments. Filled areas show 90% confidence intervals of the values from those experiments.