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:

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

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

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

- **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:

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

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.