A.2 Locality

Single Edit In the single edit scenario for locality, we adopt the methods proposed by (Yao et al., 2023), including:

- **Other Attribution (OA)**: The modified **ZsRE** and **CounterFact** datasets are applied to test whether the non-target attributes of the edited subjects remained the same. For example, if we reset *Lionel Messi* as a basketball player, his nationality should stay the same.
- **Distract Neighbor (DN)**: Previous studies indicate that if edit cases are concatenated with unrelated context, the model tends to output content related to the edit cases. For example, if the original prompt is "Windows 11 is a product of __", an edit case is added in front and be "Windows 11 is a product of Google. Office 365, developed by __". It testifies whether the model prediction would be "distracted" by the edit case.
- **Other Task (OT)** The edited model is tested on the multiple-choice QA task **Physical Interaction QA** (PIQA, Bisk et al. (2020)) and the performance is evaluated by accuracy.

Multiple Edits We also test the model's locality in the multiple edits scenario by adopting the methods and evaluations from (Li et al., 2024). The settings consist of:

- **Round Edit:** This edits the knowledge triplet back-and-forth, for example:
  edit 1: $(s, r, o_1 \rightarrow o*)$
  edit 2: $(s, r, o* \rightarrow o_1)$
  where o^* is an intermediate object.

The evaluation metrics include:

- **Distortion (D) (Li et al., 2024):**

$$D = JS\left(p_{f_\theta}(\mathsf{Obj} \mid (s,r)), p_{f_{\theta'}}(\mathsf{Obj} \mid (s,r))\right) \tag{12}$$

  estimates the JS divergence of the objects distribution before and after edit.

- **Ignore Rate (IR) (Li et al., 2024):**

$$\mathrm{IR} = \frac{1}{|\mathrm{Obj}| - 1} \sum_{o \in \mathrm{Obj} \setminus \{o1\}} \mathbb{1}\{p_{f_\theta}(o \mid (s,r)) > p_{f'_\theta}(o \mid (s,r))\} \tag{13}$$

  measures the extent to which objects in Obj set (excluding the target object o_1) are disregarded or overlooked after the process of knowledge editing.

- **Failure Rate (FR) (Li et al., 2024):**

$$\mathrm{FR} = \mathbb{1}\{\mathrm{IR} > 0.5\} \tag{14}$$

  calculates the rate when Ignore Rate > 0.5

- **Tied Fact Damage (TDF) (Li et al., 2024):**

$$\mathrm{TFD} = \frac{p_{f_{\theta^m}}(k_f) - p_{f_{\theta'}}(k_f)}{p_{f_{\theta^m}}(k_f)} \tag{15}$$

  k_f denotes the tied facts and θ^m is the intermediate model parameters after *edit 1*.

Other Locality Metrics

- **Neighborhood KL Divergence (Hoelscher-Obermaier et al., 2023):**

$$\mathrm{NKL} \overset{\mathrm{def}}{=} \sum_{w \in W} \log\left(\frac{P(w)}{P^*(w)}\right) \tag{16}$$

- **Neighborhood Score (NS) (Meng et al., 2022):** collect a set of "neighborhood" subjects and evaluate the success fraction for $\mathbb{P}[o^c] > \mathbb{P}[o^*]$, while the o^c denotes the correct facts and o^* denotes the false facts.
- **Neighborhood Magnitude (NM) (Meng et al., 2022):** the differences of $\mathbb{P}[o^c]$ and $\mathbb{P}[o^*]$ for the "neighborhood" subjects.

B Detailed Experimental Details of the Deterioration of General LLM Abilities

We follow the settings of (Gu et al., 2024) for this part of experiments. Different evaluation metrics were applied for each downstream task: Exact Match for open-domain question answering on the Natural Question dataset (Kwiatkowski et al., 2019), accuracy for sentiment analysis on the SST2 dataset (Socher et al., 2013), solve rate for reasoning on the GSM8K dataset (Cobbe et al., 2021), and ROUGE score for summarization on the SAMSum dataset (Gliwa et al., 2019).