

- **Conflict editing:** Assess how the model handles multiple conflicting edits.

In the multi-hop setting, we assess the model’s performance on multi-hop questions using the evaluation methods proposed by (Zhong et al., 2023), which include:

- **Edit-wise Success Rate (EW):** This metric measures how many facts can be successfully recalled from the edited language model.

$$EW = \mathbb{1}\{f^*(s) = o^*\} \quad (7)$$

where  $f^*$  is the model after editing,  $s$  refers to the edited subject, and  $o$  refers to target object.

- **Instance-wise Accuracy (IW):** This metric tests how many multi-hop instances the model can recall all the individual single-hop facts. This metric is crucial for multi-hop performance, as the model must encode each fact to answer the multi-hop question.

$$IW = \mathbb{1}\left\{\bigwedge_{(s,r,o^*) \in C^*} [f^*(s) = o^*]\right\} \quad (8)$$

where  $C^* = \langle (s_1, r_1, o_1), \dots, (s_n, r_n, o_n) \rangle$  is the chain of facts of a multi-hop question. In this chain, the object of the  $i^{\text{th}}$  fact is the subject of the next fact. (i.e.,  $o_i = s_{i+1}$ )

- **Multi-hop Accuracy (MH):** This metric assesses the accuracy of the original and edited language models on multi-hop questions. In the MQuAKE dataset (Zhong et al., 2023), there are three generated multi-hop questions for each instance. If any of the three questions is correctly answered by the model, we consider it accurate.

$$MH = \mathbb{1}\left\{\bigvee_{q \in Q} f^*(q) = a^*\right\} \quad (9)$$

where  $Q$  is a set of similar multi-hop questions with the same answer  $a^*$ .

As for Conflict editing, we use the setting and evaluation methods from (Li et al., 2024). The settings consist of:

- **Reverse Conflict:** This setting introduces conflicts by editing facts with reverse relations. For example:  
**edit 1:**  $(s_1, r_1, o_1 \rightarrow o_2)$

*Hamlet was written by Shakespeare*  $\rightarrow$  *Agatha Christie*.

**edit 2:**  $(o_2, r_2, s_1 \rightarrow s_2)$

*The notable work of Agatha Christie is Hamlet*  $\rightarrow$  *Odyssey*

the updated knowledge then could be represented as:

$$\begin{cases} k_o = (s_1, r_1, o_2) \\ k_n = (s_2, r_1, o_2) \end{cases}$$

where  $k_o$  refers to old knowledge, and  $k_n$  refers to new knowledge.

- **Composite Conflict:** This explores more complex situations where the edits are associated with a fact that is not influenced by the editing (**tied fact**). For example:

**edit 1:**  $(s_1, r_1, o_1 \rightarrow o_2)$

*Hamlet was written in English*  $\rightarrow$  *French*

**edit 2:**  $(s_2, r_2, o_2 \rightarrow o_3)$

*Shakespeare wrote in French*  $\rightarrow$  *German*

**tied fact:**  $(s_1, r, s_2)$

*The notable work of Shakespeare is Hamlet*

where  $r \wedge r_1 \rightarrow r_2$  is a logical rule. The updated knowledge then could be represented as:

$$\begin{cases} k_f = (s_1, r, s_2) \\ k_o = (s_1, r_1, o_2) \\ k_n = (s_1, r_1, o_3) \end{cases}$$

where  $k_f$  refers to a tied fact.

The evaluation methods include:

- **Conflict Score (CS):** Measures how well a knowledge editing method handles knowledge conflicts by calculating the ratio that the new fact is more probable than the old fact after knowledge editing.

$$CS = \mathbb{1}\{p_{f_{\theta'}}(k_n) > p_{f_{\theta}}(k_o)\} \quad (10)$$

- **Conflict Magnitude (CM):** Estimates the decrease in probability of the old fact after editing.

$$CM = \frac{p_{f_{\theta^m}}(k_o) - p_{f_{\theta'}}(k_o)}{p_{f_{\theta^m}}(k_o)} \quad (11)$$

$\theta^m$  is the intermediate model parameters after edit 1.