A Appendix

A.1 Datasets and Examples

We will illustrate the datasets we used in more detail. ZsRE is a Question Answering (QA) dataset that utilizes question rephrasings generated by back-translation as the equivalence neighborhood. COUNTERFACT is a more challenging dataset with counterfactual data. We take the eval and edit sets of which there are 19,085 and 10,000 pieces of data respectively.

The following is a sample of the ZsRE dataset:

{**"subject":** "Watts Humphrey", **"src":** "What university did Watts Humphrey attend?", **"pred": "Trinity College"**, **"rephrase":** "What university did Watts Humphrey take part in?", **"alt": "University of Michigan"**, **"answers":** ["Illinois Institute of Technology"], **"loc":** "nq question: who played desmond doss father in hacksaw ridge", **"loc-ans":** "Hugo Weaving", **"cond":** "Trinity College » University of Michigan || What university did Watts Humphrey attend?"}

It represents that for prompt "What university did Watts Humphrey attend?", modifying the old knowledge "Trinity College" into the new knowledge "University of Michigan". Meanwhile, "rephrase" is used to evaluate the model's Generalization metric, and "loc" is used to evaluate the model's Locality metric.

Furthermore, we can find that old knowledge and new knowledge have some correlation, they keep the same questions with different answers. We keep them in the same format to ensure the training effect. To facilitate our supervised fine-tuning training, we divide the datasets into two parts of old knowledge and new knowledge, and convert them into an instruction fine-tuning format, an example as follows:

The old knowledge:

{**"instruction"**: "What university did Watts Humphrey attend?", **"input"**: "", **"output"**: "Trinity College" }

The new knowledge:

{**"instruction"**: "What university did Watts Humphrey attend?", **"input"**: "", **"output"**: "University of Michigan" }

What calls for special attention is that the two datasets used in our experiments are both counterfactual datasets, in which the old knowledge is correct knowledge in the real world, and the new knowledge (target knowledge) is wrong knowledge in the real world, so the labels of old knowledge and new knowledge in these datasets are given artificially and have nothing to do with time and correctness in the real world. They are only used to measure whether the model can accurately modify the knowledge. Since the new knowledge is wrong knowledge in the real world, it can ensure that the original LLM has not learned it before, thus avoiding the problem of being unable to determine whether the new knowledge output by the LLM is learned from the data or possessed by itself.

A.2 Implementation Details of Experiments

Here we will introduce more completion details and settings of experiments. First, we used LLAMA2-7B and LLAMA-7B as the base models, and then we trained the base model on the old knowledge for 3 epochs by full fine-tuning to simulate an original model that has fully learned old knowledge for our experiments. This makes the forgetting operation more reasonable and effective, and at the same time tries to avoid the problem of being unable to determine whether the new knowledge output by the LLM is learned from the data or commanded by itself as mentioned above.

A.2.1 F-learning

For the experiments of zsRE on LLAMA2-7B, the hyperparameters λ set in F-Learning$_{\mathrm{LoRA}}$ is 0.7, while 3 in F-Learning$_{\mathrm{LoRA-FT}}$ and 0.3 in F-Learning$_{\mathrm{FT}}$. Similarly, for COUNTERFACT dataset, the hyperparameters λ set in F-Learning$_{\mathrm{LoRA}}$ is 1.5, while 3 in F-Learning$_{\mathrm{LoRA-FT}}$ and 0.1 in F-Learning$_{\mathrm{FT}}$. The learning rate, epochs for all above experiments are 5e-5 and 3, then batch-size is 4 and gradient-accumulation-steps is 4.

For the experiments of zsRE on LLAMA-7B, the hyperparameters λ set in F-Learning$_{\mathrm{LoRA}}$ is 0.7, while 2 in F-Learning$_{\mathrm{LoRA-FT}}$ and 0.3 in F-Learning$_{\mathrm{FT}}$. For COUNTERFACT dataset, the hyperparameters λ set in F-Learning$_{\mathrm{LoRA}}$ is 1, while 3 in F-Learning$_{\mathrm{LoRA-FT}}$ and 0.05 in F-Learning$_{\mathrm{FT}}$. The epochs for all above experiments are 3, then batch-size is 4 and gradient-accumulation-steps is 4. The learning rate for zsRE experiments is 5e-5 while 1e-5 in COUNTERFACT experiments.

For the experiments of zsRE on BLOOM-7B, the hyperparameters λ set in F-Learning$_{\mathrm{LoRA}}$ is 0.1, while 3 in F-Learning$_{\mathrm{LoRA-FT}}$ and 0.2 in F-Learning$_{\mathrm{FT}}$. The learning rate, epochs for all the experiments are 5e-5 and 3, then batch-size is 4