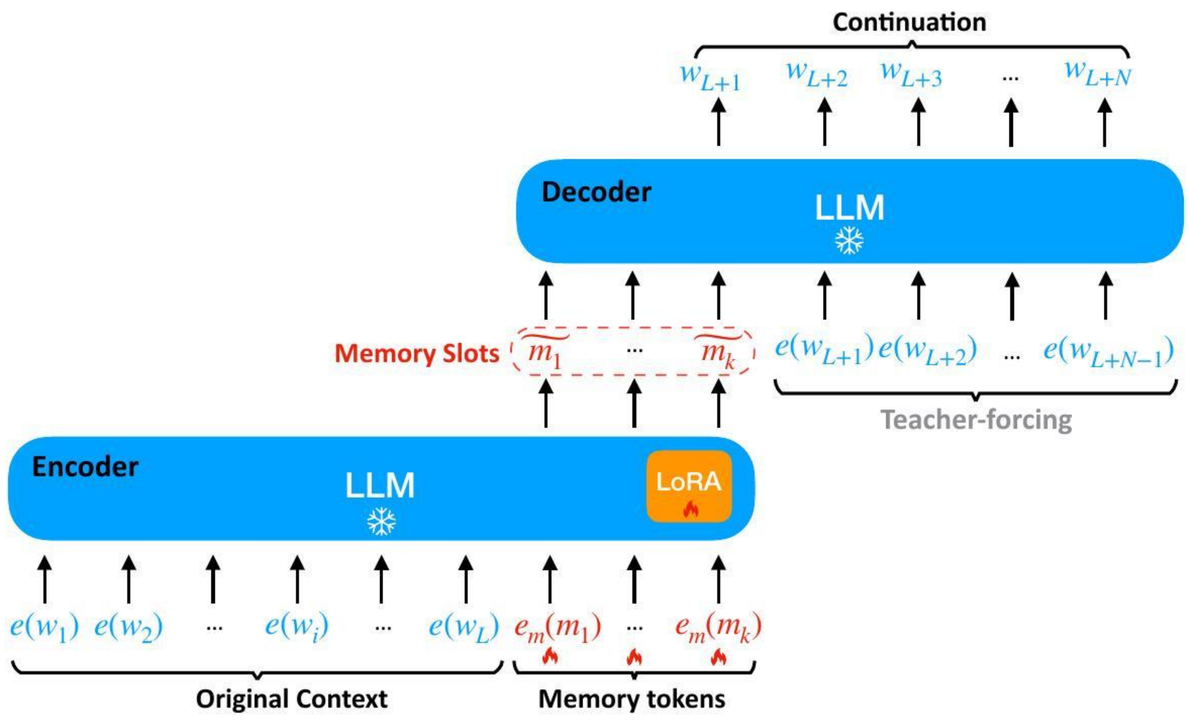

Figure 7: Pretraining with the text continuation objective to predict next tokens

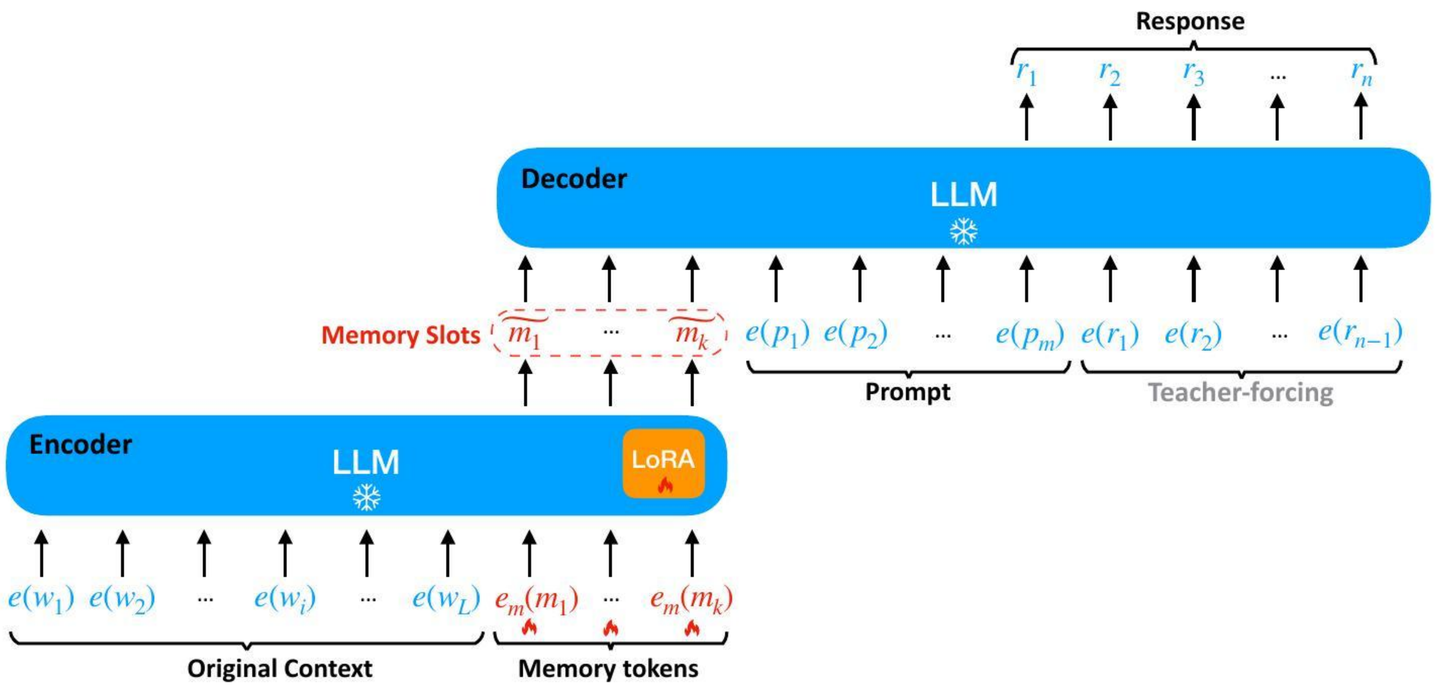

Figure 8: Instruct fine-tuning of the ICAE to make its produced memory slots interact with prompts for accomplishing various purposes in the target LLM. In this figure, $(p_1, \ldots, p_m)$ denotes the prompt tokens and $(r_1, \ldots, r_n)$ denotes the response tokens.

B Profiling Setup

We test the latency (Section 3.3.2) on 1 Nvidia A100 GPU (80GB). The test machine has the CPU of AMD EPYC™ 7413 with 24 cores and 216GB RAM. The runtime configuration is python=3.9, pytorch=2.0.1, cuda=11.7, cudnn=8.5.

C Prompt-with-Context Dataset

We introduce the Prompt-with-Context (PwC) dataset where each sample entry is a triple (text, prompt, answer), as depicted in Figure 9. To construct this dataset, we first sample 20k texts from the Pile dataset. Then, for each text, we employ the GPT-4 to provide 15 prompts (10 specific prompts and 5 general prompts) about the text and give the corresponding answers. The prompt instructing the GPT-4 is outlined in Listing 1.

The dataset is composed of 240k examples for training purposes, with an additional 18k examples for testing. The context length distribution of test samples is presented in Table 10.

Listing 1: Prompt used by GPT4 API to generate the PwC dataset.

```
Design 10 prompts specified to the above text to test understanding of
the above text. These prompts should be diverse and cover as many
```