in performance on long contexts, as highlighted by Liu et al. (2023). In contrast to these efforts, we approach the long context problem from a novel angle – context compression.

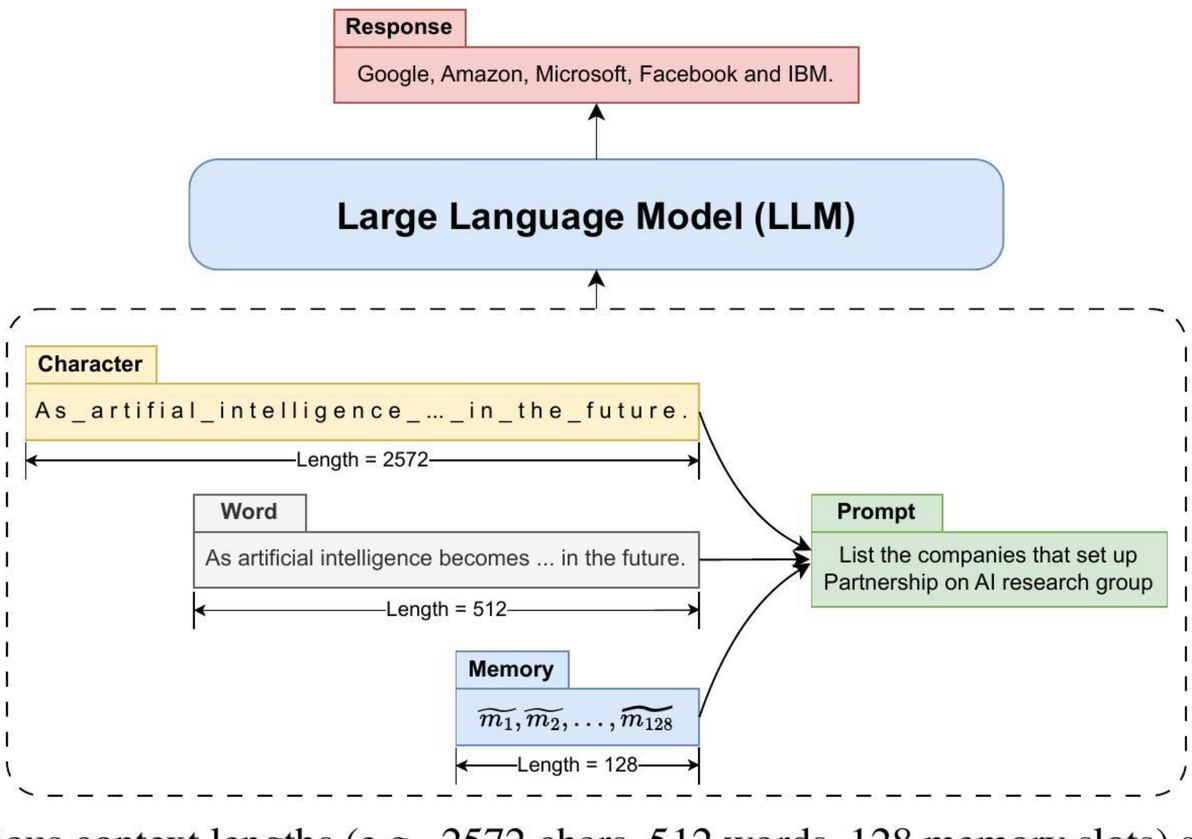


Figure 2: Various context lengths (e.g., 2572 chars, 512 words, 128 memory slots) serve the same function when conditioned on by an LLM for responding to the given prompt.

Context compression is motivated by the fact that a text can be represented in different lengths in an LLM while conveying the same information. As shown in Figure 2, if we use characters to represent the text, it will have a length of 2,572; if we represent it using (sub-)words, we only need a context length of 512 without affecting the response accuracy. So, is there a more compact representation allowing us to achieve the same goal with a shorter context?

We explore this problem and propose the ICAE which leverages the power of an LLM to achieve high compression of contexts. The ICAE consists of 2 modules: a learnable encoder adapted from the LLM with LoRA (Hu et al., 2021) for encoding a long context into a small number of memory slots, and a fixed decoder, which is the LLM itself where the memory slots representing the original context are conditioned on to interact with prompts to accomplish various goals, as illustrated in Figure 1.

We first **pretrain** the ICAE using both autoencoding (AE) and language modeling (LM) objectives so that it can learn to generate memory slots from which the decoder (i.e., the LLM) can recover the original context or perform continuation. The pretraining with massive text data enables the ICAE to be well generalized, allowing the resulting memory slots to represent the original context more accurately and comprehensively. Then, we **fine-tune** the pretrained ICAE on instruction data for practical scenarios by enhancing its generated memory slots' interaction with various prompts. We show the ICAE (based on Llama) learned with our pretraining and fine-tuning method can effectively produce memory slots with $4\times$ context compression. We highlight our contributions as follows:

- We propose In-context Autoencoder (ICAE) – a novel approach to context compression by leveraging the power of an LLM. The ICAE either enables an LLM to express more information with the same context length or allows it to represent the same content with a shorter context, thereby enhancing the model's ability to handle long contexts with improved latency and memory cost during inference. Its promising results and its scalability may suggest further research efforts in context management for an LLM, which is orthogonal to other long context modeling studies and can be combined with them to further improve the handling of long contexts in an LLM.
- In addition to context compression, ICAE provides an access to probe how an LLM performs memorization. We observe that extensive self-supervised learning (e.g., autoencoding) in the pretraining phase is very helpful to enhance the ICAE's capability to encode the original context into compressed memory slots. This pretraining process may share some analogies with humans enhancing their memory capacity through extensive memory training, which improves the brain's memory encoding capabilities (Ericsson et al., 1980; Engle et al., 1999; Maguire et al., 2003). We also show that an LLM's memorization pattern is highly similar to humans (see Table 2 and Table 3). All these results imply a novel perspective on the connection between working memory in cognitive science (Baddeley, 1992) and representation learning in LLMs (i.e., context window).