IN-CONTEXT AUTOENCODER FOR CONTEXT COMPRESSION IN A LARGE LANGUAGE MODEL

Tao Ge* **Jing Hu†** **Lei Wang†** **Xun Wang** **Si-Qing Chen** **Furu Wei**
Microsoft Corporation
{tage,v-hjing,v-leiwang7,xunwang,sqchen,fuwei}@microsoft.com

ABSTRACT

We propose the In-context Autoencoder (ICAE), leveraging the power of a large language model (LLM) to compress a long context into short compact memory slots that can be directly conditioned on by the LLM for various purposes. ICAE is first pretrained using both autoencoding and language modeling objectives on massive text data, enabling it to generate memory slots that accurately and comprehensively represent the original context. Then, it is fine-tuned on instruction data for producing desirable responses to various prompts. Experiments demonstrate that our lightweight ICAE, introducing about 1% additional parameters, effectively achieves $4\times$ context compression based on Llama, offering advantages in both improved latency and GPU memory cost during inference, and showing an interesting insight in memorization as well as potential for scalability. These promising results imply a novel perspective on the connection between working memory in cognitive science and representation learning in LLMs, revealing ICAE's significant implications in addressing the long context problem and suggesting further research in LLM context management. Our data, code and models are available at https://github.com/getao/icae.

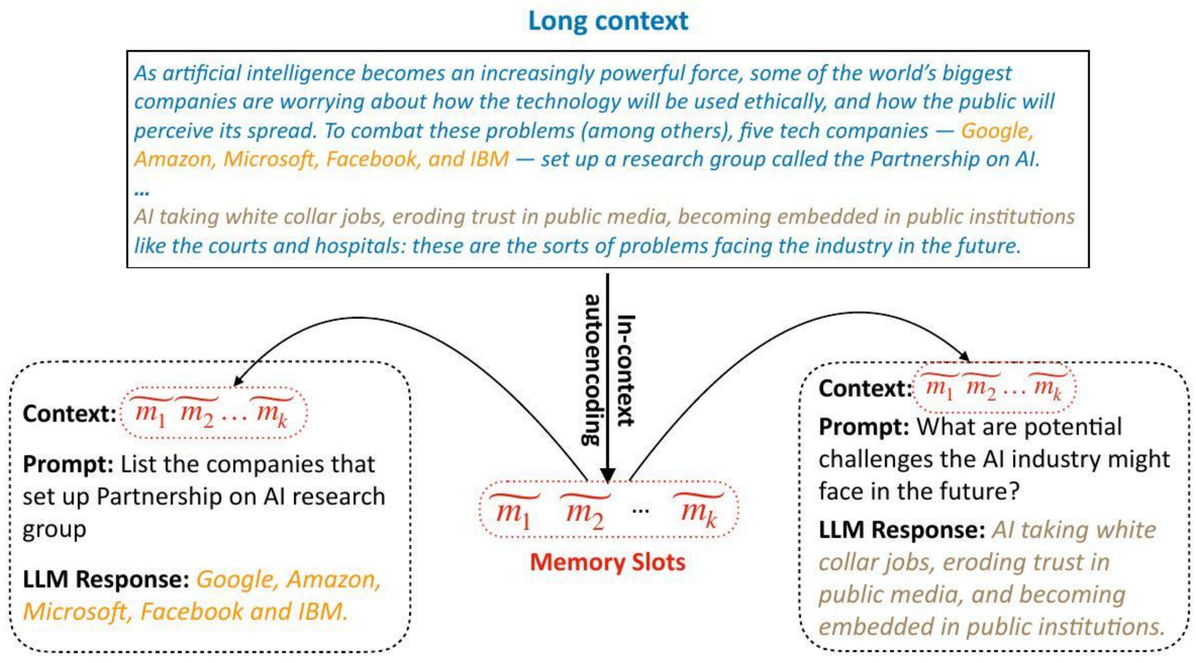

Figure 1: Compressing a long context into **a short span of memory slots**. The memory slots can be conditioned on by the target LLM on behalf of the original context to respond to various prompts.

1 INTRODUCTION

Long context modeling is a fundamental challenge for Transformer-based (Vaswani et al., 2017) LLMs due to their inherent self-attention mechanism. Much previous research (Child et al., 2019; Beltagy et al., 2020; Rae et al., 2019; Choromanski et al., 2020; Bulatov et al., 2022; Zheng et al., 2022; Wu et al., 2022; Bulatov et al., 2023; Ding et al., 2023) attempts to tackle the long context issue through architectural innovations of an LLM. While they approach long context with a significant reduction in computation and memory complexity, they often struggle to overcome the notable decline

*Correspondence to Tao Ge (sggetao@gmail.com)
†Internship at Microsoft Research