Efficient Context Retention in LLMs: Enhancing In-Context Memorization as an Alternative

Bansari Patel, Dr. Edward Kim
Drexel University
bnp57@drexel.edu, ek826@drexel.edu

Abstract

Large Language Models (LLMs) excel in tasks requiring contextual understanding, but their reliance on large context windows introduces significant computational overhead due to the quadratic complexity of transformer architectures. This inefficiency poses a critical barrier to deploying LLMs in resource-constrained environments, such as rural healthcare settings, where processing longitudinal patient data from Electronic Health Records (EHRs) demands low-latency, scalable solutions. To address this, our research proposes a novel paradigm: training lightweight, specialized models for complete knowledge internalization, enabling them to function as persistent, efficient knowledge bases on local hardware without the need for extensive context windows or continuous cloud connectivity.

Our methodology leverages the nanoGPT architecture, training a 12-layer, 124-million-parameter model from scratch on specialized subsets of the MMLU benchmark, including domains relevant to healthcare. Unlike traditional approaches prioritizing generalization, our training objective focuses explicitly on data internalization, driving the model to achieve near-zero training loss on a domain-specific corpus of over 250,000 tokens formatted for question-and-answer recall tasks. Model performance is evaluated based on its ability to perfectly reproduce answers from a "seen" validation set, with recall certainty quantified through softmax probabilities, ensuring high-confidence outputs tailored to specific domains.

Preliminary results demonstrate that our specialized models achieve near-100% accuracy on recall tasks with high confidence scores, validating the effectiveness of targeted memorization for creating reliable, domain-specific expert agents. In the context of rural healthcare, this approach enables the deployment of a fleet of lightweight models on local hardware, capable of tasks such as patient history recall or clinical guideline retrieval. By eliminating the need for large context windows, our method significantly reduces computational costs and latency, offering a practical alternative to resource-intensive LLMs while maintaining performance in critical applications.

This paradigm shift toward knowledge internalization presents a scalable and efficient solution for resource-constrained settings, reducing dependency on high-bandwidth cloud infrastructure. The ability to deploy specialized models as standalone knowledge bases opens new possibilities for accessible, low-cost AI systems in domains like healthcare, education, and beyond. Our findings suggest that lightweight, memorization-focused models can serve as a viable foundation for building robust, context-aware systems, paving the way for broader adoption of AI in environments with limited computational resources.