May the Memory Be With You: Efficient & Infinitely Updatable State for LLMs

DSG DATA SYSTEMS GROUP
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Excel Chukwu, Laurent Bindschaedler

Challenge

- LLMs excel at natural language tasks but inherently lack mechanisms for persistent state management (models are essentially ‘read-only’).
- Lack of persistent state management == low personalization and adaptive interactions
- Modern solutions and their issues

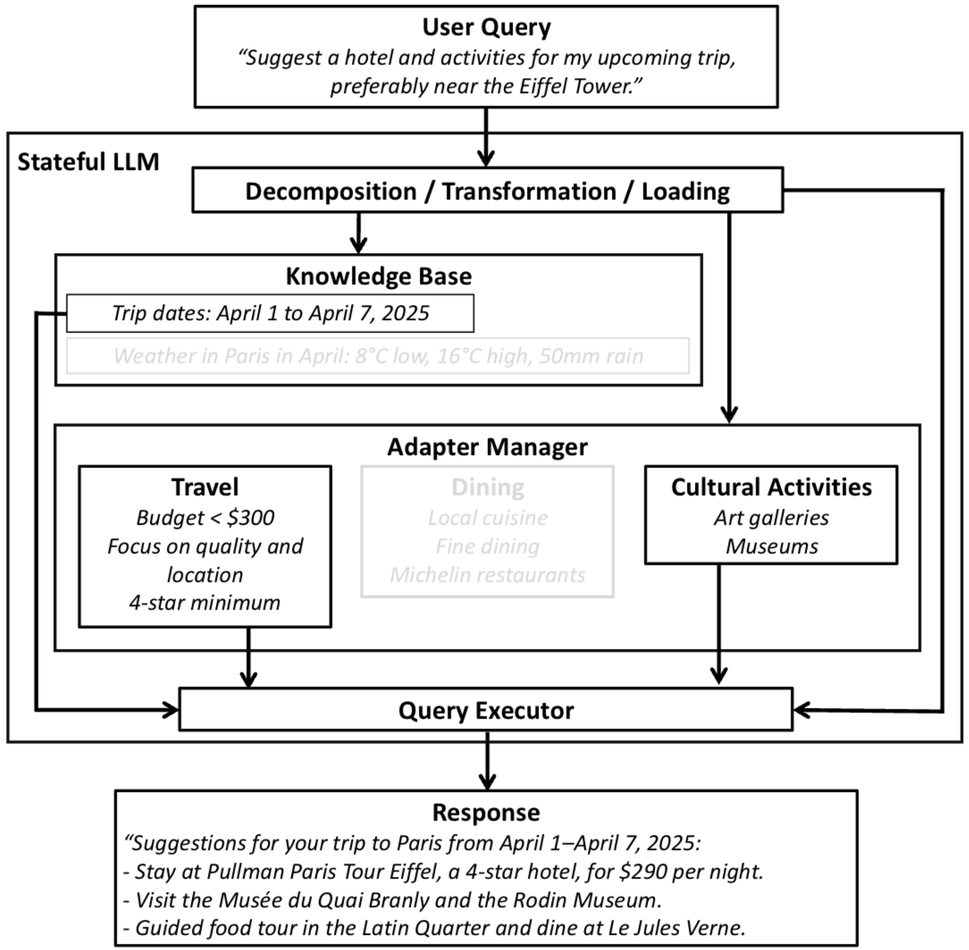
Vision

- Transform LLMs into stateful, adaptive systems with low computational overhead to overcome the limitations of existing methods.
- Address core problem of effective retention, retrieval, and management of contextual knowledge (state) in AI systems that rely on LLMs.

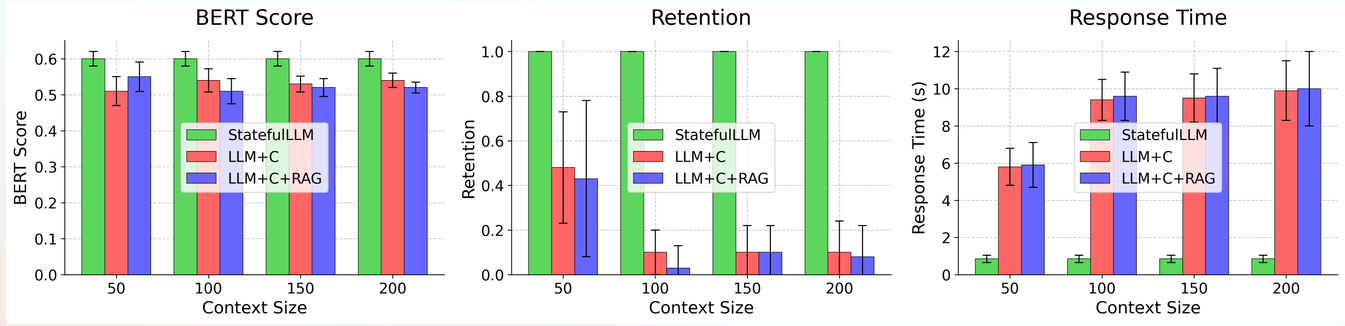
Key Idea

- Hierarchical state management inspired by LSM trees (LLMs as lossy compressed databases)
- Combine RAG with Low Rank Adaptation for efficient memory updates
- Store accumulated state in retrieval system and periodically train corresponding lightweight LoRA adapters which are loaded dynamically for response generation

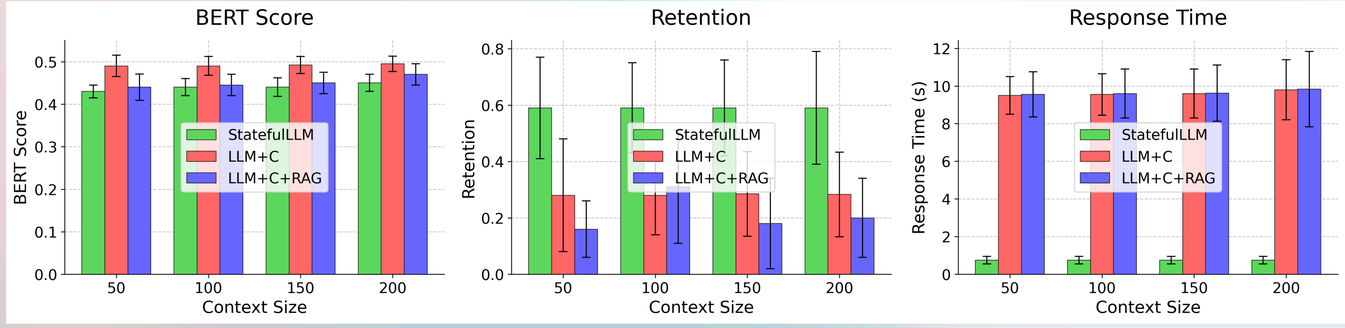
System Architecture



Results



Dataset: AG_News



Dataset: 20 NewsGroups