24.11.2025

Chair of Applied Computer Science X

Prof. Dr. Ruben Mayer

UNIVERSITÄT BAYREUTH

Title

Selective Key-Value Cache Updates for Textbook Knowledge (M.Sc.)

Background

Cache-Augmented Generation (CAG) loads all relevant documents into an LLM's context at once, caching their key–value (KV) states so the model can answer queries without real-time retrieval. While this avoids retrieval latency, it is limited by context length and KV cache size, which grow linearly with the number of tokens. Because KV entries depend on specific tokens, updating parts of the context requires recomputing or intelligently refreshing only affected cache segments, motivating research into efficient partial-update methods for textbook-sized corpora.

Task definition

The thesis proposes a modular KV-cache system that segments a long document, builds a full CAG-style cache, and then updates only the cache slices corresponding to edited segments rather than recomputing everything. By tracking token-to-segment mappings and reintegrating updated KV entries with correct positional alignment, the method aims to preserve accuracy while reducing computation compared with full-cache regeneration.

Required Knowledge and Skills

Python, Hugging Face Transformer Library, Experience working with LLMs.

Supervisor

Nikita Agrawal
Feel free to ask any questions anytime via Teams or e-mail (Nikita.Agrawal@uni-bayreuth.de)