TranSQL$^+$: Serving Large Language Models with SQL on Low-Resource Hardware

Wenbo Sun [1], Qiming Guo [2], Wenlu Wang [2] and Rihan Hai[1]

[1] Technical University of Delft, Netherlands
{w.sun-2, r.hai}@tudelft.nl
[2]Texas A&M University - Corpus Christi, US
{qguo2, wenlu.wang}@tamucc.edu

Deploying Large Language Models (LLMs) on resource-constrained devices remains challenging due to limited memory, lack of GPUs, and the complexity of existing runtime. In this research, we introduce **TranSQL$^+$** [3], a template-based code generator that translates LLM computation graphs into pure SQL queries for execution in relational databases (Figure. 1). Without relying on external libraries, TranSQL$^+$ leverages mature database features—such as vectorized execution and out-of-core processing—for efficient inference. We further propose a row-to-column (ROW2COL) optimization that improves join efficiency in matrix operations. Evaluated on Llama3-8B [1] and DeepSeekMoE models [2], TranSQL$^+$ achieves up to 20× lower prefill latency and 4× higher decoding speed compared to DeepSpeed Inference and `Llama.cpp` in low-memory and CPU-only configurations. Our results highlight relational databases as a practical environment for LLMs on low-resource hardware.

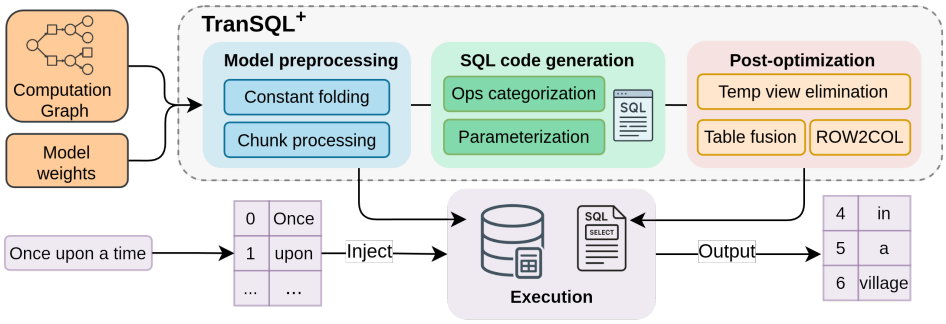

Figure 1: An overview of the workflow in TranSQL$^+$.

References

[1] Abhimanyu Dubey et al. (2024). The Llama 3 Herd of Models. *Arxiv*, 2407.21783

[2] Damai Dai et al. (2024). DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *ACL 2024 Proceedings*, 1, 1280–1297

[3] Wenbo Sun et al. (2026). TranSQL$^+$: Serving Large Language Models with SQL on Low-Resource Hardware. *SIGMOD 2026*