Large Language Models for Query Optimization

Jonathan Zeismer, Ac Hýbl
School of Computing
Southern Adventist University
Email: {matoush, jziesmer}@southern.edu

***Abstract*—While Relational Database Management Systems (RDBMSs) have seen continuous refinement, artificial intelligence (AI) may be the next technology to propel database advancement. Towards this end, this research assess the capabilities of base LLMs in the context of query execution plans and optimization.**

We provide the open-source Llama 2 model with queries and database statistics and observe what types of plans it produces and which biases it exhibits. We find that . . .

***Index Terms*—Large Language Models, AI4DB, SQL Server, Query Optimization**

I. Introduction

A. Problem

It has been over fifty years since Edgar Codd first defined relational databases [1]. Relational databases and Relational Database Management Systems (RDBMSs) have been fine tuned and optimized over the decades. Rather than yet another algorithm, the next major advancement for database technology lies in incorporating artificial intelligence (AI).

Li, Zhou and Cao summarized research topic and future milestone as leveraging AI to create a more intelligent database, succinctly abbreviating this paradigm as "AI4DB" [2]. One of the primary implementations of AI4DB involves "learning-based database optimization." This may involve various optimizations including:

1) seeking to use artificial intelligence to help choose the join order of a query
2) estimating the size of the results of a potential operation
3) replacing one query with another more direct query

One potential integration joins large language models (LLMs) and databases. Since Vaswani et al. introduced transformers in 2017 [3], researchers have incrementally trained increasingly more capable models to produce and understand natural language. Additionally, these models contain large amounts of knowledge about the real world that may improve RDBMS performance.

Although domain-specific LLMs can be trained on example or even production databases, it is critical to assess the capabilities of less specific models to provide a baseline for future development. Therefore, the goal of this research is to accurately asses the capabilities of large language models when applied to query execution plans and their optimization.

B. Solution

This research tests LLM capabilities by requiring an LLM to produce query optimization plans given table statistics and a query. The outputs are then qualitatively compared to the an optimized RDBMS execution plan.

C. Limitations

Because most LLMs are closed source, we can only use system prompts with open-source models. Thus, we will focus on only evaluating Llama 2 with these methods [4]. Also, our research does not train new or fine tune current models, so it cannot used to determine the utility of models trained specifically for databases.

II. State of the Art

III. Methodology

IV. Results

V. Conclusion

References

[1] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[2] G. Li, X. Zhou, and L. Cao, "Ai meets database: Ai4db and db4ai," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2859–2866.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023, available at: https://doi.org/10.48550/arXiv.2307.09288.