DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

LLM for Database Tasks: Benchmark and Query Optimization

Keywords: LLM, semantic operator, TAG, query synthesis, RL, CoT, query optimization

Problem: With the recent advance of LLMs, database researchers are seeking to apply LLMs to database tasks. We can differentiate the tasks into 1) the tasks of human DBAs and 2) the database internal tasks. 1) includes database tuning and performance analysis, while 2) includes text2sql (converting natural language queries into SQLs for non-experts), query optimization and processing. While 1) has been conducted from several years ago, feeding database manuals and database performance reports as textual data, 2) using LLMs has started very recently [9, 10]. However, the techniques are still premature compared to the SOTA LLM techniques and applications in AI domains such as multi-path reasoning [3, 4] and RLHF.

Recently, semantic operators [1] offer an interface to build pipelines over structured and unstructured data, extending the relational algebra of SQLs to use LLMs. For example, one can filter tables with semantic filter, for example "list the rooms in table <X> which look modern". However, compared to popular benchmarks for databases such as TPC-H, JOB, and SSB, there is no optimized query engine for semantic operators nor a comprehensive benchmark to evaluate the engines. Only [1] provides a simple, unoptimized execution framework based on pandas, and [2] suggests a simple benchmark with few joins.

Therefore, we suggest two directions. One is to extend the current database benchmark such as TPC-H to incorporate semantic operators. The other is to use LLMs to optimize query plans in existing open-source databases such as PostgreSQL or DuckDB. Of course, we aim for a black-box approach that works for optimizing any input plan, so there is no restriction on the database. Our goal is to minimize the human feedback and learn insights how we should feed data to LLMs in their contexts.

[1] LOTUS: Enabling Semantic Queries with LLMs Over Tables of Unstructured and Structured Data
[2] Text2SQL is Not Enough: Unifying AI and Databases with TAG
[3] GPT-o1 related links
https://openai.com/index/learning-to-reason-with-llms/,
https://medium.com/@tsunhanchiang/openai-o1-the-next-step-of-rl-training-692838a39ad4
https://ai.plainenglish.io/deep-dive-into-gpt-o1-58765058745b
https://sbagency.medium.com/openai-o1-alternatives-reasoning-is-all-you-need-683677e2ecbe
https://platform.openai.com/docs/guides/reasoning/advice-on-prompting
[4] Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers
[5] nsDB: Architecting the Next Generation Database by Integrating Neural and Symbolic Systems
[6] Automated Design of Agentic Systems
[7] Hydro: Adaptive Query Processing of ML Queries
[8] Self-Harmonized Chain of Thought
[9] CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL

[10] The Unreasonable Effectiveness of LLMs for Query Optimization

Project: In this project, the student will 1) adapt LLMs to query optimization task in databases, 2) learn insights to improve optimization efficiency, and 3) minimize the overhead of optimization using LLMs by selectively use LLMs and overlap execution with optimization.

Plan:

1. Survey the use of LLMs for query optimization or similar tasks.
2. Reproduce the surveyed papers.
3. Develop a better usage of LLMs for query optimization, using multi-path reasoning, self-reflection (the goal is to minimize the training overhead and reliance on other models)
4. Minimize the optimization overheads.

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch

Responsible collaborator(s): Kyoungmin Kim, kyoung-min.kim@epfl.ch

Duration: 3-6 months