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: https://www.epfl.ch/labs/dias/

Optimizing Semantic Operators for Structured and Unstructured Data

Keywords: LLM, batch processing, semantic operator

Problem: LLMs have changed our daily lives of interacting with data, from complicated machine-friendly languages to natural language. Chat bots enable interactive sessions that refine the information in multi-hop way, and prompting techniques help us communicate better with LLMs.

However, most applications and use cases focus on processing a single, ad-hoc query. This is somewhat different from how we batch process tabular data with relational databases. To fill this gap, semantic operators [1] have been proposed to enable batch processing with LLMs, managing LLM input and output in tables and exploiting optimization techniques in databases. Semantic operators can operate on both unstructured and structured data, allowing filtering, joins, and other database operators to be extended semantically.

Semantic operators offer declarative, high-level interface for batch processing LLMs, as in other recent studies [2, 3]. This project aims to survey and build applications using semantic operators, especially the ones where we can seek opportunities to improve performance. As this trend has just started, as a natural combination of LLM and databases, we expect to see lots of optimization opportunities using machine learning and database techniques.

[1] LOTUS: Enabling Semantic Queries with LLMs Over Tables of Unstructured and Structured Data
[2] A Declarative System for Optimizing AI Workloads
[3] DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines
[4] AgentKit: Flow Engineering with Graphs, not Coding
[5] ALTO: An Efficient Network Orchestrator for Compound AI Systems
[6] Parrot: Efficient Serving of LLM-based Applications with Semantic Variable
[7] Efficiently Programming Large Language Models using SGLang
[8] Teola: Towards End-to-End Optimization of LLM-based Applications
[9] CAESURA: Language Models as Multi-Modal Query Planners
[10] Compound AI: https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Project: In this project, the student will 1) survey and build applications using semantic operators, 2) analyze the performance bottlenecks, and 3) mitigate the bottlenecks.

Plan:

1. Survey and design applications using semantic operators (workload definition is the most important part).
2. Implement and run applications, analyze performance bottlenecks (accuracy, degree of hallucination, latency).

3. Improve performance using machine learning and database techniques (once steps 1-2 are concrete, optimization is not much challenging).

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s): Kyoungmin Kim, kyoung-min.kim@epfl.ch

Duration: 3 months