Agents for Data Science: From Raw Data to AI-Generated Notebooks using LLMs and Code Execution (Invited Talk)

Jiahao Cai
Google
USA
jhcai@google.com

ABSTRACT

Data science tasks involve a complex interplay of datasets, code and code outputs for answering questions, deriving insights, or building models from data. Tasks and chosen methods may require specialized data domain or scientific domain knowledge. Queries range from high-level (low-code) or highly technical (high-code). Code execution results, such as plots and tables are artifacts used by data scientists to interpret and reason about the current and future states of a solution towards completing the task. This presents unique challenges in designing, deploying and evaluating LLM-based agents for automating data science workflows. In this talk we will introduce an end-to-end, autonomous Data Science Agent (DSA) built around Gemini and available as an experiment at labs.google/code. DSA leverages agentic flows, planning and orchestration to tackle open-ended data science explorations. It uses LLMs for planning, task decomposition, code generation, reasoning and error-correction through code execution. DSA is designed to streamline the entire data science process, enabling users to query data in natural language, and get from a dataset and prompt to a fully AI-generated, populated notebook. We'll discuss design choices (prompting, SFT, orchestration), iterative development cycles, evaluation, lessons learned and future challenges. Where applicable, we will showcase real-world case studies demonstrating how DSA can assist with bootstrapping the analysis of data from complex scientific domains.

ACM Reference Format:
Jiahao Cai. 2024. Agents for Data Science: From Raw Data to AI-Generated Notebooks using LLMs and Code Execution (Invited Talk). In *Proceedings of the 1st ACM International Conference on AI-Powered Software (AIware '24), July 15–16, 2024, Porto de Galinhas, Brazil.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3664646.3676276

AIware '24, July 15–16, 2024, Porto de Galinhas, Brazil

ACM ISBN 979-8-4007-0685-1/24/07
https://doi.org/10.1145/3664646.3676276