LLM-based AI Assistants for Data Engineering

Is your degree almost in sight? Are you looking for a challenging internship in the field of software development/data engineering? Then we might be a good match! At Blenddata, we are looking for a graduation intern who can start around July 2024.

About your thesis

In your thesis, you will investigate the best way to leverage LLM-based AI assistants for various data engineering tasks. If you are successful in finding a possible solution to this complex problem, you will get the opportunity to implement it on a suitable project!

Why this subject

Blenddata aims to be at the forefront of technology. With your research, you can help us further and better understand how to use LLM-based AI assistants to support data engineering tasks. Your research and proposed solution will help us facilitate our customers with data engineering related questions and implementations.

What is Data Engineering?

Data engineering involves designing, constructing, and maintaining systems and infrastructure that collect, store, and analyze data. This field is crucial for enabling organizations to leverage their data for insights and decision-making.

How does that come together?

The integration of LLM-based AI assistants in data engineering involves several promising research directions:

- **Role-Playing Multi-Agents for Data Engineering Tasks**: Inspired by the works on multi-agent systems (MAS), the concept can be extended to data engineering. MAS, an established research area, can be leveraged to handle complex tasks in data pipeline design and management. For example, a data pipeline requires collaboration between data engineers, cloud experts, security experts, domain experts, and SQL experts. These scenarios can be modeled using MAS, with some agents being LLMs and others traditional agents. Research can focus on systematically identifying roles, designing agents, establishing communication protocols, and assessing the impact of different parameters on task achievement.

- **Task-Specific Assistants**: Focusing on specific data engineering tasks such as data understanding, pipeline code summarization and documentation, pipeline code migration (e.g., from Python to Polars, Spark to DBT), pipeline code refactoring, data quality test generation, data architecture design, and security threat discovery. The challenge is to ensure that the generated code is of high quality and maintainable, avoiding technical debt.

- **Robustness of AI Assistants in Data Engineering**: Investigating the robustness and reliability of LLM-based assistants in generating data engineering scripts. For example, examining the accuracy and reliability of SQL generators. Research can focus on identifying common pitfalls and improving the resilience of these AI-generated artifacts.

- **Detection and Tuning of Misconfigurations**: Using LLMs to detect misconfigurations in data engineering pipelines and providing recommendations for tuning them. Research can explore how LLMs can extract best practices from user manuals and apply them to detect and correct misconfigurations automatically.

- **Policy-Driven Data Architecture Design**: Utilizing domain-driven design principles to develop microservice architectures for data engineering that are policy-compliant from the outset, with LLMs aiding in the design and enforcement of policies throughout the data lifecycle.

Who are we looking for?

You:

- Are seeking a graduate internship starting in June, July, August 2024;
- Are pursuing a degree in the field of Data Science, Computer Science, or similar;
- Live in the Eindhoven, Den Bosch area;
- Are enthusiastic, eager to learn, and ambitious;
- Enjoy working together in a team.

What to expect from us?

- A compensation of €500,- per month based on a full-time commitment;
- A friendly environment to develop yourself;
- You will become part of an enthusiastic team who are eager to help you develop further;
- The opportunity to join Blenddata upon successful completion of your thesis;
- Daily provided lunch with the team at the office;
- An inspiring environment with our office in the centre of Eindhoven;
- Monthly team events and weekly Friday afternoon drinks.

Are you interested? What are you waiting for!

We would love to get in touch with you! Apply by using the apply button below, or send us your resume with some brief information about yourself. We will contact you after this.

Any questions? We're happy to help you!

Please feel free to contact us:

- Vincent Fokker (+31 6 13 83 58 58 / vincent.fokker@blenddata.nl)
- Roel Smits (+31 6 81 58 02 99 / roel.smits@blenddata.nl)