Chair of Space Mobility and Propulsion (Prof. Manfletti)
TUM School of Engineering and Design
Technische Universität München

Semester Thesis

Large Language Models for Export Control Compliant Research

Topic

One of the greatest yet most underestimated challenges of any engineering project is knowledge management. This is doubly the case in academia, where data is constantly produced by undergraduate students, professors, and everyone in between. Large Language Models (LLMs) have given rise to a suite of research frameworks that help synthesize, interpret, and even critically review the written word as well as numerical data. Tools like Deep Research and CO-STORM allow researchers to cover ground at breakneck speeds, while the collation of data into ranked vector databases allows for existing research to be quickly indexed and searched.

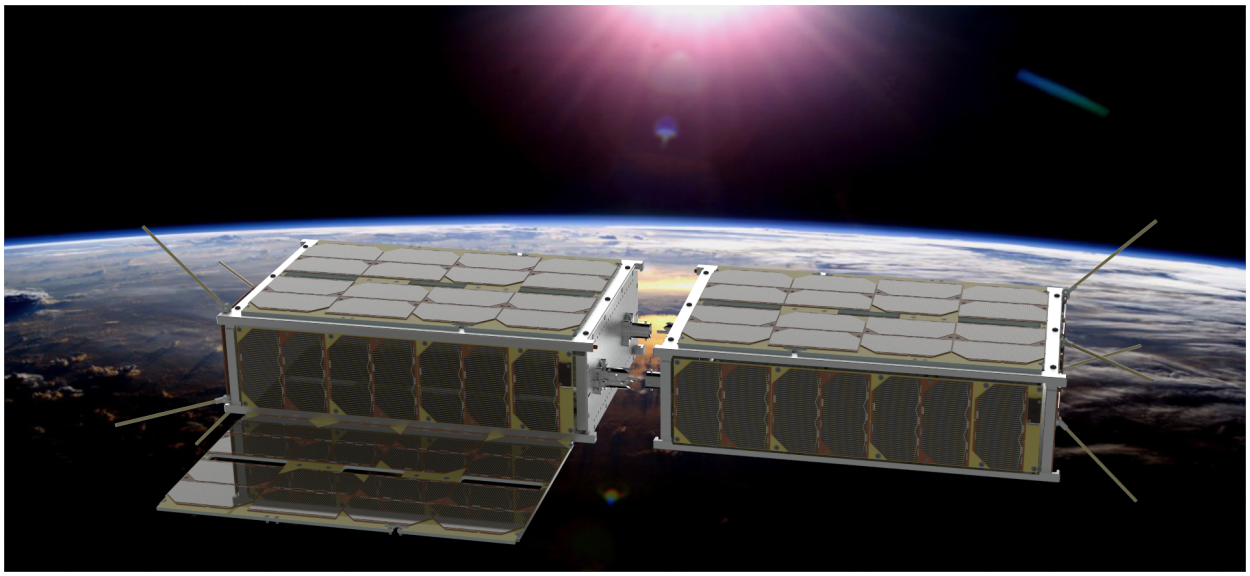

Figure 1: CubeSat Spacecraft in development at the Chair of Space Mobility and Propulsion

In this new paradigm, researchers working with sensitive or export-controlled research are immediately placed at a major disadvantage - their data cannot leave the country. As a result, the vast majority of existing tools are out of the question, as is working with most commercial partners. The solution to this problem must be local, and better yet tailored to the particular kind of data produced and processed at the Chair.

While open-source frameworks like Ollama and OpenWebUI have made local LLMs increasingly more accessible, they still lack tuning and optimization compared to commercial software. Out of the box, they do not handle content extraction well and are poorly optimized for the hardware and data present at the Chair. Your task would be to propose, implement, and optimize an open-source versatile research and knowledge management platform based on open-source software and models.

Chair of Space Mobility and Propulsion (Prof. Manfletti)
TUM School of Engineering and Design
Technische Universität München

Tasks

- Review of state-of-the-art of LLMs for synthesis and critical review of scientific data, with particular focus on **open-source solutions**.
- Review and evaluate best approaches (fine-tuning a model vs. RAG, comparing reasoning models, MoE) to augmenting a LLM given **a database of scientific texts.**
- Propose and implement methods to pre-process existing data.
- Implement a user-friendly, maintainable solution to interacting with a LLM, with **entirely offline** operation in mind to comply with export control requirements.
- Evaluate and implement **semantically accurate** content extraction (including tables, figures, equations) tailored to scientific material.
- Evaluate performance of existing solutions on **benchmark tasks provided.**

Your Profile

- Interest in and passion for informatics/data science.
- Knowledge of machine learning fundamentals.
- Experience with Linux desirable.

Contact

Tomas Mrazek
tomas.mrazek@tum.de
www.asg.ed.tum.de/spm