Internship - Development of software for an astronomical application and preparation of the corpus for an AI Large Language Models

Point de contact : S. Antier (IJCLAB, antier@ijclab.in2p3.fr), C. Douzet (IJCLAB)
In collaboration with T. Gerald (LISN) and his team
Duration: 5 to 6 months (niveau M1, M2) – Starting Date: From February to July 2026

Context

Nowadays, astronomers capture a plethora of transient phenomena such as stars devouring their partners, objects captured by supermassive black holes, dying stars, or even collisions of dead stars. Among those, high-energy phenomena can illuminate the sky over several hours to days. These rare events are in a very small minority but are of immeasurable value to several scientific fields, such as the origin of heavy elements found on Earth or the origin of Dark Energy. This is why we need to collect data sets using coordination from space and on the ground. In order to achieve that, we use a web application named SkyPortal (https://skyportal.io/), used both in the USA and Europe.

In this internship, in the context of the inter-disciplinary project MAFORAI (Monitoring Astronomical Follow-up Of Rare events with AI) and using Skyportal (a marshal and data science platform for time-domain astronomy), we aim to address the challenge of follow-up coordination in astronomy using an AI-based pipeline powered by a large language model (LLM). The long-term objective is to provide automated assistance to astronomers by using information collected in SkyPortal to suggest observational strategies but also better analysing the images collected during these campaigns.

For this internship, the work will explore the first stage of the project: creating a first corpus and data infrastructure behind to build the corpus. This includes the architecture needed to extract, organize, and label information from past observational campaigns found in Skyportal. These data products are currently dispersed across multiple sources, formats, and modalities, including heterogeneous data types such as text, images, and command logs, communications (such as instantaneous mission updates or informal reports). A key challenge is dealing with this heterogeneity and ensuring that the extracted information can be reproducible to train AI LLM.

Depending on the intern's interests, the corpus may focus more heavily on: a) text data, such as conversations, decision logs, with a focus on the evolution of the data regarding the timeline of the campaign b) vision, such as images combined with expert comments / metadata. We require prior experience with Python, GitHub, and LLM API (e.g. openAI).

Overall, the internship provides an opportunity to define the baseline of an AI-assisted follow-up coordination system, with emphasis on corpus creation, data organization, and early-stage system architecture.