Title: Automatic labeling from medical text data

Project Duration: 3 months (full-time) / 6 months (part-time)

Keywords: LLMs, RAG, Prompt Engineering, LLMOps

The focus of this project lies in creating a robust pipeline for extracting labels for hematological malignancies (medical diagnoses) and other information from doctor letters, i.e. reports by a clinician that describe the result of the examination. While the label information is available in the doctor letter, it is usually not available in a simple, computer-readable form. In order to train ML models, this information needs to be extracted into a structured format (e.g., `json`) that can be used in further data processing pipelines. This process suffers from hallucinations often produced by LLMs but could be robustified further using principles/ideas from AI Engineering such as

- Prompt engineering, providing answer templates and including additional relevant meta-information.
- Collaboration between LLMs and querying diverse LLMs (Medical LLMs and General purpose LLMs).
- Including relevant additional meta-information via RAG.

Data:

The data you will be working on are anonymized doctor letters. For this project, we will provide a collection of ~20.000 cases that can are partially labelled. Labels for a subset of these datapoints will be provided (1) as a validation set and (2) as a test set.

The core deliverables of this project are:

- Create a robust pipeline for automatically labeling potentially multilingual doctor letters with a predetermined set of labels.
- Investigate and build different improvements, e.g., based on prompt engineering, querying diverse LLMs or inclusion of meta-information.
- Extract further information from doctor letters where available, such as blood counts or other measurements, into a structured format.

This project is for you if

- You are highly motivated and self-sufficient
- You are a good python programmer
- You have a strong background in ML
- You are interested in working with LLMs, have some prior exposure to the topic and want to play around with prompts, RAG or other types of agentic workflows.
- You have a basic interest in medicine/biology and ML in the medical domain.

About us Blood is a unique window into your immune system, but analysis of these millions of tests today is done manually and subjectively. We at hema.to build a commercially available AI platform for clinical blood cancer testing. Our goal is to improve testing by providing faster, better and more reliable testing of immune disease with the mission of digital precision medicine for millions of patients, plug&play to the existing global cytometry infrastructure.

Literature

- https://proceedings.mlr.press/v225/goel23a
- https://arxiv.org/abs/2402.05129
- https://aclanthology.org/2023.findings-emnlp.603.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10690809/

Contact

Dr. Florian Pfisterer

florian@hema.to

Hemato - www.hema.to

If you are interested, please send an email to florian@hema.to, preferably with a short CV and some info on your background and motivation.