Take-Home Task: Extracting Structured Metadata from a Scientific Poster

Your task is to envision and demonstrate how LLMs can be used to extract structured metadata from scientific posters. Having structured metadata from posters can be useful to index them in a registry for instance to search and filter across multiple posters. Structured metadata of posters in a machine-friendly format (e.g., JSON), can also be useful to conduct automated analysis and discoveries from posters.

You will create a small GitHub project to showcase both your design thinking and practical implementation. Email bpatel@calmi2.org if you have any questions.

1. Create a GitHub repository for this Task.
2. In the README of the repository, describe your envisioned pipeline to extract the metadata listed in Table 1 below from the PDF file of a scientific poster using an LLM and any other required tools and resources (ideally all open source). Be concise and include:
    a. Key steps and components
    b. Tools, models, cloud infrastructure, and/or APIs you would ideally use
    c. Any assumptions or dependencies
    d. How you would evaluate your pipeline (approach, metrics, etc.)
3. Implement an example of your pipeline for this poster (or one of your choice). Write your code in a Jupyter notebook and share it on the GitHub repository. The code should take the path of the PDF of the poster on your computer as input and output a JSON file with the metadata from Table 1. Make the code modular and commented so it is easy for us to review.

    Note: We are just looking for a small reasonable snippet of your envisioned pipeline. It is fine if your code is not extracting all the metadata from Table 1 or not extracting them properly. You can explain possible improvements and adjustments you would make with more time (see step 5). If the pipeline described in step 2 requires specific hardware or resources that you do not have, you can use alternative solutions (e.g., a smaller LLM, commercial API, etc.) just for the purpose of this implementation. Document changes between your ideal pipeline from step 2 to the one implemented in step 3 (how it differs and why).
4. Document how to run your code in the README so we can test it.
5. Document in the README anything you deem useful for us to evaluate your exercise, like your process, assumptions, limitations, future work, etc.

Table 1. Examples of useful metadata for posters

Title of the poster
Authors (with affiliations)
Summary of the poster
Keywords
Methods
Results (main findings)
References
Funding source