Large Language Models for Analyzing Agricultural Research Station Project Descriptions

Scope of work - Samuel Carton, Computer Science

University of New Hampshire

Challenge/Goals

The Northeastern Regional Association of State Agricultural Experiment Station Directors (NERA) would like to develop a topical (and possibly methodological) expertise profile for each of its 15 agricultural experimentation stations, as well, possibly, as its 15 extension units. The basis for these profiles are a corpus of ~800 project titles and descriptions belonging to the various stations. The challenge is to use techniques from natural language processing (NLP) to extract the topics covered by each project, to then aggregate these topics into a topical (or methodological) profile for each station indicating its specialty area(s) relative to other stations. The ultimate purpose of these topical profiles would be to perform downstream tasks such as matching proposal solicitations to the most appropriate station based on topical (or methodological) expertise.

There are three key topics of interest: develop resilient, sustainable, and equitable food systems; lead effective adaptation and mitigation for our changing climate; and promote environmental, human, animal, and community health and well-being. However, it is still desirable to understand other topics touched on outside this set.

The proposed solution to this challenge is to develop a protocol for applying large language models (LLMs) to this task in an inductive “grounded theory” manner, involving iteratively extracting topics from each project description and accumulating a global list of topics from the full corpus. Then, the list of topics covered by each station can be aggregated and analyzed to produce a topical profile for each station (such as that station X works on topic Y at a rate much higher than the average for other stations). I also proposed to repeat the protocol for methods in order to develop a corresponding methodological expertise profile for each station. Other dimensions of analysis are also possible, but I proceed assuming we will cover only topics and methods.

Deliverables

There are three main deliverables to be produced:

1. **Topical and methodological profiles**. The primary deliverable will be a list of topics covered, and methods used, for every project description in the corpus, as well as an aggregation of this data into a topical and methodological profile for each station.
2. **Grounded theory protocol.** The secondary deliverable will be a relatively simple, repeatable protocol for applying large language models to this extraction/aggregation

process. This protocol will allow the topic/method data to be regenerated, extended, or replicated using a different dimension of analysis, at a later date if desired.

3. **Scholarly work.** I plan for the protocol described above to presented in at least one research paper, to be submitted to either an NLP conference such as the Association for Computational Linguistics (ACL), or a computational social science (CSS) conference such as the International Conference on Computational Social Science (IC2S2).

Administration

The project will be primarily supervised by Dr. Samuel Carton, with additional supervision from Dr. Anton Bekkerman and Dr. Rick Rhodes. The primary labor will be performed by a single research assistant, to be determined, possibly under the direct supervision of one of Dr. Carton's PhD students.

The meeting structure I envision is weekly or semi-weekly meetings between Dr. Carton and the primary research assistant, paired with biweekly or triweekly meetings between the entire team (Drs. Carton, Bekkerman, Rhodes and the research assistant).

Timeline

I envision this as a roughly six-month project for an advanced undergraduate, masters, or PhD student in the computer science department. A hypothetical timeline assuming the IC2S2 2025 as the target venue might look like the following:

Aug 2024: Recruit research assistant via email

Sep-Dec 2024: Engineering and experimentation work

Dec 2024: Initial results presented to Drs. Rhodes and Bekkerman

Jan-March 2024: Writing & further experimentation for robustness & improvement over baselines

March 2024: Submission to IC2S2 2025 conference

Budget

Research assistant compensation

A copy of the UNH ratesheet can be found here: unh_rate_sheet.pdf

I envision this project constituting a minimum of 10 hrs/week for the primary research assistant throughout the entire project period. At an hourly rate of $30/hr, this would constitute a minimum of 6 months * 4 weeks * 10 hours = $7,200.

Computational resources

The most powerful model available via the OpenAI application programming interface (API) costs $60.00 / 1M tokens for access (prices listed here). If we estimate each project description at 10 pages/2500 words, then this would cost $120.00 to run the model once over the entire

corpus. To accommodate overhead, experimentation, and varying dimensions of analysis, I would conservatively estimate an overall budget of $3000 for access to the API for the project.