Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA21129

Applicant name: Ana Jovančević

Details of the STSM

Title: Work Plan for Short-Term Scientific Mission – Systematic Biases build in LLMs

Start and end date: 11/09/2024 to 21/09/2024

Description of the work carried out during the STSM

We planned activities for three research projects (two main ones and one spin off project) on LLM biases, and one project related to continuous work from OPINION working group meetings on annotation opinion-related research outputs. Activities included grantee (me), hosts (Dr Damian Trilling and Dr Johannes Gruber), and Dr Aleksandra Urman present during the visit.

1. LLM Functionalities (main project)

Project Overview: This project examines LLM functionalities by constructing high quality prompts, by computer scientists working on the project, led by 'prompt engineering' literature. The goal is to analyse how prompt quality affects LLM outputs. We will test different models (GPT, Mistral, Meta) under varying settings (e.g., temperature).

Models and Initial Testing: We conducted pilot tests on several models.

Follow-up Activities: A Slack channel was established, and tasks divided. I will lead the literature review and contribute to data collection, model setup, and writing results. Model setup will be led by Dr Johannes Gruber and Dr Aleksandra Urman (also participating in the projects), while data analysis will be led by Dr Damian Trilling.

2. LLM functionalities (spin-off smaller research)

Project Overview: This project examines LLM functionalities by crowdsourcing prompts from scientists with diverse backgrounds (qualitative/quantitative, cultural, linguistic).

Models and Initial Testing: We developed a survey for crowdsourcing prompts and conducted pilot tests on several models.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Follow-up Activities: Slack channel was established for this project as well. Planned activities include me as a lead for survey part of the research, with other activities divided as in project 1.

2. LLM Biases: Gender, Age, and Cultural Biases

Project Overview: We discussed the specific biases (gender, age, cultural) to be examined and agreed on methodologies. Tasks were divided among participants. The methodology involves refining existing 'association tests' by having LLMs generate stories about specific groups to analyse inherent biases in the text.

Models and Initial Testing: A decision grid for various LLMs and settings was set up, initial prompts were also constructed. and initial tests were conducted (on ChatGPT 4.0 model).

Follow-up Activities: A Slack channel was created for ongoing communication. I will focus on literature review, data analysis, and observing model setup for learning purposes (Dr Damian Trilling will lead textual analysis, and Dr Johannes Gruber and Dr Alexandra Urman will lead model setting up).

Note. This project will be conducted on reduced grid from project one, on the grid showing the best results in terms of output quality.

3. Setting up automatic review of ‘opinion’ abstracts

Project overview: We analysed the issue of automatically annotating opinion related abstracts to retrieve information whether opinion was targeted in each paper, and which analysis was used. With definition of opinion as follows ‘Opinion is a target (object) and evaluation (what do you think about it)’.

We managed to find a working code, lead by Dr Johannes Gruber and made a step closer to solving this issue.

Follow-up Activities: continue checking and refining the code.

Description of the STSM main achievements and planned follow-up activities

Main results

Comprehensive Understanding of LLMs Biases: Through an exhaustive literature review and on-site pilot testing, we have advanced our understanding of LLM biases. Our work will reveal whether all LLM models exhibit similar biases across settings or if some are less biased. This study extends the existing literature, which primarily focuses on gender bias, by examining age and culture-related biases as well.

During the visit, we discovered that LLMs not only exhibit biases toward gender groups but also other social groups, with different prompts revealing hidden biases. We also found that model settings, such as temperature, significantly affect responses, with higher variation leading to a greater chance of uncovering biases.

Novel Methodological Contributions: by planning to test different LLMs with different settings and different prompts (written both by experts in computer science and regular academics) we are expanding current methodologies and previously used ‘association tests’.

Contribution to OPINION work: by uncovering code able to annotate opinion-related research automatically we also contribute to work conducted on working group meetings.

Network building: This visit strengthened collaboration among OPINION action members and will lead to continued future partnerships.

Planned follow-up activities

Follow up activities will be facilitated by created Slack channels and online meetings. Follow up activities will be caried out for three research projects planned.

LLM functionalities (main project):

1. Finalize prompt development.
2. Set up the models.
3. Conduct testing.
4. Analyse data.
5. Write and submit the article for publication.

LLM functionalities (spin-off project):

6. Fine-tune the survey.
7. Obtain ethical approval for conducting research on human participants.
8. Collect voluntary participants from academia.
9. Set up the models (smaller grid based on results from the first project).
10. Conduct testing.
11. Analyse data.
12. Write and submit the article for publication.

LLM biases:

1. Complete a literature review on LLM biases beyond gender.
2. Fine-tune the prompts aimed at uncovering biases.
3. Conduct testing.
4. Analyse texts to uncover biases.
5. Write and submit the article for publication.

Contribution to Action MoU Objectives and Deliverables

This mission is closely aligned with the Action MoU objectives in the following ways.

This STSM aligns closely with multiple research objectives of the OPINION Action:

1) *Bring together the necessary theoretical perspectives;*
2) *Design interdisciplinary Action Working Groups;*
3) *Establish a common methodological research agenda;*
4) *Set gold standards.*.

This STSM also aligned with capacity-building objectives:

1) *Organize master classes, summer schools, workshops and STSMs to train researches;*
2) *Identify and establish cooperation with other research groups and projects;*
3) *Establish the study of textually expressed opinions as an independent subfield.*

In summary, this Short-Term Scientific Mission helped advance understanding of LLM biases, thus bringing us one step closer to actionable strategies for their mitigation.