# Mitigating Large language Models (LLMs) hallucination

- *11:59 pm 27/10/2024 (End of Week 12)*
- **Contributes** 50% of your final result
- **Group Assignment – Group of 3-5 students**

## Summary

This project requires you to work in a team led by a teaching staff aiming to implement and demonstrate one or two hallucination mitigation techniques in large language models (LLMs). With the advent of many powerful large language models (LLMs), including OpenAI's GPTs, Google's Gemini and Meta's Llama 3, a key challenge remains due to these models generating content that appears factual but is ungrounded, aka "hallucination." Subsequently, multiple techniques have been developed to mitigate hallucination in LLMs. Some popular techniques include **Retrieval-Augmented Generation** (RAG), **Self-refinement through Feedback and Reasoning**, and **Fine-tuning**. For a comprehensive survey of these techniques, please refer to the article "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models" by Tonmoy et al. (https://arxiv.org/abs/2401.01313 )

The ultimate aim of the project is for the team to learn one or two techniques and reproduce them on an open-source LLM to investigate their efficacy. While there are clearly many interesting aspects of this project, there are also some risks associated with this project as the team is required to experiment with different techniques and solutions. Thus, the team will have weekly meetings with the project supervisor to discuss the outcomes from previous week (and receive feedback) and to plan the activities for the coming weeks. If you join this project, you will work under the supervision of your project leader, a teaching staff of the unit, and within a team of students (between 3 and 5 students).

## Project requirements

- The project will be done in Python. If you have zero or little experience with Python, this project is not suitable for you.
- Source code maintained on Git based VCS (Github/Bitbucket/GitLab/…). You must provide read-only access to the teaching staff supervising your project.
- Project documentation maintained on OneDrive. You must provide access to the teaching staff supervising your project.
- Researching at least one hallucination mitigation technique
  - Learning the theory behind the technique
  - Learning the required practical and programming aspects for the technique
- Implementing the technique for a selected LLM (to be approved by the project supervisor);
- Implementing a UI for the user to interact with the resulting model;
- Running illustrative demo of a working prototype. Ideally, this prototype should be run on a server for the user to access the service from their browser (similar to ChatGPT).
- Evaluation of the system/model against benchmark evaluations/datasets.
- 
- Portfolio consisting of:
  - Individual weekly worklogs;
  - Individual fortnightly reports on the work done/achievements;
  - A team **project summary report**.
- The **project summary report** (20-30 pages) includes the following sections:
  - Cover Page (with team details) and a Table of Contents (TOC),
  - Introduction,
  - Overall system architecture (including details of the external libraries or software packages used by your solution),
  - Detailed installation/user guides,

  - Details of the implementations members of the team have done,
  - Details of the experimentations and evaluation members of the team have carried out,
  - Scenarios/examples to demonstrate how the system works,
  - Some critical analysis of the implementation,
  - Analysis/Discussion of the experimental results, and
  - Summary/Conclusion.
- A video presentation of the project (12-15 minutes)

## Option C – Marking Scheme:

Each member of the team will be assigned weekly tasks. Depending on the effort put into the individual tasks + the quality of the work, the individual will receive weekly marks between **0 and 7 for the ten weeks 3-12** and feedback will be given during weekly team meetings. Each individual's final result for the Project will be the sum of the individual's marks for the ten weeks 3-12 plus the team portfolio assessment.
The team portfolio (maximum **30** marks) will be assessed as follows:

- Project report: Up to **20** marks
- Video presentation: Up to **10** marks
- Individual contributions: Up to **-20** marks (that is, even if the team portfolio is awarded 30 marks, a student who makes no/little contributions may get only 10 marks for this component)

## Submission

At least one member of the team must submit the entire project (code + report) as a .zip file to Canvas by 11:59pm of 27/10/2024. Create a single zip file with your code and a working version of your system. Standard late penalties apply – 10% for each day late, more than 5 days late is 0%.

You must also provide your project supervisor read-only access to your git repository within 1 week of forming teams.

The video **(12-15 minute duration)** should be submitted to Canvas by at least one member of the team by 11:59pm on Tuesday 29/10/2024.