# LLM Behavior Analysis & Responsible AI Evaluation

## Objective

The goal of this project is to study the behavior of Large Language Models by analyzing **hallucination patterns, reasoning ability, bias, context-length limitations, and safety constraints**. We will evaluate and compare three models: **GPT-2**, **LLaMA-3-8B**, and **FLAN-T5**.

## Methodology

## 1. Model Setup

We will load all three models using HuggingFace Transformers and create a common inference pipeline so that each model is tested under identical conditions.

## 2. Dataset Preparation

We will manually design four small evaluation datasets:

- **Factual Questions** – to measure hallucination
- **Reasoning Problems** – to test logical consistency
- **Ambiguous Prompts** – to analyze uncertainty handling
- **Ethical Dilemmas** – to study bias and safety behavior

Each dataset will contain 20–30 prompts stored in CSV/JSON format.

## 3. Evaluation Process

For every model and every dataset:

- Prompts will be passed to the model
- Generated responses will be stored
- Token length and outputs will be logged

All results will be combined into a master dataframe.

## 4. Metrics Used

We will compute the following:

- **Hallucination Rate**
  → Incorrect factual answers / Total factual questions

- **Bias Analysis**
  → Gender/job stereotype prompts (e.g., nurse/engineer) and frequency counting
- **Context-Length Performance**
  → Testing prompts at 256, 512, 1024+ tokens and measuring accuracy drop
- **Safety Constraints**
  → Checking whether models refuse or comply with harmful/ethical prompts

## 5. Visualization

We will generate:

- Hallucination comparison (bar chart)
- Bias distribution (bar/pie chart)
- Context length vs accuracy (line graph)
- Safety compliance (stacked bar chart)

## 6. Analysis & Reporting

Based on graphs and metrics, we will compare models and document:

- Which model hallucinates most
- Which performs best in reasoning
- Bias tendencies
- Context limitations
- Responsible AI behavior