OBJECTIVE

Primary Goal

To build an intelligent system that proactively assesses hallucination risk and dynamically routes queries to the most appropriate and reliable response pathway.

Specific Technical Objective

1. Design a Hybrid Confidence Module The first objective is to engineer a module that calculates a comprehensive confidence score by analysing two distinct layers of data:

- **Internal State Analysis:** Evaluate the LLM's internal uncertainty using signals such as Semantic Alignment, Internal Convergence, and Learned Confidence.
- **External Reality Check:** Assess the quality of external information before generation to identify risks. This includes detecting**:**
  - **Context Scarcity:** When retrieved information is too sparse.
  - **Context Conflict:** When source documents contradict one another.
  - **Domain Mismatch:** When retrieved documents are irrelevant to the query.

2. Develop a Dynamic Routing System The second objective is to implement a decision-making mechanism that uses the calculated confidence score to direct the user's query to the most efficient pathway.

The routing options include:

- **Direct Generation:** For high-confidence queries where the model can answer efficiently without aid.
- **Retrieval-Augmented Generation (RAG):** For queries requiring external grounding.
- **Cloud Model:** Flagging the query when high risk or conflicting data is detected.

3. Create a Custom Labelled Dataset The third objective is to construct a dataset specifically for training and evaluating the hallucination risk assessment components.

- This involves generating responses from existing QA benchmarks (such as Natural Questions and Trivia QA) and manually or automatically labeling them for hallucinations to create a ground truth for the model.

4. System Integration and Benchmarking The final objective is to integrate these components into a complete system and test it against established baselines.

- The system will be designed around a small, open-source LLM (specifically targeting models around 1B parameters) to demonstrate computational feasibility on standard hardware.

Project Scope

The project is scoped to focus specifically on text-based, knowledge-intensive Question Answering (QA) tasks using moderately-sized open-source Large Language Models.