PROBLEM DEFINITION

Large Language Models often produce answers that sound confident but are actually incorrect. These hallucinations happen because the model lacks self-awareness—it cannot reliably judge when it doesn't know enough to answer. Most existing solutions try to fix the problem *after* the model has already generated an unreliable response, which wastes computation and still allows incorrect information to appear. To address this fundamental weakness, the project focuses on building a system that can proactively assess the model's confidence before generation and route the query safely based on that assessment.

2.1 Objectives

- **Estimate Confidence Before Generation**
  Develop a mechanism that predicts whether the model is likely to hallucinate *before* it produces an output.
- **Use Multi-Signal Confidence Assessment**
  Combine semantic alignment, internal processing stability, and a learned confidence predictor into one reliable score.
- **Implement Intelligent Routing**
  Route queries to the appropriate path—local generation, retrieval augmentation, larger model, or human review—based on confidence thresholds.

- **Reduce Computational Overhead**

  Avoid unnecessary use of large models or post-hoc checks, improving overall efficiency.

- **Enhance Reliability and Accuracy**

  Improve hallucination detection rates, reduce false positives, and significantly boost F1 performance over baseline methods.