CHAPTER 4

RESEARCH / TECHNOLOGY GAPS AND CHALLENGES

Despite rapid progress in large language models, several critical gaps limit their reliability—especially when handling knowledge-intensive or safety-sensitive tasks. This project directly addresses these shortcomings, but the broader challenges remain important to highlight.

4.1. Lack of Pre-Generation Uncertainty Awareness

Most existing systems detect hallucinations *after* the model has already generated an answer. There is no reliable mechanism for an LLM to assess its uncertainty **before** responding, leading to confident but incorrect outputs. This absence of proactive evaluation is a major gap in current LLM pipelines.

4.2. Overreliance on Post-Hoc Corrections

Techniques such as RAG, self-consistency checks, and output verification are computationally heavy and often too late—they only react after the hallucination has occurred. This creates unnecessary compute costs and fails to prevent misinformation.

4.3. Limited Use of Internal Model Signals

LLMs generate rich internal activations, but most hallucination-detection approaches barely use them. Important indicators such as semantic alignment drift, unstable layer progression, or inconsistent hidden-state behaviour are **under-explored**, leaving a large gap in reliable uncertainty quantification.

4.4. Absence of Unified Confidence Scoring

Existing uncertainty-estimation techniques tend to rely on a single method—entropy, sampling, embeddings, or external classifiers. None provide a **combined, multi-signal confidence score** that captures semantic, structural, and learned aspects of reliability. This lack of integration reduces accuracy and increases false positives.

4.5. Inefficient Routing of Queries

Current LLM systems do not make smart decisions about *where* a query should be handled. Heavy models are often used unnecessarily, while low-confidence queries are not escalated properly. A major gap is the absence of **deterministic routing mechanisms** tied to confidence estimation.

4.6. Bias and Domain Sensitivity Challenges

Embedding-based alignment and internal confidence predictors depend heavily on reference models and training data. This introduces risks such as:

- domain-specific inaccuracies
- biased confidence estimation
- mis-calibration in unfamiliar contexts

4.7. Threshold Generalization Issues

Static confidence thresholds can fail across domains or user contexts. Without adaptive thresholding, routing decisions may become inconsistent or unreliable—especially with varied query types.

4.8. Limited Evaluation on Larger Models

Much of the experimentation uses relatively small models (e.g., 360M parameters), which may not fully reflect how confidence signals behave at scale. This creates a research gap in validating the approach on larger, real-world LLMs.