Dissertation on

"A Cost-Effective, Proactive Hallucination Routing System for LLMs"

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology
in
Computer Science & Engineering (Artificial Intelligence and Machine Learning)

UE23AM320A – Capstone Project Phase - 1

Submitted by:

Sourabh S Mahindrakar	**PES1UG23AM313**
Chandan R	**PES1UG23AM917**
Sreephaneesha k	**PES1UG23AM314**
Sri Charan D A	**PES1UG23AM315**

Under the guidance of

Dr. Ravi Gorripati
Associate Professor
PES University

August - December 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 feet Ring road, BSK 3rd stage, Hosakerehalli, Bengaluru – 560085

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'A Cost-Effective, Proactive Hallucination Routing System for LLMs'

is a bonafide work carried out by

Sourabh S Mahindrakar	**PES1UG23AM313**
Chandan R	**PES1UG23AM917**
Sreephaneesha k	**PES1UG23AM314**
Sri Charan D A	**PES1UG23AM315**

In partial fulfilment for the completion of Fifth-semester Capstone Project Phase - 1 (UE23AM320A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning) under rules and regulations of PES University, Bengaluru during the period Aug. 2025 – Dec. 2025. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 5th-semester academic requirements in respect of project work.

Signature **Dr. Ravi Gorripati** Associate Professor	Signature Dr.Jayashree R Chairperson	Signature Dr. K S Sridhar Dean of Faculty and Registrar

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"A Cost-Effective, Proactive Hallucination Routing System for LLMs"** has been carried out by us under the guidance of **Dr. Ravi Gorripati, Associate Professor** and submitted in partial fulfilment of the course requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering (Artificial Intelligence and Machine Learning)** of **PES University, Bengaluru** during the academic semester Aug – Dec 2025. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES1UG23AM313	**Sourabh S M**	_____
PES1UG23AM917	**Chandan R**	_____
PES1UG23AM314	**Sreephaneesha K**	_____
PES1UG23AM315	**Sri Charan D A**	_____

ACKNOWLEDGEMENT

I would like to express my gratitude to **Dr. Ravi Gorripati,** Department of Computer Science and Engineering, PES University, for his continuous guidance, assistance, and encouragement throughout the development of UE23AM320A- Capstone Project Phase – 1.

I am grateful to Capstone Project Coordinators, **Dr. Chetana Srinivas** for organizing, managing, and helping with the entire process.

I take this opportunity to thank **Dr. Jayashree R,** Professor & Chairperson, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), PES University, for all the knowledge and support I have received from the department. I would like to thank **Dr. K S Sridhar** Dean of Faculty and Registrar, PES University for his help.

I am deeply grateful to **Prof. Jawahar Doreswamy**, **Chancellor, PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University,** for providing me with various opportunities and enlightenment every step of the way. Finally, Phase 1 of the project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

Large Language Models are powerful, but they also love to make things up—confidently. Most existing solutions try to fix these hallucinations after the model has already produced a bad answer, which wastes compute and doesn't stop the damage. This paper takes a different route: instead of reacting after the fact, it judges before the model even begins generating whether the answer is likely to be trustworthy.

The proposed system estimates the model's confidence using three signals:

- How closely its internal representations match reliable reference embeddings,
- How steadily its reasoning progresses through layers, and
- A learned predictor trained directly on activation patterns.

These signals are combined into one confidence score. Based on that score, the system routes the question to one of several paths letting the small model answer if it's confident, using retrieval when certainty dips, escalating to a bigger model when needed, and falling back to a human only when the model is truly lost.

Across standard knowledge-heavy QA benchmarks, this proactive approach catches hallucinations far better than older methods while using significantly less compute.

Instead of treating hallucinations like an afterthought, this framework brings early awareness into LLM pipelines—making them faster, safer, and more reliable for real-world use

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Problem: LLMs Sound Smart but Often Guess Wrong

Large Language Models are incredible at producing fluent, convincing text—but that fluency hides a dangerous flaw: they routinely generate statements that *sound true* but are actually wrong. These hallucinations are not rare, and in high-stakes settings like medical advice or legal reasoning, they can cause real harm.

1.2 Why Current Fixes Aren't Enough

Most existing systems try to catch hallucinations *after* the model has already generated an answer. Retrieval-augmented generation, consistency checks, or post-hoc verification help, but they waste compute and fail to prevent the model from producing false content in the first place. In other words, the model has already "spoken," and now you're scrambling to clean up the mess.

1.3 A Shift in Strategy: Stop the Hallucination Before It Starts

This project flips the usual pipeline. Instead of waiting for the model to hallucinate, we analyse the model's internal signals *before* generation begins. The goal is simple: determine whether the model is likely to answer reliably. If not, the system automatically reroutes the query to a safer option—retrieval, a larger model, or a human.

1.4 How We Estimate Confidence

The confidence estimator draws from three powerful internal clues:

1.4.1. **Semantic Alignment**

We compare the model's internal representation of the query with embeddings from a trusted reference model. If alignment is weak, the model probably doesn't "understand" the query well.

1.4.2. **Internal Convergence**

We look at how stable the hidden layers are as the model processes the input. Poorly converging layers usually signal uncertainty or confusion.

1.4.3. **Learned Confidence**

A dedicated neural predictor is trained to read internal activations and estimate reliability directly.

These three signals are combined into a single confidence score.

1.5 Smart Routing Based on Confidence

Once the score is computed, the system chooses one of four paths:

- **High confidence** → small local model responds
- **Medium confidence** → use retrieval to ground the answer
- **Low confidence** → escalate to a larger, stronger model
- **Very low confidence** → hand off to a human reviewer

This makes the system faster, cheaper, and safer than post-generation fixes.

1.6 What This Achieves

Across multiple QA benchmarks, this method sharply improves hallucination detection while cutting computational cost by around 40%. The model gets better at knowing *when it doesn't know*, and that self-awareness leads to fewer mistakes and tighter control over reliability.

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.

LITERATURE SURVEY

1. Foundations: Understanding Hallucination in LLMs

Recent comprehensive surveys have established the taxonomy and root causes of hallucinations, defining the scope of the problem your project addresses.

- **Ji et al. (2023) – *Survey of Hallucination in Large Language Models***
  - **Definition:** This work defines hallucination as content generated by an LLM that is fluent and syntactically correct but factually inaccurate or unsupported by external evidence.
  - **Taxonomy:** It distinguishes between Intrinsic Hallucinations (where the generated output contradicts the source material or itself) and Extrinsic Hallucinations (where the output cannot be verified from the source or contradicts real-world knowledge).
  - **Causes:** The survey identifies that hallucinations stem from both data collection issues (misinformation in training data) and inference mechanisms (decoding strategies like top-k sampling that prioritize diversity over accuracy).
- **Huang et al. (2023) – *A Survey on Hallucination in Large Language Models***
  - **Categorization:** This survey introduces a distinction between Factuality Hallucination (deviating from established real-world facts) and Faithfulness Hallucination (diverging from the user's provided input context).

- **Relevance to Project:** It highlights that while LLMs excel at reasoning, they struggle with "knowledge-intensive" tasks due to limited parametric memory, validating your project's focus on QA tasks.

---

2. The "Reactive" Standard: Post-Hoc Detection

Current industry standards largely rely on checking the model's work *after* it has been generated. This represents the "reactive" approach your project aims to replace.

- **Manakul et al. (2023) – *SelfCheckGPT***
  - **Methodology:** This is a "zero-resource" black-box method. It detects hallucinations by sampling multiple stochastic responses (e.g., 5–10 variations) from the LLM for the same query.
  - **The Core Idea:** It operates on the principle of consistency: if an LLM "knows" a fact, its sampled responses will likely be similar. If it is hallucinating, the samples will diverge and contradict one another.
  - **Limitations (The Gap):**
    - **Computational Cost:** It requires generating multiple answers for every single query, making it roughly 4.2x more expensive than standard generation.
    - Latency: It is slow because it acts only after the initial generation is complete, making it unsuitable for real-time low-latency applications.
    -

3. The "Preventative" Standard: Retrieval-Augmented Generation

To prevent hallucinations, RAG was introduced to ground LLMs in external data.

- **Lewis et al. (2020) – *Retrieval-Augmented Generation (RAG)***
  - **Methodology:** This seminal work combined a pre-trained sequence-to-sequence generator (like BART) with a dense vector retriever (DPR). Before answering, the model retrieves relevant documents (non-parametric memory) to inform its response.
  - **Impact:** It significantly reduced hallucinations in knowledge-intensive tasks by reducing reliance on the model's internal (parametric) memory.
  - **Limitations (The Gap):**
    - **"Always-On" Inefficiency:** Standard RAG retrieves documents for *every* query, even simple ones (e.g., "What is the capital of France?"), adding unnecessary latency and computational overhead.
    - **Context Hallucination:** If the retrieved documents are irrelevant or conflicting, the model can still hallucinate by misinterpreting the context, a failure mode your project explicitly targets with its "Context Conflict" check.

4. The "Proactive" Shift: Confidence-Aware Routing (Your Base Paper)

Your project directly extends this recently proposed framework which attempts to solve the inefficiencies of the above methods.

- **N. M (2025) – *Confidence-Aware Routing for LLM Reliability***

- **Core Philosophy:** This paper proposes shifting from "reactive correction" to "proactive assessment." It predicts reliability *before* generation to decide if the model needs help.
- **Methodology:** It calculates a confidence score using three internal signals:
    1. **Semantic Alignment:** Measuring the distance between the model's internal representations and reference embeddings.
    2. **Internal Convergence:** Analysing how stable the model's hidden states are across its layers.
    3. **Learned Confidence:** A dedicated estimation of the model's own certainty.
- **Routing:** Based on this score, it routes queries to: Direct Generation (High confidence), RAG (Medium confidence), or Human Review (Low confidence).
- **Your Contribution (The Extension):** The base paper is "Self-Aware" (internal signals only). Your project makes it "Context-Aware" by adding an External Reality Check (Context Scarcity, Conflict, and Domain Mismatch) to catch cases where the model feels confident but is being misled by bad data.

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.

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.

CONCLUSION OF CAPSTONE PROJECT PHASE - 1

In this first phase of the project, we have successfully defined the scope and methodology for building a **Cost-Effective, Proactive Hallucination Routing System**. Our work to date has established the following:

- **Problem Validation:** Through an extensive literature survey, we validated that current hallucination mitigation strategies are largely reactive and computationally expensive, often attempting to fix errors after they occur.
- **Proposed Innovation:** We have designed a solution that improves upon the "self-aware" base paper by making it "context-aware." We introduced a novel **External Reality Check** layer to assess risks such as Context Scarcity, Context Conflict, and Domain Mismatch.
- **Feasibility Confirmation:** We conducted a feasibility study confirming that the system is computationally viable. By utilizing smaller LLMs (360M parameters) and open-source tools, the project can be implemented using standard university hardware.
- **Strategic Roadmap:** We have outlined a clear path for Phase 2, which includes creating a custom labeled dataset from QA benchmarks and training the confidence module to dynamically route queries between Direct Generation, RAG, and Human Review.

Final Summary: This project bridges the gap between model efficiency and reliability. By shifting from post-hoc correction to pre-generation risk assessment, we aim to create a system that allows LLMs to be safely deployed in high-stakes domains like healthcare and finance without the inefficiencies of "always-on" RAG.

PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

1. Dataset Preparation & Pre-processing

- **Objective:** Create the ground truth required to train the risk assessment model.
- **Action Items:**
    - **Acquire Benchmarks:** Download and set up existing QA benchmarks such as **Natural Questions** and **Trivia QA**[11].
    - **Generate Responses:** Run these questions through the selected small LLM (approx. 360M parameters) to generate initial responses[2].
    - **Labelling:** Create a custom dataset by labelling these generated responses for hallucinations[3333]. This involves categorizing answers as factual, hallucinated, or ambiguous to serve as training data.

2. Development of the Confidence Module

- **Objective:** Build the hybrid system that analyses both internal model states and external data quality[4].
- **Action Items:**
    - **Internal Signal Extraction:** Implement the logic to extract the three internal signals defined in the base paper: **Semantic Alignment, Internal Convergence, and Learned Confidence**[5].
    - **External Reality Check Implementation:** Develop the new "Context-Aware" layer to detect your three specific risk factors:
        - **Context Scarcity:** Algorithms to measure if retrieved context is too sparse[6].

- **Context Conflict:** Integration of Natural Language Inference (NLI) to detect contradictions in source documents[7].
- **Domain Mismatch:** Semantic similarity checks to verify relevance between query and documents[8].

3. Routing System Logic & Integration

- **Objective:** Develop the dynamic decision-making engine[99].
- **Action Items:**
  - **Score Aggregation:** Create a formula or classifier that combines the Internal and External scores into a single **Confidence Score**[10].
  - **Thresholding:** Define the thresholds that determine which path a query takes:
    - **High Confidence --> Direct Generation**
    - **Medium Confidence/Need Grounding --> RAG**
    - **High Risk/Conflict -->Human Review**.

4. System Evaluation & Benchmarking

- **Objective:** Prove the effectiveness of the system against established baselines[12].
- **Action Items:**
  - **Baseline Comparison:** Compare your "Proactive" system against "Reactive" methods (like SelfCheckGPT) and "Always-on RAG" in terms of computational cost and accuracy[1313131313131313].
  - **Performance Metrics:** Measure:
    - **Accuracy:** Reduction in hallucination rates.
    - **Efficiency:** Time taken per query and computational overhead.

- **Routing Precision:** How accurately the system identifies when to use RAG vs. Direct Generation.

5. Documentation & Final Presentation

- **Objective:** Compile findings into a research paper and final defence.
- **Action Items:**
  - Draft the final report including the methodology, experimental setup, and results.
  - Prepare the Phase 2 Final Presentation.
  - (Optional Goal) Prepare the work for publication as mentioned in your project scope references.

REFERENCES/BIBLIOGRAPHY

Primary References

- **[1]** N. M, "Confidence-Aware Routing for Large Language Model Reliability Enhancement: A Multi-Signal Approach to Pre-Generation Hallucination Mitigation," *arXiv preprint arXiv:2510.01237*, 2025.
- **[3]** P. Manakul, M. Gales, and A. Prom-on, "SelfCheck GPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in Proc. EMNLP, 2023, pp. 9004-9017.

Additional Key Works Cited in Literature Survey (From Slide 12)

- **RAG Foundation:** Lewis et al., 2020, "Retrieval-Augmented Generation (RAG)".
- **Hallucination Surveys:** Ji et al., 2023; Huang et al., 2023.

APPENDIX A DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

1. Acronyms and Abbreviations

- **ESA:** End Semester Assessment (Implied from file name context).
- **ISA:** In-Semester Assessment.
- **LLM:** Large Language Model.
- **NLI:** Natural Language Inference.
- **NLP:** Natural Language Processing.
- **QA:** Question Answering.
- **RAG:** Retrieval-Augmented Generation.
- **SRN:** Student Registration Number.

2. Definitions of Key Terms

- **Context Conflict:** A risk factor where contradictions exist between different source documents, potentially confusing the model.
- **Context Scarcity:** A condition where the retrieved context is too sparse or brief, forcing the LLM to guess and fill in blanks from its internal memory.
- **Domain Mismatch:** A situation where retrieved documents are not semantically relevant to the user's specific query.
- **Hallucination:** A phenomenon where Large Language Models produce content that appears plausible but is factually incorrect, undermining reliability.
- **Proactive System:** A proposed system designed to efficiently predict the risk of hallucination *before* generating a response, rather than correcting it afterward.

- **Reactive Methods:** Current solutions that attempt to detect or fix hallucinations only after the text generation process is complete.
- **Retrieval-Augmented Generation (RAG):** A technique that grounds LLM responses in external documents to improve factuality**.**
- **Routing System:** An intelligent mechanism that assesses risk and dynamically directs queries to the most appropriate pathway (e.g., direct generation vs. RAG vs. human review).
- **SelfCheck GPT:** A specific reactive method that checks for consistency across multiple generated answers to detect hallucinations.