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