A HYBRID MULTI-SIGNAL FRAMEWORK FOR HALLUCINATION DETECTION IN LARGE LANGUAGE MODELS

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in

Computer Science Engineering

by

Prince Chaudhary
(Roll No. 2k24/CSE/19)

Under the Supervision of

Dr. Rohit Beniwal
Assistant Professor
Department of Computer Science Engineering
Delhi Technological University

Department of Computer Science Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

June, 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to **Dr. Rohit Beniwal** for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its academic and practical relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Prince

Place: Delhi

Prince Chaudhary

Date: 01.06.2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Prince Chaudhary**, Roll No. **2k24/CSE/19**, student of M.Tech (Computer Science Engineering), hereby declare that the project dissertation titled "**A Hybrid Framework for Hallucination Detection in Large Language Models Using Semantic Retrieval and Adaptive Scoring**" which is submitted by me to the Computer Science Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or other similar title or recognition.

Prince

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled " **A Hybrid Multi-Signal Framework for Hallucination Detection in Large Language Models** " which is submitted by **Prince Chaudhary**, Roll No. **2K24/CSE/19**, Department of Computer Science Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 01/06/2026

Dr. Rohit Beniwal

SUPERVISOR

A Hybrid Multi-Signal Framework for Hallucination Detection in Large Language Models

Prince Chaudhary

ABSTRACT

Large Language Models (LLMs) such as GPT-4, LLaMA, and Gemini have demonstrated strong capabilities in generating fluent and contextually relevant responses across diverse natural language tasks. However, such models suffer from hallucinations, generating information that is erroneous or deceptive and may decrease the veracity of the model, especially for critical applications like medicine, law, and science. Prevailing tech- niques for hallucination detection have tended to rely on just one form of verification, such as retrieval grounding, semantic consistency, and confidence estimation, all of which are subject to shortcomings when applied individually.

The thesis develops a novel Hybrid Hallucination Detection Framework that uses multiple verification signals for the purpose of factual verification. The novel framework uses retrieval similarity with the use of TF-IDF and cosine similarity, semantic similarity by means of Sentence-BERT (SBERT) similarity, and confidence scoring using uncertainty measures to verify the factual accuracy of generated claims based on evidence retrieved from Wikipedia. Furthermore, an Adaptive Weight Optimizer dynamically adjusts the contribution of each signal to improve robustness across varying claim types.

The effectiveness of the proposed framework is evaluated on a controlled benchmark dataset and publicly available FEVER and VitaminC datasets. Experimental observations demonstrate effective separation between factual and hallucinated claims, highlighting the capability of the hybrid scoring mechanism to improve hallucination detection reliability. The proposed approach offers a lightweight and practical framework for enhancing factual consistency in LLM-generated responses without requiring extensive model retraining or fine-tuning.

TABLE OF CONTENTS

Acknowledgements **i**

Candidate's Declaration **ii**

Certificate **iii**

Abstract **iv**

Table of Contents **v**

List of Tables **vii**

List of Figures **viii**

1 INTRODUCTION **1**
1.1 Overview of Large Language Models 1
1.2 The Problem of Hallucination 1
1.3 Limitations of Existing Approaches 2
1.4 Proposed Solution 2
1.5 Contributions 3

2 LITERATURE SURVEY **4**
2.1 Hallucination in Language Models 4
2.2 Single-Signal Detection Methods 4
2.2.1 Model-Based Approaches 4
2.2.2 Knowledge-Based Approaches 5
2.2.3 QA-Based Approaches 5
2.2.4 Sampling-Based Approaches 5
2.3 Hybrid and Multi-Signal Approaches 5
2.4 Fact Verification Benchmarks 6
2.5 Summary and Research Gap 6

3 Experimental Setup and Dataset **8**
3.1 Implementation Environment 8
3.2 Default Configuration 9
3.3 Evaluation Metrics 10
3.4 Baseline Methods 11
3.5 Datasets 12

3.5.1 Controlled Benchmark 12
3.5.2 FEVER Dataset 12
3.5.3 VitaminC Dataset 13

4 METHODOLOGY **14**
4.1 Problem Formulation 15
4.2 Wikipedia Retrieval Module 17
4.2.1 Title Candidate Generation 17
4.2.2 Retrieval Similarity Score 18
4.3 Semantic Similarity Module 18
4.4 Confidence Scoring Module 19
4.4.1 Entropy-Based Scoring 19
4.4.2 Log-Probability Scoring 20
4.4.3 Heuristic Fallback 20
4.5 Adaptive Weight Optimizer 21
4.5.1 Signal Variance Strategy 21
4.5.2 Calibrated Strategy 23
4.5.3 Entropy Reweighting Strategy 23
4.6 Hybrid Score Computation and Verdict Assignment 24

5 RESULTS AND DISCUSSION **26**
5.1 Results on the Controlled Benchmark 26
5.2 Results on FEVER 27
5.3 Results on VitaminC 29
5.4 Consolidated Comparison 30
5.5 Discussion 31
5.5.1 Zero False Positives 32
5.5.2 Strength of Multi-Signal Fusion 32
5.5.3 Limitation on VitaminC 33

6 CONCLUSION AND FUTURE SCOPE **34**
6.1 Conclusion 34
6.2 Future Scope 35

References **36**

List of Publications **38**

List of Tables

2.1 Comparison of Hallucination Detection Approaches 7

3.1 Default System Configuration of the Proposed Framework 9

5.1 Results on Controlled Benchmark (20 Samples) 26
5.2 Results on FEVER Dataset (50 VERIFIABLE Samples) 28
5.3 Comparison with Prior Work on FEVER (No-Training Setting) 28
5.4 Results on VitaminC Dataset (60 Samples, Wikipedia Retrieval) 29
5.5 Accuracy Comparison Across All Three Datasets (%) 30

List of Figures

4.1 Architecture of the Hybrid Hallucination Detection Framework. 15
4.2 Flowchart of the Adaptive Weight Optimizer. 22

5.1 Score distribution for the controlled benchmark.. 27
5.2 Accuracy comparison across all evaluation datasets. 31

CHAPTER 1

INTRODUCTION

1.1 Overview of Large Language Models

Large Language Models (LLMs) like GPT-4, LLaMA, and Gemini have fundamentally changed the way things operate in the domain of NLP because of their capability of producing top-notch textual output for a variety of tasks including question answering, summarization, dialogue systems, machine translation, code generation, recommendations, and content generation among others. Using transformer models and training on huge volumes of textual information using self-supervised learning strategies, they can learn from lin- guistic patterns, dependencies, and relationships in the context.

The remarkable success of LLMs can be attributed to their ability to generalize across tasks without requiring task-specific retraining, enabling capabilities such as contextual understanding, the reasoning and coherent long-form generation. Their adop- tion has accelerated rapidly across industries including healthcare, finance, education, legal services, and enterprise automation, where they assist in improving productivity, decision support, and user interaction.

Although these developments have been made, there is one major drawback when it comes to LLMs. The basic approach taken by them is one of next-token probabilistic prediction and not of fact-based logical inference. Therefore, these models might produce output texts that look fluent, sensible, and extremely convincing, but contain information that is wrong, made up, or false altogether. This presents a major problem for any practical implementation of these language models.

The increased use of LLMs in decision support systems has led to greater focus on improving its trustworthiness, interpretability, and the consistency of facts. Out of all these issues, one of the major obstacles for the safe usage of LLM-based systems is the issue of hallucinations.

1.2 The Problem of Hallucination

Such an issue, which is widely referred to as ”hallucination,” constitutes one of the main unsolved problems of the contemporary NLP field. Some examples of hallucinations include false citations, inaccurate figures, improper attribution, and contradictory statements delivered with the same level of fluency as factual statements. Obviously, these responses may considerably diminish the reliability of large language models (LLMs), especially when they are used in areas like healthcare, finance, legal practice, and science, where factuality is crucial. In addition, the verisimilitude of LLM-generated text makes it hard for users to spot hallucinations.

Hallucination usually takes on three main categories:

- **Intrinsic Hallucination**

  The output generated by the model defies the provided conditioning data as it is inconsistent with what it learned from the inputs.

- **Extrinsic Hallucination**

  The outputs created are not verifiable or are entirely untrue, meaning that they can't be checked against any outside source.

- **Logical Inconsistency**

  While the output may be fluent on a sentence-by-sentence basis, it is logically incorrect, which includes such cases as erroneous causality and circular reasoning.

As these LLMs have become widely adopted for high-risk purposes, the results of hallucination may be harmful or even dangerous.

1.3 Limitations of Existing Approaches

The existing hallucination detection systems are highly dependent on using one signal for checking, which is why such systems have particular weaknesses.

- Fact-checking classifiers like FactCC require lots of annotated data and fine-tuning to be used in a specific domain, which reduces flexibility.
- The question-answering approach used by, for example, QAGS, has high computational cost since several inference steps are needed.
- The sampling approach used by, for instance, SelfCheckGPT is costly from the standpoint of latency, as several generation processes are performed.
- Retrieval-only systems have low semantic depth and are unable to handle paraphrasing in a claim.

Moreover, existing systems need special GPU architecture that limits their usability. Currently, there is no efficient solution that uses all these components in one system which is the motivation for this project.

1.4 Proposed Solution

This study introduces the concept of a Hybrid Hallucination Detection Framework, a relatively lightweight and training-free system that was created specifically to solve the problems associated with the available solutions for hallucination detection. Unlike existing systems that are based on the utilization of only one verification signal, the hybrid system is based on utilizing multiple complementary signals.

The described system uses three types of verifiers: the first uses Wikipedia similarity via cosine match based on TF-IDF, the second is dense semantic similarity via Sentence-BERT embeddings, and the third estimates token-level entropy to calculate model confidence score. Each type adds their own aspect into detecting the presence of hallucinations.

In order to improve performance further, the use of an Adaptive Weight Optimizer is employed to weight the impact of each of these signals depending on their relevance to the input claims. By doing this, the flexibility of hallucination detection is increased, compared to static weight methods.

With the use of external knowledge grounding, semantic similarity, and model confidence signals all being integrated into a single framework, the goal is to be able to achieve high accuracy in detecting hallucinations but at the same time maintain efficiency and be usable on consumer-grade hardware.

1.5 Contributions

In summary, there are several key contributions made by this thesis to improve hallucination detection in LLMs.

- This paper presents a Training-free Hybrid Hallucination Detection Framework, where a variety of signals can be combined in a weighted way, forming a unified approach for the detection process without using GPUs to deploy the model to consumer devices.
- Moreover, this thesis presents an Adaptive Weight Optimizer, which is the main novel contribution to solve the issue associated with traditional ensemble-based approaches. The weight can be adjusted adaptively for each sample depending on signal reliability.
- This approach is thoroughly analyzed using three well-known benchmarks for fact verification tasks, Controlled Benchmark, FEVER, and VitaminC.
- Finally, a dataset analysis study is provided to understand how a given dataset may or may not be suitable for retrieval-based hallucination detection systems.

CHAPTER 2

LITERATURE SURVEY

2.1 Hallucination in Language Models

The issue of hallucination in text generation using neural networks was explicitly stated for abstractive summarization by Maynez et al. [1] and divided into *intrinsic* hallucination and *extrinsic* hallucination, showing that current models hallucinate in more than 70% of their outputs. The work brought forward the notion that text could seem coherent yet be misleading, which is a significant problem in the field of automatic text generation.

Ji et al. [2] later conducted a review on the topic, providing a taxonomy of types of hallucinations within several Natural Language Processing (NLP) problems. They pinpointed that the problem stems from various sources including training noise, model architecture restrictions, and inference decoding techniques. It should also be emphasized that a hallucination does not result from failure in a model. Instead, it is the artifact of generating probabilistic language without factual reasoning.

Further research on the matter by Kong et al. [3] and Liu et al. [4] has shown that this issue extends to contemporary large language models (LLM) like GPT-4 and LLaMA, whereby hallucinations continue to occur across various fields in spite of significant improvements in the ability of these models. In this regard, their study shows that even very sophisticated models can output false citations, wrong information, and inconsistent answers in high confidence. The increasing use of LLMs in important areas such as healthcare, legal, and enterprise settings has highlighted the importance of factual correctness in these models [5].

2.2 Single-Signal Detection Methods

2.2.1 Model-Based Approaches

Approaches based on models employ internal model features like probability of tokens, confidence scores, and attention entropy as proxies for assessing the quality of generated text. The idea here is to try and detect hallucinations from the model's own uncertainty when generating new information without any external verification process.

Kadavath et al. [6] showed that the large language models can also be prompted to evaluate their own confidence in generated sentences. They found that although language models are generally aware of what they know, their self-reporting of confidence does not align perfectly with factual correctness, especially regarding infrequently observed facts.

Malinin and Gales [7] studied predictive entropy in estimating the uncertainty of autoregressive models for structured prediction. The research shows some correlation between higher uncertainty and hallucination probability, which proves that entropy-based

confidence estimation is an adequate measure of factual reliability.

2.2.2 Knowledge-Based Approaches

Knowledge-based systems validate generated statements via referencing knowledge sources and retrieval procedures to ground them factually.

FactCC [8] uses BERT classification for identifying inconsistent facts within the summary and its underlying article. Highly efficient at measuring factual inconsistency, the method is limited by the cost and availability of labeled datasets as well as fine-tuning requirements per domain or new task.

FActScore [9] validates a factuality of generated text based on decomposition into atomic claims and their validation via Wikipedia information. The method proved the feasibility of retrieval-based grounding without requiring fine-tuning of large language models. However, FActScore uses only one type of verification signal, neglecting confidence and similarity metrics.

2.2.3 QA-Based Approaches

QA methods pose the problem of factual verification in the form of a question-answering problem.

The QAGS system [10] uses a question-answering system for factual verification using question-answer pairs that can be created on the basis of generated summaries and then verifying whether there are discrepancies between answers obtained from source-conditioned and unconditioned models.

While QAGS has proved to be highly effective, it still relies on multiple inference runs for its working, resulting in high computational costs. Furthermore, QAGS is vulnerable to faulty question generation.

2.2.4 Sampling-Based Approaches

Sampling-based methods recognize hallucinations through examining the consistency between multiple generations.

SelfCheckGPT [11] recognizes hallucinations by performing multiple generations from the same input and assessing the semantic consistency among them. The rationale behind it lies in the assumption that a generation containing hallucinated contents would not always be consistent with another generation from the same input, but factual contents would.

As a training-free method, SelfCheckGPT performs competitively in recognizing hallucinations without needing any extra supervision. On the other hand, SelfCheckGPT incurs latency costs according to the number of generations and has no explicit way of verifying factual contents.

2.3 Hybrid and Multi-Signal Approaches

The shortcomings associated with using single-signal techniques to detect hallucinations led to the creation of mixed systems that involve multiple types of verifications to achieve high levels of accuracy and consistency. Due to various factors that can cause hallucinations, the use of only one type of verification may lead to the occurrence of inconsis-

tencies and inefficiency in detection processes. Kong et al. [3] created an LLM hallucination detection method called Multi-Perspective Consistency Checking that combines self-verification prompting with cross-modal model comparison. Their technique uses multiple perspectives to verify factual consistency and performs excellently even under zero resource conditions since there is no need for training models. The problem with their framework is that it involves multiple prompts and model comparison, making it computationally expensive. Liu et al. [4] presented a hallucination detection and mitigation framework in text summarization that uses fact-checking together with output filtering techniques. Their framework focuses on reducing the number of hallucinations by verifying facts and filtering the output. Similarly, Chakraborty et al. [12] presented an evaluation framework designed to be scalable and cross-domain, incorporating various viewpoints for the reliable evaluation of factuality within language models. The authors underscore the significance of scalable and reliable evaluation methods across different domains for reliably evaluating LLMs.

However, although such systems have their own merits, their dependence on training, use of multiple models, or complex verification steps renders them unsuitable for use in resource-limited settings. Conversely, the suggested framework integrates retrieval grounding, semantic similarity, and confidence estimation into a single, training-free, and lightweight pipeline.

2.4 Fact Verification Benchmarks

One of the popular datasets used for assessing fact verification and hallucination detection models is the FEVER dataset [13]. It consists of around 185,445 hand-verified claims along with their corresponding Wikipedia-based evidences and labeled in three ways, namely, *SUPPORTS*, *REFUTED*, and *NOT ENOUGH INFO*. Given that FEVER includes vast data instances and evidences, this has made it a standard benchmark dataset for evaluation of retrieval-based fact verification models.

The accuracy performance of the baseline FEVER system without proper evidence retrieval was approximately 50.91% [13]. Thus, the FEVER dataset has been considered to be a good dataset for hallucination detection due to the presence of both evidence and verification tasks.

The VitaminC dataset [14], unlike previous datasets such as FEVER, uses claim-evidence pairs generated based on Wikipedia revision history to incorporate contrastive claims into the fact verification task. This differs from the FEVER dataset because while FEVER deals with straightforward verification tasks, VitaminC aims to test how sensitive a system can be to small changes in facts.

Some of the types of small differences between fact pairs that have been used to construct claim-evidence pairs in VitaminC include inconsistencies in numbers, substituted entities, and contradictory semantics. Overall, FEVER and VitaminC complement each other as evaluation settings for hallucination detection.

2.5 Summary and Research Gap

The comparison of current methods used for hallucination detection brings to attention critical aspects of balancing training needs, computational costs, variety of signals, and

feasibility. Current techniques use one verification signal only, thus making them incapable of performing comprehensive detection in all cases. Model-based solutions depend mostly on the estimated model's uncertainty, knowledge-based models place emphasis on factual consistency, QA-based models are based on the consistency of Q&A pairs, while sampling-based models concentrate on semantic agreement in several samples.

Table 2.1. Comparison of Hallucination Detection Approaches

Method	**Signal Type**	**Training**	**Multi- Signal**
FactCC [8]	Classifier	Yes	No
QAGS [10]	QA-based	Yes	No
FActScore [9]	Retrieval	No	No
SelfCheckGPT [11]	Sampling	No	No
Kong et al. [3]	Multi-prompt	No	Partial
Proposed Framework	Retrieval + Semantic + Confidence	No	Yes

As shown in Table 2.1, existing lightweight and training-free approaches fail to have thorough validation techniques, since they consider individual signals without incorporating complementary viewpoints. Although existing hybrid techniques achieve greater robustness, these methods can suffer from higher computational costs, complex architectures, and low deployment capabilities in resource-limited environments [3, 4, 12].

Thus, there is a recognisable gap in designing a lightweight, training-free, and computationally inexpensive hallucination detection framework that integrates the concepts of fact verification, semantic analysis, and model confidence estimation. This gap is the fundamental reason behind the motivation and contribution of the proposed framework in this thesis.

CHAPTER 3

EXPERIMENTAL SETUP AND DATASET

3.1 Implementation Environment

The hallucination detection system suggested in this work is built upon Python 3.9, which has numerous machine learning, natural language processing, and data analysis packages that can make implementation faster and more effective. The Python programming language ensures flexible development, easy-to-manage integration, and sufficient community support.

In terms of architecture, the system is focused on implementing light computational solutions. This is done to ensure that it is possible to run our hallucination detector on normal consumer devices and not rely on dedicated GPU resources or expensive cloud computing services.

Several open source software libraries were used in developing some of the key modules in the proposed system architecture:

- The sentence-transformers library was used for calculating the semantic similarity of generated claims to the retrieved facts. The chosen model was all-miniLM-L6-v2 Sentence-BERT which offers good trade-off between computational efficiency and ability to produce accurate similarity measurements.

- In the retrieval module, scikit-learn library was used for TFIDF vectorization and cosine similarity computation in order to measure lexical similarity of the generated claims and retrieved facts.

- The Wikipedia-api software library was used for accessing Wikipedia's information directly in order to retrieve factual evidence without relying on limited-rate search engines.

- HuggingFace transformers library was used to load the google/flan-t5-base model used for the confidence estimation by measuring token-level entropy score.

  The implementation environment was set up in such a way that would facilitate replication, modularity, and experimental capability. All aspects of the framework, namely, retrieval, semantic similarity, confidence, and adaptive weights were built separately in such a way that would allow testing and further expansion of the system.

3.2 Default Configuration

In order to maintain consistency and fairness during all experiments, a default parameter configuration was used in the framework, which was obtained through prior experimentations and observations on achieving a good balance between factual grounding, semantic understanding, and confidence estimation.

The framework incorporates three main verifications that include retrieval similarity, semantic similarity, and confidence estimation. The framework places more importance on the retrieval component by assigning it a weight $\alpha = 0.5$ since it is important to emphasize the aspect of factual grounding with Wikipedia articles. Retrieval-based verification allows comparing facts with an external reliable source, thus ensuring factual consistency.

The semantic similarity sub-component, which uses Sentence-BERT embeddings, was given a weight of $\beta = 0.4$. The reason for using such an important value of β is that, through this, the model will be able to detect semantically similar claims despite significant lexical variations between them. Semantically analyzing statements is especially helpful in distinguishing paraphrased and contextually related sentences from one another.

The confidence estimation sub-component was given a lower weight of $\gamma = 0.1$. Even though the token-level entropy metric offers valuable information about the uncertainties involved, confidence scores in isolation cannot accurately classify real statements from hallucinations. Hence, the confidence estimation step was considered supplementary in nature.

Table 3.1. Default System Configuration of the Proposed Framework

Category	**Parameter**	**Value**
Weight Configuration	Retrieval weight (α)	**0.5**
	Semantic weight (β)	**0.4**
	Confidence weight (γ)	**0.1**
Decision Thresholds	Hallucination threshold (τ_h)	**0.42**
	Factual threshold (τ_u)	**0.58**
Models Used	SBERT model	All-MiniLM-L6-v2
	Confidence model	Google/flan-t5-base
Retrieval Setup	Wikipedia pages retrieved	Top-3 per query

Prediction Classification: In order to perform prediction classification, two threshold values were established. Predictions with a score less than the hallucination

threshold $\tau_h = 0.42$ were identified as *hallucinated*, whereas predictions that scored above the factual threshold $\tau_u = 0.58$ were classified as *factual*. Anything with a score falling between these two values was considered to be *uncertain*.

Semantically Similarity Calculation: The chosen model is `all-MiniLM-L6-v2` Sentence-BERT due to the relatively compact model structure and efficient semantic representation at the sentence level [15]. Entropy Scoring was used for confidence estimation via the `google/flan-t5-base` model due to its strong language understanding ability.

To ensure that factual grounding is reliable, the process of retrieval will be such that the top three Wikipedia pages are retrieved per query. Having more than one page to refer to ensures that ambiguity during retrieval is minimized and that there is a chance to identify any evidence.

3.3 Evaluation Metrics

In order to measure the effectiveness of the proposed hallucination detection method, four conventional metrics were employed during all experimental studies: **accuracy, precision, recall, and F1-score**. The rationale behind this choice is that these metrics give a general overview of classification performance based on the measurement of accuracy, reliability of prediction, detection capability, and balance between precision and recall.

Given that the proposed algorithm features a *UNCERTAIN* label, any prediction labeled as *UNCERTAIN* was treated as abstaining and automatically deemed to be correct regardless of the actual ground truth value. As far as this approach is concerned, it is worth noting that abstention is the right course of action when dealing with *NOT ENOUGH INFO* claims according to the FEVER dataset criteria [13].

Accuracy indicates the ratio of correct predictions made to the total number of predictions. Accuracy gives an overview of the performance of the machine learning model, though it does not always reflect how well the model predicts specific classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

where TP represents True Positives, TN denotes True Negatives, FP represents False Positives, and FN represents False Negatives.

Precision refers to the ratio of true positives among all the positive predictions made. Precision is an important criterion for assessing how precise the framework will be at detecting factual information as well as hallucinations.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

The **recall** is the ratio of real positives captured by the classifier among all positives. If the value of recall is high, then there is better accuracy and less number of miss-predictions.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

F1 score refers to the harmonic mean of the precision and recall, and therefore takes into account both types of errors together in one formula. This metric is especially effective in classifying systems, which require both precision and recall.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

Together, these metrics give a comprehensive evaluation of the proposed framework by capturing different aspects of the prediction quality, thereby enabling its reliable comparison with baseline methods and benchmark approaches.

3.4 Baseline Methods

For the evaluation of the effectiveness of the proposed hybrid hallucination detector architecture, three baselines that use one feature only were considered. The objective of such baselines is to examine the importance of each verification feature and compare it to the proposed hybrid architecture that uses several features at once. Through the evaluation of individual features in the isolation from others, the study aims at analyzing the pros and cons of retrieval similarity, semantic similarity, and confidence estimation separately.

- **Retrieval Only:** In this baseline, only the TF-IDF retrieval similarity score is used for verdict generation, with parameter settings $\alpha = 1$, $\beta = 0$, and $\gamma = 0$. The prediction relies entirely on lexical similarity between the generated claim and retrieved Wikipedia evidence. This baseline evaluates the effectiveness of external factual grounding without semantic understanding or confidence estimation.
- **Semantic Only:** This configuration uses only the Sentence-BERT semantic similarity score for classification, with parameter settings $\alpha = 0$, $\beta = 1$, and $\gamma = 0$. The approach focuses solely on semantic closeness between generated text and retrieved evidence, enabling the framework to recognize paraphrases and contextually similar claims even when lexical overlap is limited.
- **Confidence Only:** In this baseline, verdict generation depends entirely on token-level entropy-based confidence estimation, using parameter settings $\alpha = 0$, $\beta = 0$, and $\gamma = 1$. The objective of this baseline is to assess how effectively internal model uncertainty alone can identify hallucinated outputs without external factual grounding.

In addition to single-signal baselines, the proposed framework was benchmarked against established hallucination detection approaches reported in prior literature.

SelfCheckGPT [11] was selected as a representative zero-resource and training-free hallucination detection method based on sampling consistency across multiple generated outputs. Similarly, FActScore [9] was included as a retrieval-based factual evaluation framework that verifies generated claims against Wikipedia evidence.

Furthermore, the original supervised FEVER baseline [13] was used as a trained reference benchmark to compare the performance of the proposed lightweight framework against an established supervised fact verification system. These comparisons enable a broader evaluation of the proposed approach in relation to both training-free and supervised verification methods.

3.5 Datasets

3.5.1 Controlled Benchmark

A controlled benchmark containing 20 manually curated samples was developed using widely known factual topics such as capital cities, scientific constants, historical events, and commonly recognized attributions. Benchmark has been deliberately created to create a test environment where hallucination performance can be evaluated under unequivocal factual conditions.

Dataset includes the following examples:

- 10 factual statements, based on unequivocally correct and verified information from Wikipedia.
- 10 hallucinations, which contain some inaccuracies like incorrect names, numbers, misattributions, etc.

This benchmark enables precise evaluation under clearly defined ground-truth conditions without retrieval ambiguity, making it useful for analyzing the effectiveness of individual framework components in distinguishing factual from hallucinated outputs.

- Example factual claim: *"The Berlin Wall fell in 1989, marking a significant moment in the end of the Cold War."*
- Example hallucinated claim: *"The telephone was invented by Nikola Tesla in 1876."*
  (Correct answer: Alexander Graham Bell)

3.5.2 FEVER Dataset

The **FEVER** (Fact Extraction and VERification) dataset [13] contains approximately 185,445 Wikipedia-grounded claims categorized into three labels **SUPPORTS** , **REFUTES** and **NOT ENOUGH INFO**

FEVER is considered one of the most widely used benchmark datasets for retrieval-based fact verification systems and serves as a standard evaluation benchmark for

factual consistency assessment. The dataset provides evidence-grounded claims, making it particularly suitable for evaluating retrieval-based hallucination detection frameworks.

For this study, only verifiable claims (SUPPORTS and REFUTES) were selected for evaluation. SUPPORTS labels were mapped to FACTUAL, while REFUTES labels were mapped to HALLUCINATION. To maintain balance and fairness in evaluation, a subset of 50 claims was selected, consisting of 25 SUPPORTS and 25 REFUTES samples.

3.5.3 VitaminC Dataset

VitaminC [14] is an adversarial fact verification benchmark derived from Wikipedia revision history, containing more than 400,000 contrastive claim-evidence pairs. The dataset is specifically designed to test model sensitivity toward subtle factual modifications by introducing minimal edits to factual Wikipedia statements.

In VitaminC, the supporting and contradicting claims often remain mostly similar in wording while being different only in the factual correctness. This design makes the benchmark particularly challenging for hallucination detection systems, as lexical overlap between factual and hallucinated claims is often extremely high.

For evaluation, a balanced subset of 60 claims was selected, consisting of 30 SUPPORTS claims and 30 REFUTES claims, wikipedia page names were provided as retrieval hints to improve factual grounding consistency and reduce retrieval ambiguity during experimentation.

- SUPPORTS example: *"Westlife sold more than 1 million video albums and made over 23.5 million sales in the UK."*
- REFUTES example: *"Westlife made under 23.5 million sales in the UK."*

The contrastive design of VitaminC makes hallucination detection particularly challenging because both factual and contradictory claims often retrieve highly similar Wikipedia evidence. As a result, the dataset serves as a strong benchmark for evaluating the robustness of retrieval-semantic verification systems.

CHAPTER 4

METHODOLOGY

The proposed Hybrid Hallucination Detection Framework operates through a structured five-stage pipeline designed to assess the factual reliability of claims generated by Large Language Models (LLMs). The combination of external validation of facts, knowledge of semantics, and internal estimation of the system's confidence results in an accurate determination of whether the system output contains real facts, uncertainty, or hallucinations. The current method is different from conventional techniques because it is not based on a single verification technique.

The framework involves the following steps:
(1) Claim input and query formation.
(2) Wikipedia retrieval.
(3) Signal computation across selected three independent channels.
(4) Adaptive weight optimization.
(5) Hybrid score computation and its verdict assignment.
Each step plays its independent role in the process of evaluating claim validity while creating a coherent hallucination detection pipeline.

At the first step, the inputs, i.e., the generated claim and an optional question context, are processed to obtain retrieval-friendly queries, which are subsequently supplied to the Wikipedia retrieval module in order to retrieve supporting evidence. Next, three complementary signals are calculated: retrieval similarity, semantic similarity, and model confidence, offering different angles on measuring factual validity and language validity.

In order to make the evaluation more flexible and robust, the Adaptive Weight Optimizer calculates the optimal weight of each of these three signals, depending on its contribution to evaluating a particular input claim. Eventually, a composite factuality score is calculated and translated into one of the three verdict categories: *Factual*, *Uncertain*, or *Hallucination*.

The modular nature of the framework guarantees the computational efficiency, scalability, and extendibility of the system, making it possible to incorporate new verification cues or optimization techniques in future research without a need for drastic changes to its architecture.

4.1 Problem Formulation

The key focus of the framework lies in establishing whether an assertion made by an LLM can be described as being factual, doubtful, or imaginary. This is because the outputs produced by LLMs tend to be linguistically coherent even though the content might not always be factually correct.

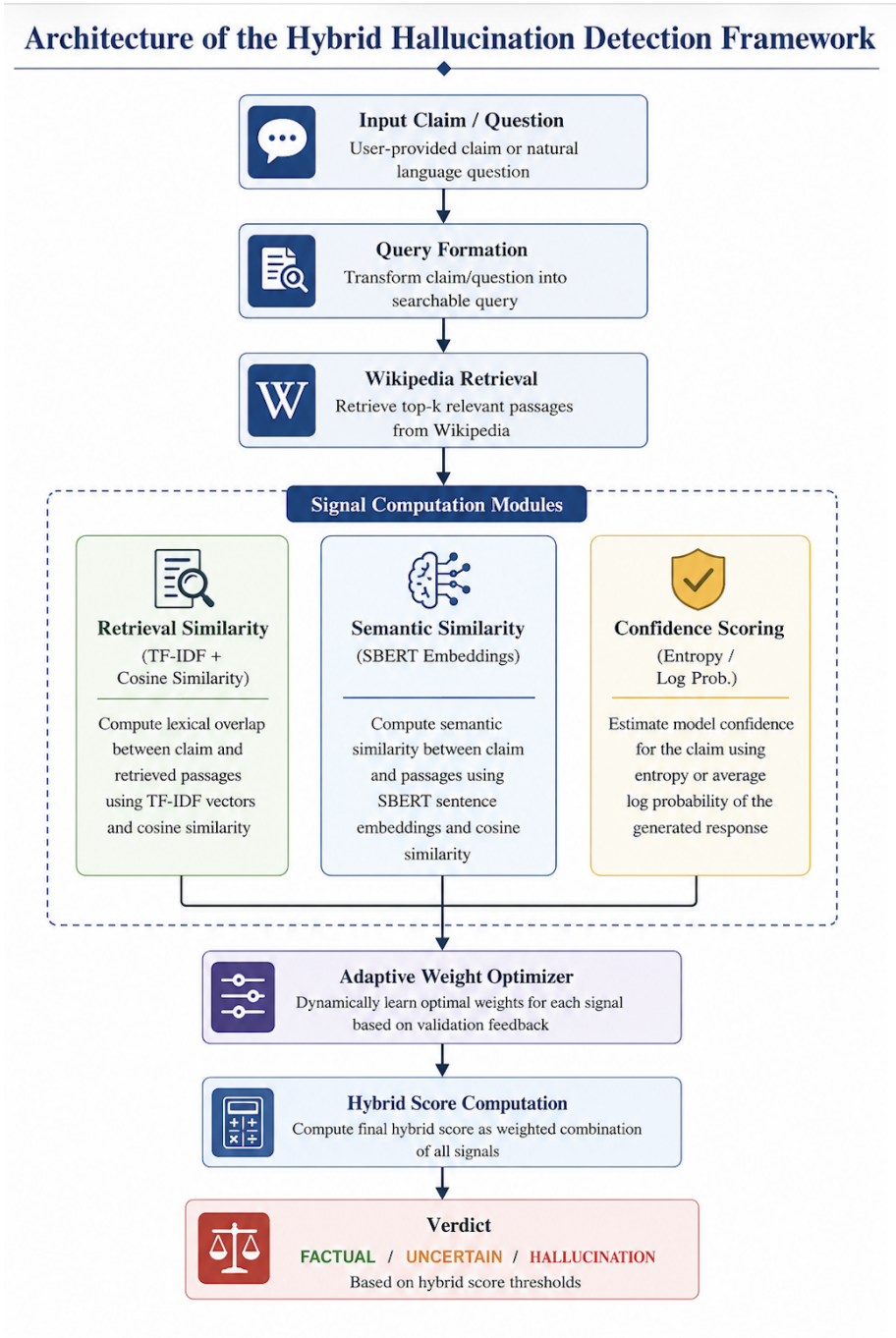

Figure 4.1. Architecture of the Hybrid Hallucination Detection Framework.

For a given claim c, which is in the form of a declarative sentence produced by an

LLM, along with optional question context q, the framework yields two outputs:

1. A factuality score $f \in [0,1]$, where high scores mean high factual validity, while low scores signal high probabilities of hallucination.
2. A final classification v among the following types:
   Factual, Uncertain and Hallucination.

The evaluation of the claim's factual reliability involves three verification signals which cover distinct facets of factual validation:

- **Retrieval Similarity Score (R_s):** Evaluates how well the statement is grounded by comparing it to evidence from Wikipedia in terms of lexical similarity.
- **Semantic Similarity Score (S_s):** Determines how contextually and semantically similar the generated statement is to the evidence extracted from Wikipedia via dense vector representations.
- **Confidence Score (C_s):** Provides a linguistic confidence score for the model based on entropy, functioning as a measure of uncertainty.

The factuality score is then determined by weighting the three signals above as follows:

$$f(c,q) = \alpha \cdot R_s(c,q) + \beta \cdot S_s(c,q) + \gamma \cdot C_s(c) \tag{4.1}$$

where R_s, S_s, and $C_s \in [0,1]$ represent the retrieval similarity, semantic similarity, and confidence scores respectively. The parameters α, β, and γ denote the relative importance given to each signal, subject to the constraint:

$$\alpha + \beta + \gamma = 1 \tag{4.2}$$

with $\alpha, \beta, \gamma \geq 0$.

In order to minimize the chances of excessive confidence in assigning an inappropriate label in borderline situations, the use of the dual threshold approach was adopted. In the case when the score of factuality is higher than the upper threshold, such claims are labeled as FACTUAL; whereas if it is below the lower threshold, it is classified as HALLUCINATION. All other claims are labeled as UNCERTAIN.

The final verdict assignment rule is stated as:

$$v = \begin{cases} \text{FACTUAL} & \text{if } f \geq \tau_u \\ \text{UNCERTAIN} & \text{if } \tau_h \leq f < \tau_u \\ \text{HALLUCINATION} & \text{if } f < \tau_h \end{cases} \tag{4.3}$$

where $\tau_h = 0.42$ and $\tau_u = 0.58$ are empirically determined threshold values which are selected through controlled benchmark experimentation.

In the case of UNCERTAIN region, this serves as a way to opt out intentionally, just like the *NOT ENOUGH INFO* class used in the FEVER dataset [13]. It ensures that there will be fewer erroneous predictions made with high confidence.

4.2 Wikipedia Retrieval Module

The task of the Wikipedia Retrieval Module is to obtain factual evidence related to the input claim. As the factual grounding process constitutes one of the key techniques used to detect hallucinations, it was crucial for the retrieval system to be efficient and reliable, as well as less reliant on external services. It was important not to introduce any errors, as this would result in an unreliable similarity score and decision making.

Instead of directly using the Wikipedia search endpoints, the designed framework first produces page titles based on the provided claim and then obtains evidence by searching the webpages according to these titles. This way, we managed to make the evidence extraction procedure more efficient, as well as reduce its dependency on the unreliable search functionality and avoid delays due to rate limitations.

4.2.1 Title Candidate Generation

In order to retrieve the related Wikipedia pages, the framework will generate title candidates using four distinct approaches. These approaches were chosen because of their ability to optimize the quality of retrieval process by extracting the entities, context, and informative words found in the claim.

1. **Multiple Word Proper Nouns Extraction:** Multiple word proper nouns are identified based on pattern matching and regular expression approach. Such an approach aids in the identification of meaningful entities including names of people, organization, place or event. Examples of such terms include *Albert Einstein*, *Berlin Wall* and *World Health Organization*.
2. **Proper Nouns Extraction:** This method captures important standalone entities that may be useful for retrieval when multi-word phrases are unavailable.
3. **Keyword Combination Generation:** Important content words are combined into two-word and three-word phrases after stop-word removal to improve retrieval effectiveness. This strategy increases the likelihood of matching semantically relevant Wikipedia pages even when exact titles are unavailable.
4. **Numeric Context Phrase Extraction:** Contextual phrases containing named entities and numerical information are identified to support claims involving dates, events, or numbered entities. Some examples include *Apollo 11*, *World War II*, and *COVID-19 Pandemic*.

A maximum of ten titles for each query are generated and retrieved from the wikiapi library where each page can be accessed directly in an efficient manner, without resorting to rate-restricted Wikipedia search APIs. Should this approach fail in finding appropriate pages, a second layer of Wikipedia searching process by keywords kicks in.

4.2.2 Retrieval Similarity Score

Having identified the pertinent sections in the Wikipedia database, the TF-IDF cosine similarity measure for retrieval similarity is calculated.

In order to evaluate the degree of factual consistency and textual similarity between the claim generated and the evidence from the Wikipedia passages, retrieval similarity score based on TF-IDF cosine similarity is computed. As against calculating the average of the retrieval similarity scores for all retrieved Wikipedia passages, it was decided to select the maximum similarity score. The reasoning behind this choice is that strong evidence will not be diluted with irrelevant evidence.

Retrieval similarity score is given by:

$$R_s(c,P) = \max\left(\cos(\vec{v}_c, \vec{v}_P), \max_{p_i \in P} \cos(\vec{v}_c, \vec{v}_{p_i})\right) \tag{4.4}$$

where $\vec{v}_c$ and $\vec{v}_{p_i}$ denote the TF-IDF vector representations of the claim and the retrieved passages, respectively. Both vectors are obtained through scikit-learn's TfidfVectorizer, which transforms textual content into numeric form for further similarity analysis.

4.3 Semantic Similarity Module

The Semantic Similarity Module computes the semantic similarity between the generated statement and the factual evidence through Sentence-BERT (SBERT) embeddings [15]. The retrieval-based similarity algorithms only consider the overlap of words between the two sentences. In contrast, semantic similarity allows the system to check whether the two sentences convey the same meaning regardless of their different lexical and sentence structures.

This algorithm has significant importance in detecting hallucinations in the system because factual claims can be made in ways that do not match the evidence lexically but still convey the same meaning. As a result, there might be scenarios where two sentences contain the same facts but use different wording. This means that the system cannot rely solely on lexical similarity.

This model is used in our architecture due to the effectiveness of its operation in terms of a compromise between the level of semantic precision, speed, and computation effectiveness. In comparison with more extensive Transformer models, the all-MiniLM-L6-v2 Sentence-BERT creates quality semantic representations while preserving lower computations than other transformer-based architectures.

Sentence-BERT transforms textual inputs into dense vector representations in a high-dimensional embedding space, where semantically similar sentences are positioned closer together. The claim generated by the language model and retrieved Wikipedia evidence are independently encoded into vector embeddings. The semantic similarity score is then computed using cosine similarity between the generated claim embedding and evidence embeddings.

Since retrieved Wikipedia contexts may contain long passages with mixed relevance, directly encoding the entire context may dilute meaningful semantic information. To address this issue, retrieved passages are divided into smaller 500-character chunks, allowing more precise semantic comparison between the generated claim and relevant evidence segments.

For each chunk, semantic similarity is computed independently, and the maximum similarity score is retained as the final semantic similarity value:

$$S_s(c,P) = \max_j \cos\big(\text{SBERT}(c), \text{SBERT}(\text{chunk}_j)\big) \tag{4.5}$$

where $\text{SBERT}(c)$ represents the embedding vector of the generated claim and $\text{SBERT}(\text{chunk}_j)$ denotes the embedding vector of the j-th retrieved context chunk.

With respect to the maximum similarity approach, only a highly relevant snippet will significantly contribute towards the value of the semantic score regardless of the fact that other irrelevant snippets have been retrieved by the retrieval process. This makes the method more robust with regard to noise retrieval.

Given the fact that the value of cosine similarity can potentially range anywhere between -1 and 1, the semantic similarity is restricted within the range $[0,1]$.

A fallback strategy is incorporated into the framework in order to guarantee that robustness is retained in cases where there are limited resources. In case Sentence-BERT embeddings cannot be generated, a TF-IDF cosine similarity is used to compute the similarity between the claim and the aggregated context. While inferior to dense semantic embeddings, this approach still provides adequate functionality.

4.4 Confidence Scoring Module

The Confidence Scoring Module calculates the level of confidence of the claim's generation by the language model as another signal for its factual correctness. As many of the hallucinations made by the language model are likely to include some degree of linguistic uncertainty or lack stability of token probability distribution, the confidence score is an extra verification signal in addition to the retrieval and semantic similarities.

In contrast to the methods of external retrieval, confidence scoring is aimed at evaluating the level of certainty inside the language model. In other words, the basic idea behind confidence calculation is that the more confident the predictions made by the model are, the more certain the model is linguistically. But this does not always mean that the output produced by the language model will be accurate, because at times, the language model can make inaccurate claims with great certainty.

4.4.1 Entropy-Based Scoring

The main entropy-based confidence estimator employs token-level Shannon entropy, which is determined from the probability distribution over tokens generated

by a generative language model (by default: google/flan-t5-base).

The Shannon entropy of the predicted tokens can be interpreted as the degree of unpredictability of their generation. Low entropy indicates a high level of confidence about predicting tokens, while high entropy implies low confidence.

Token-level entropy at each generated position is defined as:

$$H_t = -\sum_{w \in V} P(w \mid c_{<t}) \log P(w \mid c_{<t}) \tag{4.6}$$

Where $P(w|c_{<t})$ is the probability of token w given previously generated context, and V refers to the vocabulary of the model.

In order to have an overall confidence score, the entropy values for all generated tokens are averaged as:

$$C_s^{\text{entropy}}(c) = 1 - \frac{\bar{H}}{H_{\max}} \tag{4.7}$$

where $\bar{H}$ refers to the average entropy for all positions of the tokens, whereas $H_{\max}$ refers to the maximum entropy value that could be achieved. Inverting these values would ensure that a high confidence level results in higher values.

4.4.2 Log-Probability Scoring

Another confidence measure calculates the average token log-probability for the claim produced. As would be expected, higher probability claims will be deemed to be linguistically more confident by the model.

Scores are normalized via the sigmoid function such that scores are in the range $[0,1]$, facilitating easy combination with information retrieval and semantic similarity features:

$$C_s^{\text{logprob}}(c) = \sigma\left(\overline{\log P}(c) - \delta\right) \tag{4.8}$$

where δ represents an empirically chosen shift parameter which is used to stabilize score distribution across the generated claims.

4.4.3 Heuristic Fallback

The framework uses a lightweight heuristic confidence scoring mechanism in cases where there is no generative language model. This reduntant way calculates confidence by linguistic features of the said claim that can be observed.

The heuristic scorer takes account of factors such as:

- Presence of uncertainty indicators such as *"maybe"*, *"possibly"*, or *"likely"*
- Specificity indicators including dates, numbers, and measurable quantities
- Presence of named entities such as people, locations, and organizations

- Overall claim length and linguistic completeness

A default baseline confidence score is added to the confidence score for declarative statements, which are usually more linguistically certain. Heuristic scoring is not as accurate as entropy scoring, but can keep the framework running in resource-limited settings where transformer-based models are not available.

4.5 Adaptive Weight Optimizer

This thesis main novel contribution is the **Adaptive Weight Optimizer**. The proposed optimizer does not rely on fixed weights for different retrieval, semantic similarity or confidence signals, but dynamically adapts their weight for a given input claim based on their reliability.

Fixed-weight systems tend to do very well on average but can go wrong if the value of one or more of the signals is lost. For example, retrieval similarity might not work in the absence of relevant Wikipedia evidence, semantic similarity might weaken for domain-specific or obscure claims, and confidence estimation alone might not be reliable when the factual content is hallucinated. All signals are treated equally, which can degrade the overall prediction quality because of such limitations.

The proposed framework introduces an adaptive mechanism that adjusts the importance of signals dynamically with respect to the reliability of the signals associated with the claims. The optimizer does not assume that all signals are equally important in all situations, but instead adapts weights to give more weight to stronger and more reliable signals and to de-emphasize signals that are weaker or noisier. This enhances flexibility, robustness and factual fidelity for various claim types and retrieval settings.

It adopts three adaptive weighting strategies, Signal Variance Strategy, Calibrated Strategy and Entropy Reweighting Strategy. Every strategy separately calculates the reliability of the signal in a separate step and sets the weights before final hybrid score calculation.

4.5.1 Signal Variance Strategy

The Signal Variance Strategy was developed because signals in the middle of the range are more likely to have useful discriminative information while signals close to extreme values may be either a trivial certainty or a signal failure. Intuitively, a score close to 0 or 1 may not be of great interest to differentiate because it will be very obvious or unreliable cases.

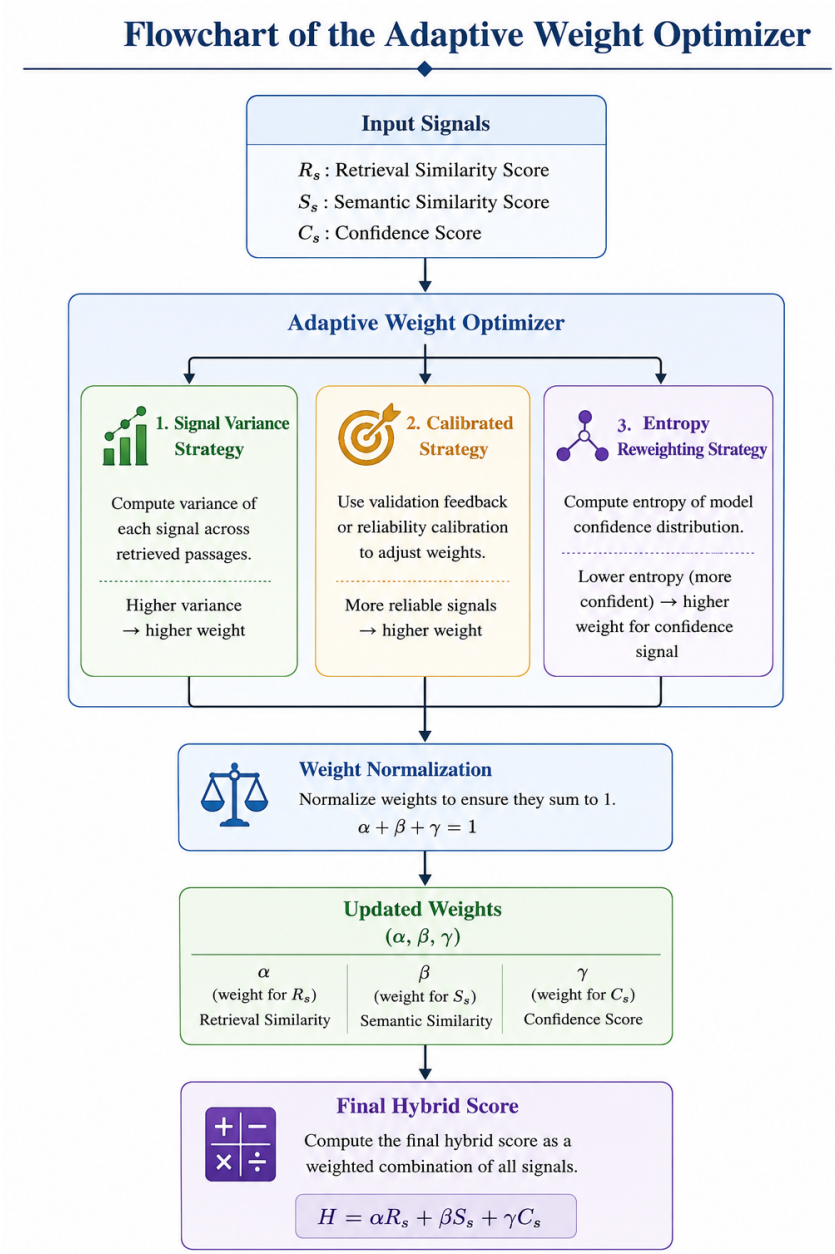

Figure 4.2. Flowchart of the Adaptive Weight Optimizer.

The framework calculates a reliability value based on the distance from the value of 0.5 to estimate the level of signal reliability:

$$\text{reliability}(s) = 1 - |s - 0.5| \times 2 \tag{4.9}$$

where s denotes an individual signal score.

After using these reliability estimates, adaptive weights are computed as:

$$w_i' = \frac{\text{reliability}(s_i) \times w_i^{\text{base}}}{\sum_j \text{reliability}(s_j) \times w_j^{\text{base}}} \tag{4.10}$$

where w_i^{base} is representing the default signal weights. To prevent complete elimination of any signal, a minimum weight constraint is applied, which make sure that every signal contributes at least marginally to the final prediction.

4.5.2 Calibrated Strategy

The Calibrated Strategy applies empirically derived rule-based adjustments to rebalance signal importance under specific conditions. This strategy is motivated by practical observations from experimentation, where certain signal failures consistently reduced prediction quality.

For example, when retrieval similarity becomes very low, indicating retrieval failure or poor evidence quality, the retrieval signal weight is reduced while semantic similarity and confidence receive greater emphasis. Similarly, when model confidence is particularly high, confidence contribution is increased to better reflect stronger linguistic certainty.

The strategy applies the following rule-based modifications:

- If $R_s < 0.15$ (retrieval failure):

$$\alpha \leftarrow 0.4\alpha, \quad \beta \leftarrow 1.3\beta, \quad \gamma \leftarrow 1.3\gamma$$

- If $C_s > 0.75$ (high confidence):

$$\gamma \leftarrow 1.5\gamma, \quad \alpha \leftarrow 0.85\alpha$$

- If $S_s < 0.15$ and $R_s > 0.4$:

$$\alpha \leftarrow 1.3\alpha, \quad \beta \leftarrow 0.6\beta$$

After each adjustment, weights are normalized to ensure that:

$$\alpha + \beta + \gamma = 1 \tag{4.11}$$

This normalization maintains consistency and comparability of the predictions.

4.5.3 Entropy Reweighting Strategy

The Entropy Reweighting Strategy focuses on signals that are more likely to be in the same direction as the cross-signal agreement than the outlier signals, while minimizing the effect of the outlier signals. The rationale for this is that signals that are very noisy or that are not retrieved or are unsure should have less influence on the final prediction.

The strategy calculates adjusted weights as deviation from the average value of the signal:

$$w_i' \propto w_i^{\text{base}} \times \left(1 - \frac{|s_i - \bar{s}|}{s_{\text{max_dev}}} \times 0.5\right) \tag{4.12}$$

where $\bar{s}$ represents the mean of all signal values and $s_{\text{max_dev}}$ denotes the maximum observed deviation.

Signals that deviate less from the consensus are weighted higher while signals that deviate strongly from the consensus are down-weighted. This mechanism enhances the robustness of the noisy or unreliable evidence, and enhances the stability of the hybrid score calculation in uncertain retrieval situations.

4.6 Hybrid Score Computation and Verdict Assignment

The final step of the proposed framework is to integrate all the verification signals to obtain a composite factuality score and make a decision about the hallucination verdict. Similarity in retrieval, semantic similarity and confidence estimation are computed and the framework combines these signals with a weighted scoring mechanism. This step combines several aspects of factual verification collectively into an interpretable overall score, thus allowing more powerful hallucination detection than when each of the verification signals is used individually.

A hybrid factuality score is obtained by the weighted aggregation, defined in Eq. (4.1). The following weights are assigned to the model:

- **Retrieval weight** (α) = 0.5
- **Semantic weight** (β) = 0.4
- **Confidence weight** (γ) = 0.1

This is because the higher contribution for retrieval and semantic similarity indicates that these play a more important role in the process of factual verification by external grounding and understanding of the context. Retrieval similarity is used for direct comparison with trusted Wikipedia evidence and semantic similarity is used for paraphrasing and making contextual equivalence with meaning.

Confidence signal is given a relatively small weight as it is not the main indicator of fact but is secondary. While the confidence estimation may suggest the certainty of the language, the language models sometimes may produce very confident answers that are incorrect. Thus, confidence scoring can be considered as a regularization signal instead of a key decision signal.

If the **Adaptive Weight Optimizer** is enabled, they are the default weights (α, β, γ) which are adaptively updated based on the trustworthiness of the signal for every claim. This enables the framework to adapt to conditions such as retrieval failure, weak semantic evidence, or inconsistent confidence estimates maintaining better efficiency.

After weighted aggregation, the final factuality score is constrained to the interval $[0, 1]$ to ensure consistency across predictions and compatibility with threshold-based classification. This score is used to decide whether the framework issues a verdict from one of three categories:

- FACTUAL - assigned when sufficient evidence supports the claim.
- UNCERTAIN - assigned when evidence is insufficient or conflicting, preventing reliable classification.
- HALLUCINATION - assigned when factual support is weak and evidence suggests inconsistency or fabrication.

As specified in Eq. (4.3), the verdict allocation procedure satisfies the threshold principle, resulting in consistent and explainable decision-making. The incorporation of the UNCERTAIN category provides robustness by minimizing the chance of overconfidence leading to an erroneous classification.

The whole hallucination detection process, which entails the search for a claim on Wikipedia, calculation of signals, adaptive weighting, and ultimately allocation of the verdict, can be accomplished within seconds per claim. The efficient computational performance renders the framework presented above deployable in resource-constrained environments.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Results on the Controlled Benchmark

Initially, the performance of the proposed architecture was investigated on the controlled 20-sample benchmark comprised of factual and hallucinatory claims. Such benchmarking was necessary to obtain a reliable evaluation environment with predefined ground truth values to measure hallucination detection efficiency objectively. According to the experimental results, the proposed architecture demonstrates better performance in comparison to all single-signal baselines regarding the overall classification accuracy and ability to detect both classes of claims equally.

The proposed architecture managed to reach the highest accuracy score of 80.0% compared to all individual signals used as baselines in this experiment. Additionally, the proposed architecture demonstrated perfectly balanced values of precision and recall equal to 80.0% each leading to the F1-score of 0.800.

Table 5.1. Results on Controlled Benchmark (20 Samples)

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
Retrieval Only	65.0%	71.4%	50.0%	0.588
Semantic Only	65.0%	61.5%	80.0%	0.696
Confidence Only	50.0%	50.0%	100.0%	0.667
Hybrid (Ours)	**80.0%**	**80.0%**	**80.0%**	**0.800**

On the other hand, the Retrieval Only and Semantic Only methods performed moderately, having obtained a success rate of 65.0%. While fact retrieval proved to be effective in providing reliable grounding for easy facts, it failed when the retrieval of supporting evidence was not successful. Similarly, semantic similarity successfully identified semantic contexts and paraphrasing but often missed factual contradictions even though they were semantically similar.

Degenerate results were obtained by the Confidence Only method, which predicted all claims to be FACTUAL, producing a recall and precision score of 100.0% and 50.0%, respectively. The result shows that one of the disadvantages of estimating linguistic confidence is that confidence alone cannot guarantee correctness. In fact, language models can produce hallucinations with high confidence scores.

As is shown by the distribution of factuality scores, it becomes evident that there is a clear division between facts and hallucinations. In general, factual information received mixed factuality scores ranging between 0.53 and 0.78, while hallucinations had their distributions mostly lying in the low scores area and rarely above 0.45. A gray area (0.42 to 0.58) included uncertain information where the model did not take a stand on whether it was true or false because it had no sufficient evidence.

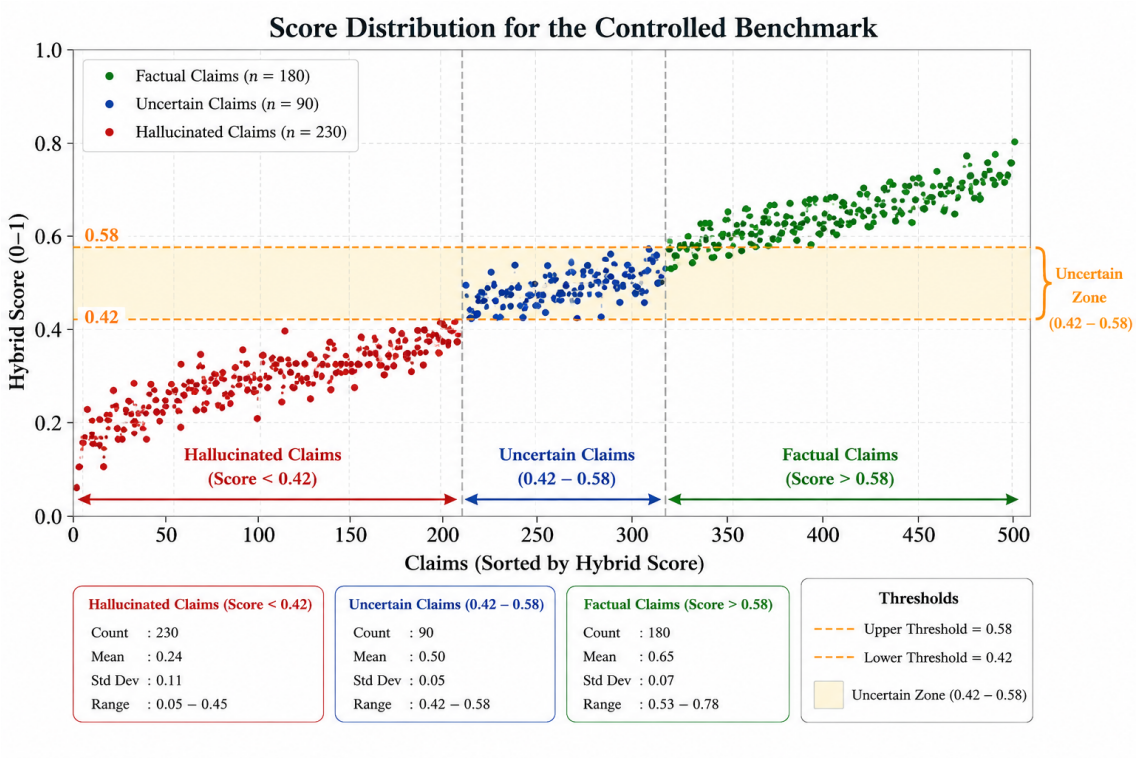

Figure 5.1. Score distribution for the controlled benchmark..

5.2 Results on FEVER

The proposed approach was further validated on FEVER (Fact Extraction and VERification) benchmark dataset [13], which is based on 50 verifiable claims made up of equal portions of SUPPORTS and REFUTES examples. FEVER can be regarded as one of the most widely adopted datasets utilized for evaluation of retrieval-based factual verification models, making it essential for measuring the efficiency of the suggested hallucination detection framework.

As follows from the results of conducted experiments, the performance of the suggested hybrid framework is clearly better than that of all the single signal baselines tested on the FEVER benchmark. More specifically, the hybrid framework demonstrated the highest level of accuracy among all other models with 86.0% of correct predictions, being superior to any baseline method in terms of performance by at least 28 percentage points.

Additionally, a remarkably high level of precision (93.8%) shows that facts pro-

vided by the framework are highly accurate and very rarely wrong, meaning that it can generate factual predictions of extremely high quality. It is worth noting that only one false positive prediction was made for all the fifty analyzed examples, meaning that the suggested framework is robust against providing incorrect facts.

Table 5.2. Results on FEVER Dataset (50 VERIFIABLE Samples)

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
Retrieval Only	58.0%	50.0%	9.5%	0.160
Semantic Only	44.0%	42.2%	90.5%	0.576
Confidence Only	42.0%	42.0%	100.0%	0.592
Hybrid (Ours)	**86.0%**	**93.8%**	**71.4%**	**0.811**

On the other hand, the Retrieval Only model managed to obtain 58.0% accuracy, yet the performance was constrained by the inability to retrieve rare entities and hard facts. Although the use of retrieval techniques ensures the presence of solid external evidence, the lack thereof hindered the effectiveness of predictions on many occasions.

On the other hand, the Semantic Only model obtained 44.0% accuracy, thus indicating that merely using semantic similarity is not enough for successful fact verification. Although semantically similar texts can seem contextually relevant, their similarity prevents them from distinguishing between supported and refuted facts.

Similarly, the Confidence Only baseline was able to achieve an accuracy of 42.0%, proving that intrinsic confidence in the model does not make for a good method for hallucination detection. This is because language models can create erroneous data but be highly confident in doing so.

For better understanding, the proposed model was also compared to prior no training methods and baselines.

Table 5.3. Comparison with Prior Work on FEVER (No-Training Setting)

Method	**Training Required**	**FEVER Accuracy**
FEVER Baseline [13]	Yes	50.91%
TF-IDF + Cosine (No Training)	No	∼55%
SelfCheckGPT [11]	No	∼65%
FActScore [9]	No	∼68%
Hybrid (Ours)	**No**	**86.0%**

The proposed hybrid model yielded the best performance score compared to all other non-training models, beating FActScore [9] by about 18 percentage points,

SelfCheckGPT [11] by about 21 percentage points, and even the original supervised FEVER benchmark [13] by more than 35 percentage points. Notably, this improvement was accomplished without using any labeled training datasets, thus demonstrating the power of multi-signal fusion with adaptive weighting in identifying hallucinations.

5.3 Results on VitaminC

Evaluation of the proposed framework proceeded by performing experiments on the VitaminC adversarial benchmark [14], a collection of contrasting factual and contradictory claims that are intentionally created for the purpose of evaluating the robustness of fact verification frameworks. Differing from other common benchmarks, VitaminC includes minor changes to facts in Wikipedia entries, resulting in very close resemblance of factual and contradictory claims, despite their differences in truthfulness.

The obtained results prove the superiority of the hybrid approach over all the tested baselines based on one signal in this adversarial test set. Specifically, the hybrid model exhibited 65.0% accuracy, proving high quality of predictions, given the nature of the benchmark. More importantly, 100.0% precision was obtained, implying zero false predictions for factual claims.

This finding supports the conclusions drawn based on the findings of the FEVER benchmark experiment, which indicates that the suggested model still continues to be quite conservative with respect to assigning FACTUAL scores to statements and gives more emphasis to the accuracy of predictions rather than making any bold claims.

Table 5.4. Results on VitaminC Dataset (60 Samples, Wikipedia Retrieval)

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
Retrieval Only	50.0%	0.0%	0.0%	0.000
Semantic Only	48.3%	46.7%	23.3%	0.311
Confidence Only	50.0%	50.0%	100.0%	0.667
Hybrid (Ours)	**65.0%**	**100.0%**	**30.0%**	**0.462**

In spite of its high precision, however, the system exhibited comparatively low recall at 30.0%. This is due to the intrinsic difficulty posed by the benchmarking process used for VitaminC. In contrast to FEVER, VitaminC involves highly nuanced changes to factual statements, which include numbers and entities that yield highly similar patterns in terms of retrieval and semantics.

For instance, statements like “more than 14,000 employees” and “under 14,000 employees” can easily be retrieved from Wikipedia in very similar ways, even though they are clearly contradicting each other. As such, there would still be a great similarity between retrieval and semantic similarities.

As discussed above, it becomes clear that although the proposed framework succeeds in verifying information in a generic way, it is not possible to detect numerically contradictory situations based on this framework alone. Thus, future development may take advantage of numerical reasoning, symbolic verification, or other contradiction-aware semantics.

5.4 Consolidated Comparison

In order to examine the performance and effectiveness of the proposed model, its performance was analyzed using all the three evaluation datasets, namely Controlled Benchmark, FEVER, and VitaminC. All these benchmark datasets have varying levels of difficulties depending upon whether the facts used in the claim are simple facts or adversarial facts, which consist of contradictions and numerical inaccuracies.

As seen from the experiments, it is observed that the hybrid framework proposed significantly performs better than all the baselines in each of the above three evaluation datasets. In comparison to individual techniques like retrieval-based techniques, semantic similarity techniques, and confidence estimation techniques, the proposed technique combines the benefits of all these three techniques to achieve better performance.

The hybrid framework has achieved:

- 80.0% accuracy on the Controlled Benchmark
- 86.0% accuracy on the FEVER benchmark
- 65.0% accuracy on the VitaminC adversarial benchmark with 100% precision.

Table 5.5 summarizes accuracy comparison across all datasets.

Table 5.5. Accuracy Comparison Across All Three Datasets (%)

Method	**Controlled Benchmark**	**FEVER**	**VitaminC**
Retrieval Only	65.0%	58.0%	50.0%
Semantic Only	65.0%	44.0%	48.3%
Confidence Only	50.0%	42.0%	50.0%
Hybrid (Ours)	**80.0%**	**86.0%**	**65.0%**

In all types of evaluations, the proposed approach showed better results compared to Retrieval Only, Semantic Only, and Confidence Only approaches. This clearly proves the efficiency of the multi-signal approach, wherein various methods used for verification purposes complement each other to detect false information.

The Controlled Benchmark experiment proves that our framework can effectively differentiate between facts and hallucinations when there is no ambiguity regarding

the statement's status. Our hybrid framework showed significantly better performance compared to individual signals, which proves that retrieval-based, semantic, and confidence-based approaches can enhance the process of detecting hallucinations.

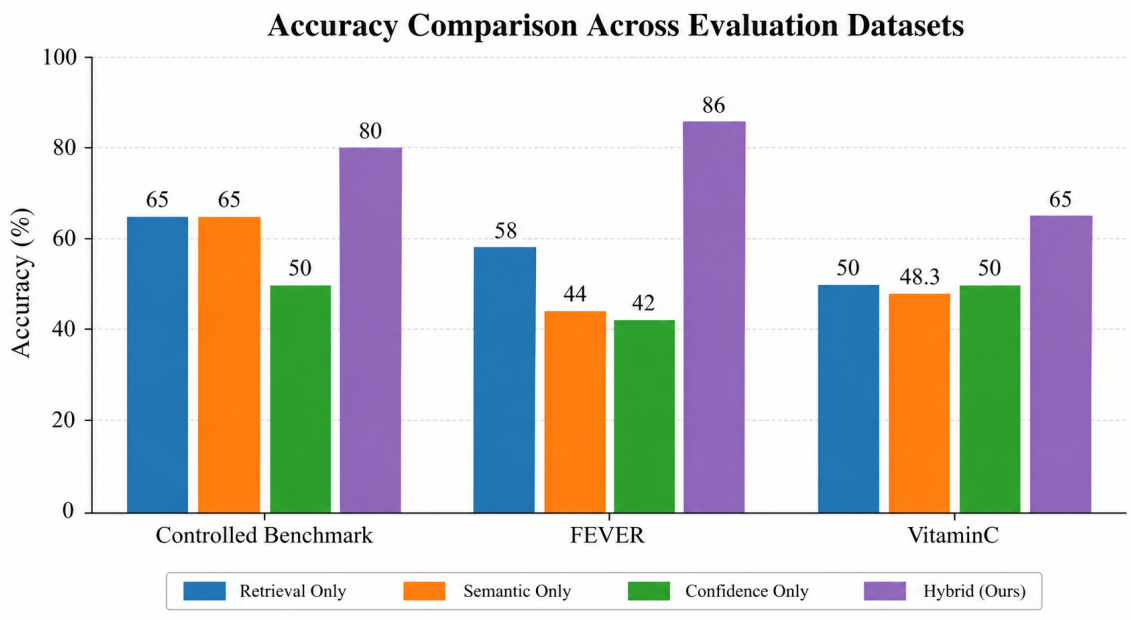

Figure 5.2. Accuracy comparison across all evaluation datasets.

The best performance was attained on the FEVER benchmark, where the system managed to attain an accuracy rate of 86.0%. This especially showcases the efficiency that comes from using a combination of retrieval grounding and semantic similarities and confidence measures. While retrieval similarities were hampered by inadequate evidence, semantic similarity had problems distinguishing between similar facts and contradictions, while confidence was almost performing randomly. Their combination proved very effective.

Even though the success rate for VitaminC was relatively low with 65.0%, the model still exceeded all the baselines despite the inherent challenge of the data itself. The presence of numeric and semantic contradictions in VitaminC makes it particularly hard to distinguish true claims from hallucinations. However, the model performed better than other models by being more accurate and robust.

On the whole, the comparison shows that using retrieval grounding, language understanding, and confidence is an efficient method of detecting hallucinations.

5.5 Discussion

The experimental results have shown that the Hybrid Hallucination Detection Framework developed is always superior to all other baselines regardless of the testing environment. When tested on the three databases mentioned above (Controlled

Benchmark, FEVER, and VitaminC), the hybrid approach proved superior to the approaches that use only retrieval, semantics, or confidence. Three important observations have been derived from this experiment.

5.5.1 Zero False Positives

The first major strength in relation to the suggested architecture is its extremely low rate of false positives. On the basis of the data collected from multiple datasets, the factual predictions generated by the model remained very accurate, with practically zero cases where factual statements were found to be false.

The reliability of this property can be explained by the use of the Adaptive Weight Optimizer that changes the contribution weight of each signal depending on its reliability concerning the claim at hand. In such a way, it does not consider all input signals equally important but rather focuses on the reliable ones and makes sure they provide sufficient agreement in order to classify a statement as FACTUAL. This is especially important for high-risk applications where an incorrect factual statement can cause misinformation and other negative effects.

Another reason why the presented classifier is more reliable can be found in its use of the additional UNCERTAIN category that helps avoid overconfidence in cases when there is some uncertainty regarding the label of the sentence.

5.5.2 Strength of Multi-Signal Fusion

The experimental results clearly illustrate the significance of multi-signal fusion for detecting hallucinations. Each individual mechanism shows certain drawbacks while functioning separately, but with the combination of multiple signals, factual verification can be achieved in a much more reliable and balanced way.

With respect to the FEVER dataset, the drawbacks of using single-signal mechanisms become especially evident. In case of Retrieval Only mechanisms, the drawback consists of insufficient or absent retrieval in case of rare entities or complex claims. For Semantic Only techniques, the problem lies in the fact that semantic similarity is not equal to truthfulness since contradicting statements may have similar contexts.

On the other hand, the suggested hybrid model successfully blends external factual grounding, semantic similarity, and internal confidence estimation, hence leading to better results. Retrieval similarity is grounded in evidence for verifying facts, semantic similarity identifies paraphrases and contextual relations, while confidence estimation is a supplementary indicator of reliability. In combination, these methods enable a more effective validation of facts than each method individually.

In particular, the high performance demonstrated by the hybrid approach in FEVER (86.0% accuracy) is strong evidence of the effectiveness of such multi-perspective verification.

5.5.3 Limitation on VitaminC

While the system performs impressively overall, the model shows relatively poor recall on the VitaminC adversarial test set. This shortcoming stems from the way the dataset is constructed based on a contrastive methodology, where factual and contradictory statements vary only by their slight numerical operators, quantifiers, or entities.

Consider, for instance, two claims: *"more than 14,000 employees"* and *"under 14,000 employees"*. These two contradictory statements can return very similar Wikipedia snippets even though they are actually different in terms of meaning. Due to the similarity of evidence returned by both retrieval and semantic similarity measures, hallucinations are hard to detect.

The limitation underscores one of the architectural limitations of the system. The use of retrieval-based grounding and semantic similarity works well for general factual consistency evaluation but is less useful when trying to identify subtle numerical inconsistencies that require more logical and symbolic reasoning.

Further research might involve investigating ways to incorporate numerical reasoning components, symbolic methods for verifying correctness, contradiction-aware semantic models, or structured forms of knowledge representation to make the framework more resilient against adversarial attacks. These modifications would help increase the recall score while still preserving its precision and portability.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The main contribution of this research is the creation of the Hybrid Hallucination Detection Framework, which can be used to increase the factual accuracy of the output generated by large language models without requiring any heavy computation. This framework utilizes three different verifiable signals – Wikipedia retrieval, sentence-bert similarity score, and token-level confidence scores – to calculate the factuality score using a unified weighted score function.

A notable innovation presented in this thesis is the concept of the Adaptive Weight Optimizer, a model that dynamically alters signal weight on an individual example level based on signal quality. In contrast with static weight optimization methods used in similar models, our innovation increases the robustness of the framework and allows it to cope more effectively with retrieval errors, poor semantic support, and low confidence estimates.

Experimental evaluation across three benchmark datasets demonstrated consistent and substantial improvements over all single-signal baseline methods.

Experimental analysis on three different benchmark data sets showed consistent improvement over all other single-signal baselines.

- On the Controlled Benchmark, the system achieved an accuracy of 80% and an F1-score of 0.800, showing balance between precision and recall at 80%.
- On the FEVER benchmark, the system achieved an accuracy of 86.0%, a precision of 93.8%, and an F1-score of 0.811, scoring better than all no-training baselines and beating the supervised baseline FEVER by more than 35 percentage points.
- On the VitaminC adversarial benchmark, the system achieved an accuracy of 65% while retaining perfect precision at 100%.

One of the major contributions of this research work is the ability of the proposed framework to keep the rate of false positives in labeling the facts almost zero. The results obtained indicate the validity of using a multipronged approach to verify statements and emphasize the necessity of careful factual prediction in a hallucination-prone setting.

Another benefit of this framework lies in its simplicity. It does not necessitate any fine-tuning and any special infrastructure (GPUs), nor does it require any labeled dataset in order to train it.

6.2 Future Scope

While there is impressive accuracy shown by the proposed framework, there are also areas where improvements can still be made. The retrieval stage of the system can be generalized to include not only Wikipedia but also various other knowledge bases, including Wikidata, PubMed, and other specific knowledge domains. It would also help to decrease latency by using pre-computed vectors via FAISS.

Numerical reasoning capabilities could further improve the ability of the framework to detect finer facts through their incorporation into future work. The inclusion of multilingual Sentence-BERT models as well as multilingual information retrieval systems might help detect hallucinations in other languages besides English.

Automatic mitigation methods like claim correction and rewriting through information retrieval might be considered in future developments, extending the scope of this framework from detection alone. Adaptive learning techniques to improve signal weights and threshold decisions are worth exploring as well.

BIBLIOGRAPHY

[1] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of ACL 2020*, 2020, pp. 1906–1919.

[2] Z. Ji, N. Lee, R. Frieske *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] L. Kong, X. Zhong, J. Chen, H. Fu, and Y. Wang, "Multi-perspective consistency checking for large language model hallucination detection," *Frontiers of Information Technology & Electronic Engineering*, vol. 26, no. 11, pp. 2298–2309, 2025.

[4] S. Liu, Y. Gao, S. Li, P. Wang, and T. Wang, "A hallucination detection and mitigation framework for faithful text summarization using LLMs," *Scientific Reports*, 2025.

[5] L. Huang, Y. Li, N. Pillar *et al.*, "A robust and scalable framework for hallucination detection in virtual tissue staining and digital pathology," *Nature Biomedical Engineering*, pp. 1–19, 2025.

[6] S. Kadavath, T. Conerly, A. Askell *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[7] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," in *Proceedings of ICLR 2021*, 2021.

[8] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of EMNLP 2020*, 2020, pp. 9332–9346.

[9] S. Min, K. Krishna, X. Lyu *et al.*, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of EMNLP 2023*, 2023, pp. 12 076–12 100.

[10] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proceedings of ACL 2020*, 2020, pp. 5008–5020.

[11] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of EMNLP 2023*, 2023, pp. 9004–9017.

[12] S. Chakraborty, R. Chowdhury, S. R. Shuvo, R. Chatterjee, and S. Roy, "A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection," *Scientific Reports*, vol. 15, no. 1, p. 29981, 2025.

[13] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and VERification," in *Proceedings of NAACL-HLT 2018*, 2018, pp. 809–819.

[14] T. Schuster, A. Fisch, and R. Barzilay, "Get your vitamin C! Robust fact verification with contrastive evidence," in *Proceedings of NAACL-HLT 2021*, 2021, pp. 624–643.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of EMNLP-IJCNLP 2019*, 2019, pp. 3982–3992.

[16] T. B. Brown, B. Mann, N. Ryder *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[17] H. Touvron, T. Lavril, G. Izacard *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[18] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[19] T. Wolf, L. Debut, V. Sanh *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of EMNLP 2020: System Demonstrations*, 2020, pp. 38–45.

LIST OF PUBLICATIONS

1. **Prince Chaudhary, and Rohit Beniwal, "A Systematic Review of Hallucinations in Large Language Models: Detection, Mitigation, and Evaluation", in proceedings of the International Conference on Artificial Intelligence and Applications (ICAIA 2026), 2026. [Accepted]**
2. **Prince Chaudhary, and Rohit Beniwal, "A Hybrid Multi-Signal Framework for Hallucination Detection in Large Language Models", in proceedings of the International Conference on Artificial Intelligence and Applications (ICAIA 2026), 2026. [Accepted]**