

A Trust-Aware Framework for Hallucination Detection and Accuracy Paradox Analysis in Large Language Models

Mr. Vedeshvar L

vedesh0212@gmail.com

Rathinam Technical Campus

Dr. Sudha SV

Rathinam Technical Campus

Research Article

Keywords: Trust-Aware AI, Hallucination Detection, Accuracy Paradox, Multi-Model Evaluation, Trust Calibration, Adversarial Robustness

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10408506/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Trust-Aware Framework for Hallucination Detection and Accuracy Paradox Analysis in Large Language Models

Vedeshvar L¹, Dr. S. V. Sudha²

¹Assistant Professor, Department of Computer Science and
Engineering, Rathinam Technical Campus,
Coimbatore, Tamil Nadu, India.

Vedesh0212@gmail.com, vedeshvar.soc@rathinam.in

²Professor and Dean, School of Computing Rathinam Technical
Campus, Coimbatore, Tamil Nadu, India.
svsudha.mvenki@gmail.com

Abstract

Large language models (LLMs) have achieved remarkable success in natural language understanding and generation, enabling applications across scientific research, healthcare, education, and decision support systems. Despite these advances, LLMs frequently generate fluent but factually incorrect responses, a phenomenon commonly referred to as hallucination. Such behavior raises critical concerns regarding reliability, accountability, and safe deployment of generative AI systems. Existing approaches to mitigate hallucinations including retrieval-augmented generation, multi-agent debate frameworks, and confidence calibration techniques primarily focus on improving response quality but often lack a systematic mechanism for estimating the trustworthiness of model outputs. In particular, many current methods fail to address the problem of overconfident incorrect predictions, where models produce highly confident answers despite limited or inaccurate knowledge, a phenomenon we refer to as the accuracy paradox.

In this research paper, we introduce **TrustSLM**, a trust-aware evaluation framework designed to assess and regulate responses generated by multiple large language models. The proposed framework integrates outputs from multiple base models and evaluates them using a set of reliability metrics that capture semantic consensus, response uncertainty, evidence consistency, answerability, manipulation signals, and overconfidence detection. These metrics are aggregated into a trust feature representation that is processed by a lightweight classifier to estimate an overall trust score for each generated response. Based on this trust score, the system

determines whether the response should be accepted, hedged with uncertainty, or rejected through a controlled abstention policy.

We evaluate the proposed framework using multiple prompt datasets and compare it against baseline approaches including single-model generation and majority voting across models. Experimental results demonstrate that TrustSLM significantly improves hallucination detection and trust calibration while reducing overconfident incorrect responses. The proposed architecture introduces a trust-aware control layer for large language model systems and contributes toward safer, more interpretable, and reliable deployment of generative AI technologies in high-stakes domains.

Keywords: Trust-Aware AI, Hallucination Detection, Accuracy Paradox, Multi-Model Evaluation, Trust Calibration, Adversarial Robustness

Introduction

Recent advances in large language models (LLMs) have significantly transformed the field of natural language processing and artificial intelligence. Transformer-based architectures trained on massive datasets have demonstrated remarkable performance across a wide range of tasks, including text generation, question answering, reasoning, summarization, and dialogue systems. These models have enabled the development of intelligent assistants capable of generating human-like responses and assisting users in various domains such as education, healthcare, research, and decision support systems.

Despite these impressive capabilities, the reliability of language model outputs remains a major concern. One of the most widely discussed challenges associated with modern generative models is the phenomenon known as *hallucination*. In the context of language models, hallucination refers to the generation of information that appears plausible but is factually incorrect, unsupported by evidence, or inconsistent with known knowledge. Unlike traditional errors in classification systems, hallucinated outputs often exhibit fluent linguistic structure and confident tone, making them difficult for users to identify as incorrect. As a result, hallucinations pose significant risks when language models are deployed in environments where accuracy and accountability are essential.

The problem becomes even more critical when language models are used in high-stakes domains such as healthcare, legal analysis, financial advisory systems, and scientific knowledge discovery. In such scenarios, incorrect or fabricated information may lead to misleading conclusions, poor decision making, or erosion of trust in AI technologies. Consequently, ensuring the reliability and trustworthiness of generative AI systems has become a central research challenge in the field of machine learning.

One important factor contributing to hallucinations is the training objective of language models. Most generative models are trained to predict the next token in a sequence based on the statistical patterns present in large-scale text corpora. While this objective enables models to produce coherent and contextually appropriate text, it does not explicitly enforce factual correctness or knowledge verification. When confronted with questions that fall outside their training distribution or require precise factual knowledge, language models may generate plausible responses based on linguistic patterns rather than grounded information.

Another critical issue is the tendency of language models to produce responses with high confidence even when their predictions are incorrect. This mismatch between confidence and correctness reflects a broader phenomenon of miscalibration in machine learning models. In the context of language models, this behavior leads to what can be described as an accuracy paradox, where the model appears confident and authoritative while providing inaccurate or unsupported information. The presence of such overconfident errors complicates the task of evaluating the reliability of generated responses and highlights the need for mechanisms that explicitly assess trustworthiness.

Several approaches have been proposed in recent years to address the problem of hallucinations in language models. Retrieval-augmented generation techniques attempt to incorporate external knowledge sources into the generation process in order to improve factual grounding. Multi-agent reasoning frameworks enable multiple language models to collaborate or debate in order to improve reasoning quality. Other methods focus on uncertainty estimation, confidence calibration, or external verification tools. While these approaches provide valuable improvements, they often focus on enhancing generation itself rather than introducing a systematic mechanism for evaluating the reliability of generated responses before they are delivered to users.

Furthermore, many existing methods rely heavily on a single model's internal signals, which may still be misaligned with factual correctness. For example, a language model may produce a high probability score for a response that is syntactically coherent but factually incorrect. This limitation suggests that evaluating reliability solely based on model-internal metrics may be insufficient for identifying hallucinated outputs.

An alternative strategy involves comparing responses generated by multiple language models. When independent models produce similar responses to the same prompt, this agreement may provide a signal of higher reliability. Conversely, disagreement among models may indicate uncertainty or lack of knowledge. Multi-model consensus can therefore serve as a useful indicator for detecting unreliable responses. However, consensus alone cannot guarantee correctness, as multiple models may still share similar biases or knowledge gaps. A comprehensive evaluation framework must therefore incorporate additional signals beyond model agreement.

To address these challenges, this paper proposes TrustSLM, a trust-aware evaluation framework designed to assess and regulate responses generated by multiple large language

models. The core idea of the proposed framework is to introduce an explicit trust evaluation layer that operates between the generation stage and the final response delivered to the user. Rather than relying solely on the output of a single model, the framework aggregates responses from multiple language models and analyzes them using a set of reliability metrics.

These trust metrics capture several aspects of response reliability, including semantic consensus between model outputs, response uncertainty, evidence consistency, answerability estimation, and detection of overconfidence signals. The metrics are aggregated into a feature representation that is processed by a lightweight classifier responsible for estimating an overall trust score for each generated response. Based on this trust score, the system determines whether the response should be accepted as reliable, hedged with uncertainty statements, or rejected through an abstention mechanism.

The introduction of an abstention policy is particularly important in safety-critical environments. Rather than forcing the system to provide an answer in all situations, the proposed framework allows the model to decline answering when the reliability of the response falls below a predefined threshold. This strategy aligns with the principle that in certain situations, providing no answer may be preferable to delivering potentially incorrect information.

In addition to improving reliability, the TrustSLM framework also provides transparency by exposing trust indicators that can help users understand the reasoning behind system decisions. By combining multiple trust metrics and providing interpretable outputs, the framework aims to enhance user confidence in AI systems and facilitate responsible deployment of generative technologies.

The contributions of this work can be summarized as follows. First, we introduce a trust-aware evaluation framework for large language model outputs that integrates multiple reliability signals into a unified trust scoring mechanism. Second, we propose an approach for detecting the accuracy paradox, identifying cases in which models exhibit high confidence despite limited evidential support. Third, we develop a decision policy that determines whether responses should be accepted, hedged, or rejected based on estimated trust scores. Finally, we evaluate the proposed framework through experiments comparing it with baseline approaches including single-model generation and simple consensus strategies.

The remainder of this paper is organized as follows. Section 2 reviews related work on hallucination detection, reliability estimation, and trust evaluation in language models. Section 3 describes the architecture of the proposed TrustSLM framework and its core components. Section 4 presents the mathematical formulation of the trust metrics and scoring mechanism. Section 5 details the experimental setup, datasets, and evaluation metrics used in this study. Section 6 reports experimental results and provides quantitative and qualitative analysis. Section 7 discusses the implications, limitations, and potential extensions of the proposed approach. Finally, Section 8 concludes the paper and outlines directions for future research.

1.1 Background and Motivation

Recent advances in large language models (LLMs) have transformed the field of natural language processing by enabling machines to generate coherent, contextually relevant, and human-like text. Transformer-based architectures such as GPT-style models and encoder–decoder systems have demonstrated strong performance across tasks including question answering, summarization, translation, and reasoning. As these systems are increasingly deployed in real-world applications ranging from medical decision support to scientific research assistance the reliability of their outputs has become a critical concern.

Despite their impressive capabilities, LLMs frequently generate responses that appear fluent and convincing but are factually incorrect. This phenomenon, commonly referred to as *hallucination*, occurs when models produce information that is unsupported by evidence or inconsistent with known facts. Because such outputs are often expressed with high confidence and linguistic coherence, users may mistakenly interpret them as reliable.

1.2 The Hallucination Problem in Language Models

Hallucinations arise from several factors inherent in the design of modern generative models. Language models are trained to predict the next token in a sequence rather than to verify factual correctness. Consequently, they may generate plausible but incorrect statements when faced with incomplete knowledge or ambiguous prompts. Hallucinated responses can take multiple forms, including fabricated citations, incorrect factual statements, logical inconsistencies, and invented entities.

In high-stakes domains such as healthcare, legal analysis, and scientific communication, hallucinated responses can lead to serious consequences. A model that confidently generates incorrect information may mislead users, undermine trust in AI systems, and introduce risks in decision-making processes.

1.3 Overconfidence and the Accuracy Paradox

A particularly concerning aspect of hallucination behavior is the tendency of language models to express incorrect answers with high confidence. This miscalibration between model confidence and factual correctness leads to what can be described as an accuracy paradox, where the apparent certainty of the model masks underlying inaccuracies. In such cases, traditional confidence metrics are insufficient for determining whether a response should be trusted.

The accuracy paradox highlights the need for mechanisms that go beyond simple confidence estimation and instead evaluate responses using multiple reliability signals. Identifying situations in which a model appears confident but lacks supporting evidence is essential for building trustworthy AI systems.

1.4 Limitations of Existing Hallucination Mitigation Techniques

Several approaches have been proposed to address hallucinations in language models. Retrieval-augmented generation integrates external knowledge sources into the generation process to improve factual grounding. Multi-agent reasoning frameworks allow multiple models to debate or critique each other's responses in order to improve reasoning quality. Other approaches focus on confidence calibration, uncertainty estimation, or tool-based verification.

While these methods provide valuable improvements, they often suffer from important limitations. Retrieval-based approaches depend heavily on the quality and coverage of external knowledge bases. Multi-agent reasoning systems may introduce significant computational overhead and do not always guarantee reliable outcomes. Confidence calibration techniques typically rely on model-internal signals that may still be misaligned with factual correctness.

More importantly, most existing approaches attempt to improve generation itself rather than explicitly evaluating the trustworthiness of generated responses.

1.5 The Need for Trust-Aware Evaluation Frameworks

To ensure the safe deployment of language models, it is essential to introduce mechanisms that explicitly evaluate the reliability of model outputs before presenting them to users. Rather than relying solely on the internal confidence of a single model, a trust-aware evaluation framework should incorporate multiple signals reflecting the consistency, plausibility, and evidential support of generated responses.

Such a framework should be capable of:

- detecting hallucinated responses,
- identifying overconfident predictions,
- estimating response reliability,
- and determining when a model should refrain from answering.

Introducing a trust evaluation layer between model generation and user interaction provides a practical pathway toward safer AI systems.

1.6 Multi-Model Consensus as a Reliability Signal

One promising approach for improving reliability involves comparing responses generated by multiple language models. When independent models produce consistent answers to the same prompt, this agreement may serve as a signal of higher reliability. Conversely, significant disagreement among models may indicate uncertainty or lack of knowledge.

Multi-model consensus can therefore serve as a useful mechanism for detecting potentially unreliable outputs. However, consensus alone is insufficient; models may still agree on incorrect information if they share similar biases or training data limitations. A comprehensive trust evaluation system must therefore combine consensus analysis with additional reliability indicators.

1.7 Research Objectives

The primary objective of this work is to develop a trust-aware framework for evaluating the reliability of responses generated by large language models. Specifically, this research aims to answer the following questions:

1. How can multiple signals be integrated to estimate the trustworthiness of LLM outputs?
2. How can systems detect cases where models are overconfident despite limited knowledge?
3. How can AI systems determine when to answer, hedge uncertainty, or abstain entirely?

Addressing these questions is essential for improving the safety and reliability of generative AI systems.

1.8 Overview of the TrustSLM Framework

To address the challenges described above, we propose TrustSLM, a trust-aware evaluation framework designed to assess and regulate responses generated by multiple language models. The proposed architecture integrates outputs from several base models and evaluates them using a set of trust metrics that capture semantic agreement, uncertainty signals, evidence consistency, answerability, and manipulation indicators.

These metrics are combined into a feature representation that is processed by a lightweight classifier responsible for estimating an overall trust score. Based on this trust score, the system determines whether the generated response should be accepted, hedged with uncertainty, or rejected through an abstention policy.

1.9 Contributions of This Work

The contributions of this paper can be summarized as follows. First, we introduce TrustSLM, a trust-aware evaluation framework designed to assess the reliability of responses generated by large language models. Unlike conventional approaches that rely solely on a single model’s internal confidence signals, the proposed framework integrates outputs from multiple language models and evaluates them using a combination of trust metrics that capture semantic agreement, uncertainty, and evidential consistency. Second, we address the problem of overconfident incorrect predictions, referred to in this work as the *accuracy paradox*, by developing mechanisms that identify situations in which language models produce highly confident responses despite limited or unreliable supporting evidence. Third, the framework proposes a multi-metric trust scoring mechanism that combines signals such as model consensus, response uncertainty, evidence verification, answerability estimation, and manipulation detection into a unified feature representation used for trust estimation. Fourth,

we introduce a decision policy based on controlled abstention, allowing the system to determine whether a generated response should be accepted, hedged with uncertainty, or rejected entirely when the estimated trust level falls below a predefined threshold. Finally, we conduct a comprehensive empirical evaluation comparing the proposed framework with baseline generation strategies, demonstrating that the TrustSLM architecture improves hallucination detection, reduces overconfident incorrect responses, and provides a more reliable mechanism for evaluating large language model outputs.

1.10 Practical Implications

The TrustSLM framework provides a practical mechanism for improving the reliability of generative AI systems in real-world applications. By introducing a trust evaluation layer between language models and end users, the system can reduce the risk of hallucinated outputs and provide more transparent decision-making.

Such mechanisms are particularly valuable in domains where factual correctness and accountability are essential.

2. Related Work

The rapid advancement of large language models (LLMs) has generated substantial research interest in understanding and mitigating the problem of hallucinations in generative AI systems. Hallucinations occur when language models produce responses that appear coherent and plausible but are factually incorrect, unsupported by evidence, or inconsistent with established knowledge. Early investigations into hallucination behavior have suggested that this phenomenon arises largely from the probabilistic training objective of language models, which focuses on predicting the most likely next token in a sequence rather than verifying factual correctness. As a result, models often generate linguistically plausible statements even when they lack sufficient knowledge to produce accurate answers (Kalai and Nissim, 2025; Geng et al., 2024). Subsequent studies have further explored how hallucinations emerge due to limitations in training data coverage, incomplete knowledge representation, and the inability of models to differentiate between factual and speculative information (Passerini et al., 2024; Salehi et al., 2025). These findings have motivated a growing body of research dedicated to detecting and mitigating hallucinated outputs in language model systems.

Several approaches have been proposed to detect hallucinations by verifying the consistency of generated responses with external knowledge sources. Early work in this area relied on rule-based fact-checking systems and structured verification pipelines that compare generated responses with curated knowledge bases. More recent approaches employ neural verification models and semantic similarity techniques to assess whether the generated response aligns with supporting evidence (Feng et al., 2024; Huang et al., 2024). In these approaches, generated responses are analyzed using semantic embeddings and external retrieval mechanisms in order to determine whether they are grounded in verifiable information. Other studies have explored probabilistic uncertainty estimation and token-level confidence analysis as indicators of response reliability, suggesting that high uncertainty in model predictions may signal a higher likelihood of hallucinated outputs (Epstein et al., 2025; Tian et al., 2025). While these methods improve the detection of unreliable responses, they often rely heavily on internal confidence signals or external verification resources that may not always be available. Another important direction in the literature focuses on retrieval-augmented generation (RAG), where generative language models are combined with information retrieval systems that provide relevant contextual knowledge during the generation process. By incorporating external documents or structured knowledge bases, retrieval-

augmented models can ground their responses in verifiable information and thereby reduce hallucination rates. Several studies have demonstrated the effectiveness of retrieval-based approaches in domains such as medical question answering, scientific literature analysis, and clinical decision support (Wada et al., 2025). However, retrieval-based approaches remain dependent on the completeness and accuracy of the underlying knowledge sources. When relevant evidence cannot be retrieved or when retrieved documents contain ambiguous information, the model may still generate incorrect responses.

In parallel with retrieval-based approaches, researchers have explored collaborative reasoning systems in which multiple language models interact in order to refine their responses. Multi-agent frameworks enable models to critique each other's outputs and iteratively improve reasoning quality through structured interactions. Debate-based systems, for example, allow multiple models to generate competing arguments and evaluate alternative explanations before selecting a final response (Sun et al., 2025; Liang et al., 2026). These collaborative approaches have demonstrated promising improvements in reasoning accuracy, as multiple models can collectively identify logical inconsistencies or unsupported claims. Nevertheless, multi-agent systems often introduce increased computational complexity and may still produce incorrect results when participating models share similar biases or knowledge gaps.

Confidence calibration has also been widely studied as a mechanism for improving the reliability of machine learning models. In a well-calibrated model, the predicted confidence of an output should correspond closely to the true probability of correctness. However, numerous studies have shown that modern deep neural networks, including large language models, frequently exhibit significant miscalibration, producing predictions with high confidence even when the outputs are incorrect (Geng et al., 2024; Ryser et al., 2025). This problem becomes particularly concerning in generative AI systems, where overconfident incorrect responses can mislead users into believing that the generated information is reliable. Several techniques have been proposed to address calibration issues, including temperature scaling, Bayesian uncertainty estimation, and ensemble-based confidence estimation (Epstein et al., 2025; Tian et al., 2025). Although these techniques improve the alignment between model confidence and actual correctness, they typically rely on internal model signals and may not capture deeper semantic inconsistencies present in generated text.

Beyond hallucination detection and confidence calibration, recent research has also examined broader behavioral characteristics of language models, including bias, sycophancy,

manipulation susceptibility, and prompt sensitivity. Studies have demonstrated that language models may sometimes generate responses that align with user expectations rather than objective truth, particularly when prompts contain persuasive or leading language (Passerini et al., 2024). Such behavior highlights the importance of evaluating not only factual correctness but also the broader reliability and robustness of generative models. As a result, the concept of trustworthy artificial intelligence has gained increasing attention, emphasizing the need for systems that are reliable, interpretable, and robust under uncertainty (Ryser et al., 2025; Asgari et al., 2025).

Ensemble learning and multi-model consensus mechanisms have also been explored as potential indicators of response reliability. In traditional machine learning systems, combining predictions from multiple models often improves predictive accuracy and robustness. In the context of language models, agreement among independent models may serve as a signal of higher reliability, while disagreement may indicate uncertainty or lack of knowledge. However, consensus alone does not guarantee factual correctness, as multiple models may still share similar training biases or misconceptions.

Despite the significant progress made in hallucination detection, retrieval augmentation, confidence calibration, and collaborative reasoning frameworks, many existing approaches focus primarily on improving the generation process itself rather than explicitly evaluating the trustworthiness of generated responses before they are presented to users. Furthermore, systems that rely solely on internal confidence signals or retrieval-based verification may fail to detect certain classes of hallucinated or overconfident outputs. These limitations highlight the need for frameworks that integrate multiple reliability signals into a unified evaluation mechanism capable of assessing the trustworthiness of language model outputs.

To address this challenge, the present work proposes **TrustSLM**, a trust-aware evaluation framework designed to assess and regulate responses generated by multiple large language models. The proposed framework aggregates responses from several models and evaluates them using a set of reliability metrics that capture semantic consensus, uncertainty estimation, evidence consistency, answerability assessment, and manipulation signals. By integrating these signals into a unified trust scoring mechanism, the proposed approach introduces a systematic method for determining whether a response should be accepted, hedged with uncertainty, or rejected through an abstention policy. This trust-aware evaluation layer

operates independently of the generation process and therefore complements existing approaches that focus on improving generation quality. Through this mechanism, the TrustSLM framework aims to provide a more reliable and interpretable evaluation strategy for large language model outputs, contributing to the broader goal of building safer and more trustworthy generative AI systems.

3. Proposed Methodology: TrustSLM Framework

3.1 Overview of the Proposed Framework

The proposed system, referred to as TrustSLM, introduces a trust-aware evaluation framework designed to assess and regulate the reliability of responses generated by large language models. The primary objective of this framework is to detect hallucinations, identify overconfident responses, and determine whether a generated answer should be accepted, hedged with uncertainty, or rejected entirely. Unlike traditional approaches that rely on a single model's internal confidence scores, the TrustSLM framework incorporates multiple language models and evaluates their outputs using a set of reliability metrics.

The core idea behind TrustSLM is to introduce a trust evaluation layer between the generation stage and the user interaction stage. This trust layer acts as a supervisory component that examines the outputs generated by several language models and determines their reliability before presenting them to the user. By evaluating the agreement between models, analyzing uncertainty patterns, verifying evidence consistency, and detecting overconfidence signals, the system can provide a more robust and transparent evaluation of generated responses.

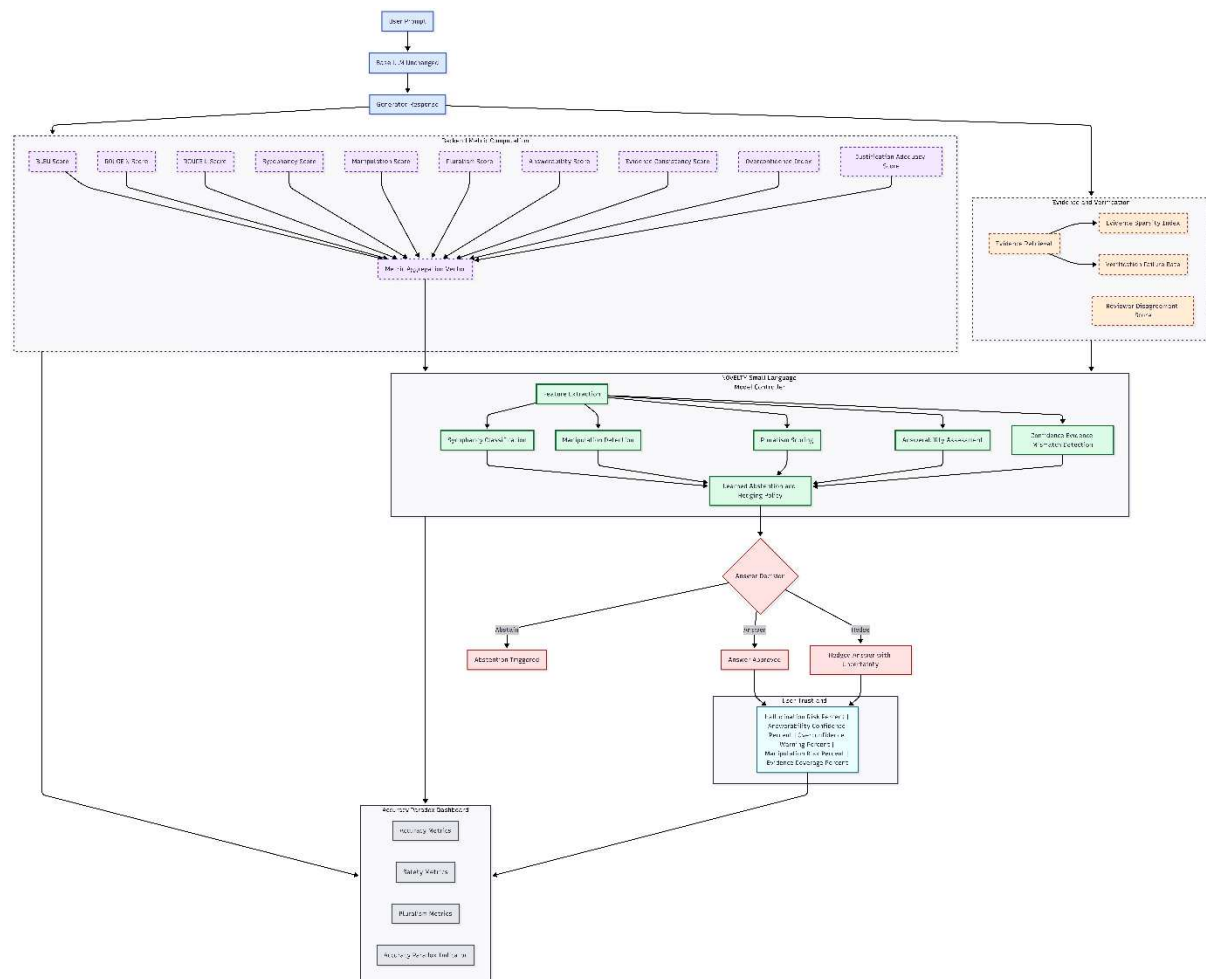
The overall workflow of the framework consists of several stages, including prompt processing, multi-model response generation, response embedding and representation, trust metric computation, feature aggregation, trust classification, and decision-based response delivery. Each stage plays a critical role in evaluating the reliability of generated responses and ensuring that potentially incorrect information is filtered before reaching the end user.

3.2 System Architecture

The architecture of the TrustSLM framework consists of five major components:

1. Prompt Processing Module
2. Multi-LLM Response Generation Module
3. Response Representation and Embedding Module
4. Trust Metric Computation Module
5. Trust Classification and Decision Module

These components work sequentially to evaluate the reliability of responses produced by LLM.



3.1: Architecture of the TrustSLM Framework

Figure 3.1 illustrates the overall architecture of the proposed TrustSLM framework. The system receives a user prompt, generates responses from multiple language models (CHI, PHI, and T5), converts responses into semantic embeddings, computes multiple trust metrics including consensus, uncertainty, evidence consistency, and overconfidence, and finally applies a trust classifier to determine whether the response should be accepted, hedged, or rejected.

3.2.1 Prompt Processing Module

The first stage of the framework involves receiving a user query or prompt. The prompt may originate from various application domains, including general knowledge queries, scientific information requests, reasoning tasks, or domain-specific problem-solving scenarios. The prompt processing module performs initial preprocessing operations, which may include tokenization, normalization, and removal of extraneous characters.

The goal of this stage is to ensure that the prompt is properly formatted and compatible with the input requirements of the language models used in the framework. In some cases, prompts may also be analyzed to determine whether they contain ambiguous or manipulative language that could potentially influence the behavior of language models. Detecting such patterns is useful for identifying prompts that may lead to biased or misleading responses.

3.3 Multi-LLM Response Generation

3.3.1 Motivation for Multi-Model Generation

A key design principle of the TrustSLM framework is the use of multiple language models to generate responses for the same prompt. Relying on a single model for response generation may lead to unreliable outputs, particularly when the model lacks sufficient knowledge about the query. By incorporating multiple models, the framework can analyze the diversity and agreement among generated responses, which provides valuable signals for evaluating reliability.

Multi-model generation also helps mitigate the risk of model-specific biases or limitations. Different models may have been trained on different datasets or may employ different architectural variations, resulting in diverse responses. When multiple independent models produce similar responses, the likelihood of the answer being correct increases.

3.3.2 Base Language Models Used

The TrustSLM framework incorporates several generative language models to produce candidate responses. In the current implementation, the system utilizes models such as:

- CHI model
- PHI model
- T5 transformer model

Each model processes the input prompt independently and generates a response based on its internal language representation.

3.3.3 Response Collection

Once the prompt is processed, it is forwarded simultaneously to all participating language models. Each model generates a response, resulting in a set of candidate responses:

$$\text{Response Set} = \{r_1, r_2, r_3, \dots, r_n\}$$

where r_i represents the response generated by the i -th model.

These responses form the basis for subsequent reliability analysis.

3.4 Response Representation and Embedding

3.4.1 Semantic Embedding of Responses

After responses are generated, the framework converts them into semantic vector representations using sentence transformer models. Sentence embeddings capture the semantic meaning of a piece of text in a numerical vector space. This allows the system to compare the similarity between different responses and analyze the relationships among them.

The embedding process transforms each response into a high-dimensional vector:

$$e_i = \text{Embedding}(r_i)$$

where r_i is the response text and e_i represents its embedding vector.

3.4.2 Importance of Embedding-Based Analysis

Embedding representations enable the system to compute semantic similarity between responses generated by different models. By measuring the similarity between embedding vectors, the framework can determine whether different models are producing similar answers or whether they are generating inconsistent responses.

Embedding-based analysis also provides a robust method for capturing semantic relationships between responses even when the textual phrasing differs.

3.5 Trust Metric Computation

The TrustSLM framework evaluates response reliability using several trust metrics that capture different aspects of model behavior.

3.5.1 Semantic Consensus Score

The semantic consensus score measures the degree of agreement between responses generated by different models. This metric is calculated by computing the average semantic similarity between response embeddings.

If multiple models generate highly similar responses, the consensus score increases, indicating a higher likelihood of correctness. Conversely, significant disagreement among responses suggests uncertainty or lack of knowledge.

3.5.2 Response Uncertainty Metric

The uncertainty metric measures the variation among response embeddings. High variance indicates that models are producing diverse and potentially conflicting responses, suggesting lower reliability.

Uncertainty is estimated by computing the variance of embedding vectors relative to their centroid.

3.5.3 Evidence Consistency Metric

Evidence consistency measures whether the generated response aligns with available supporting information. In this stage, external knowledge sources or retrieval mechanisms may be used to verify the factual correctness of the response.

3.5.4 Answerability Score

The answerability score estimates whether a given prompt can be reliably answered by language models. Certain prompts may require domain-specific expertise or may contain incomplete information.

By identifying such cases, the system can avoid generating unreliable answers.

3.5.5 Overconfidence Detection

Overconfidence detection identifies cases where models produce responses with high confidence despite limited or unreliable evidence. This phenomenon is referred to as the accuracy paradox, where models appear confident but provide incorrect information.

Detecting overconfidence is essential for preventing misleading responses.

3.5.6 Manipulation Detection

Manipulation detection identifies prompts that attempt to bias the language model through persuasive or leading language. Such prompts may encourage the model to generate responses that align with user expectations rather than factual correctness.

3.6 Feature Aggregation

After computing individual trust metrics, the framework aggregates them into a unified feature vector representing the trust state of the response.

Feature Vector = [Consensus, Uncertainty, Evidence Consistency, Answerability, Overconfidence, Manipulation]

This feature vector serves as the input to the trust classification model.

3.7 Trust Classification Model

The TrustSLM framework employs a lightweight machine learning classifier to estimate the trustworthiness of generated responses.

3.7.1 Classifier Selection

In this work, logistic regression is used as the trust classifier due to its simplicity, interpretability, and efficiency.

3.7.2 Training Data

The classifier is trained using labeled examples of reliable and unreliable responses. Training data may include:

- verified correct responses
- hallucinated responses
- partially correct responses

3.7.3 Trust Score Estimation

The classifier processes the feature vector and produces a trust score indicating the reliability of the generated response.

3.8 Decision Layer

The final stage of the TrustSLM framework involves determining how the system should respond to the user based on the estimated trust score.

3.8.1 Accept Response

If the trust score exceeds a predefined acceptance threshold, the response is considered reliable and delivered directly to the user.

3.8.2 Hedge Response

If the trust score falls within an intermediate range, the system provides a hedged response that includes uncertainty indicators.

3.8.3 Abstain Response

If the trust score falls below a rejection threshold, the system refuses to answer the prompt. This abstention mechanism prevents the dissemination of potentially incorrect information.

3.8.4 Adversarial Prompt Detection and Handling

Adversarial prompts pose a significant challenge to the reliability of large language models, as they are intentionally designed to exploit weaknesses in model reasoning, induce hallucinations, or bypass safety constraints. Such prompts often include misleading premises, logical inconsistencies, or instructions that attempt to override system behavior. To ensure

robustness, the proposed TrustSLM framework incorporates an adversarial prompt handling mechanism within the trust evaluation pipeline.

Let p denote the input prompt. An adversarial score $S_{adv}(p) \in [0,1]$ is computed to quantify the likelihood that the prompt is adversarial. This score is derived from a combination of signals including semantic inconsistency, instruction manipulation patterns, answerability estimation, and overconfidence risk. Formally,

$$T(p) = \alpha C(p) + \beta M(p) + \gamma A_s(p) + \delta O_r(p) \quad (3.1)$$

where $C(p)$ denotes semantic contradiction, $M(p)$ captures instruction manipulation patterns, $A_s(p)$ represents answerability, and $O_r(p)$ measures overconfidence risk. The weights satisfy $\alpha + \beta + \gamma + \delta = 1$.

$$T'(p) = T(p) \cdot (1 - S_{adv}(p)) \quad (3.2)$$

where $T(p)$ is the original trust score and $T'(p)$ is the adversarial-adjusted trust score. This formulation ensures that prompts identified as adversarial reduce the effective trust assigned to the generated response.

3.9 Importance of the Proposed Framework

The TrustSLM framework offers several advantages over existing hallucination mitigation approaches.

First, it introduces a trust evaluation layer that operates independently of the generation process. Second, the use of multiple language models enables more reliable consensus analysis. Third, the integration of multiple trust metrics provides a comprehensive evaluation of response reliability. Finally, the decision-based response mechanism ensures that unreliable responses are either hedged or rejected, reducing the risk of misinformation.

3.10 TrustSLM Evaluation Algorithm

The overall workflow of the proposed TrustSLM framework is summarized in Algorithm 1. The algorithm describes the step-by-step process of generating responses from multiple language models, computing trust metrics, incorporating adversarial detection, and producing a final decision based on trust scores.

The proposed algorithm ensures that unreliable or adversarial responses are filtered through a trust-aware mechanism before being presented to the user, thereby improving robustness and reliability.

Algorithm 1: TrustSLM Evaluation Pipeline

Input: Prompt p

Output: Decision $D(p) \in \{\text{Accept}, \text{Hedge}, \text{Abstain}\}$

1. **Receive input prompt p**
2. **Generate responses from multiple models**

$$R = r_1, r_2, \dots, r_k \quad (3.3)$$

3. **Compute semantic embeddings**

$$e_i = \phi(r_i), \forall i \in 1, \dots, k \quad (3.4)$$

4. **Compute trust metrics**

- Semantic consensus $C(p)$
- Uncertainty $U(p)$
- Evidence consistency $E(p)$
- Answerability $A(p)$
- Overconfidence $O(p)$

5. **Construct feature vector**

$$F(p) = [C(p), U(p), E(p), A(p), O(p)] \quad (3.5)$$

Compute trust score

$$T(p) = \sigma(w \cdot F(p)) \quad (3.6)$$

Compute adversarial score

$$S_{adv}(p) = \alpha C(p) + \beta M(p) + \gamma A(p) + \delta O(p) \quad (3.7)$$

Adjust trust score

$$T'(p) = T(p)(1 - S_{adv}(p)) \quad (3.8)$$

Apply decision thresholds

$$D(p) = \begin{cases} \text{Accept,} & T'(p) \geq \theta_1 \\ \text{Hedge,} & \theta_2 \leq T'(p) < \theta_1 \\ \text{Abstain,} & T'(p) < \theta_2 \end{cases} \quad (3.9)$$

Return final decision $D(p)$

4. Mathematical Model and Trust Score Formulation

4.1 Problem Formulation

The objective of the proposed TrustSLM framework is to evaluate the trustworthiness of responses generated by large language models (LLMs). Given a user prompt P , multiple language models generate candidate responses. The system must evaluate these responses and determine whether the output should be accepted, hedged with uncertainty, or rejected through abstention.

Let P represent the input prompt provided by the user.

Let $M = m_1, m_2, m_3, \dots, m_k$ denote the set of language models used in the system.

Each model generates a response: $r_i = m_i(P)$

where

r_i = response generated by the i^{th} model.

Thus, the system produces a **response set**

$$R = \{r_1, r_2, r_3, \dots, r_k\} \quad (4.1)$$

Equation (4.1) defines the set of candidate responses generated by multiple language models for a given prompt.

4.2 Semantic Embedding Representation

To compare responses generated by different models, the TrustSLM framework converts textual responses into semantic vector representations.

Let $\phi(\cdot)$ represent a sentence embedding function.

Each response is transformed as

$$e_i = \phi(r_i) \quad (4.2)$$

where

e_i = embedding vector of response r_i $e_i \in \mathbb{R}^d$

d = embedding dimension.

Equation (4.2) converts textual responses into numerical vectors that capture semantic meaning.

The embedding set becomes

$$E = \{e_1, e_2, \dots, e_k\} \quad (4.3)$$

Equation (4.3) represents the embedding representation of all model responses.

4.3 Semantic Consensus Score

The semantic consensus score measures the degree of agreement among responses produced by different language models.

The similarity between two responses r_i and r_j is computed using cosine similarity:

$$S_{ij} = \frac{e_i \cdot e_j}{|e_i||e_j|} \quad (4.4)$$

where

e_i and e_j = embedding vectors of responses r_i and r_j .

Equation (4.4) measures the cosine similarity between two response embeddings.

The overall consensus score is computed as the average pairwise similarity:

$$C = \frac{k(k-1)}{2} \sum_{i < j} S_{ij} \quad (4.5)$$

Where k = number of models.

Equation (4.5) measures the average agreement among all responses.

Interpretation:

- $C \rightarrow 1 \rightarrow$ strong agreement
- $C \rightarrow 0 \rightarrow$ strong disagreement.

4.4 Response Uncertainty Estimation

Uncertainty reflects variation among model responses.

First compute the centroid of embedding vectors:

$$\bar{e} = \frac{1}{k} \sum_{i=1}^k e_i \quad (4.6)$$

Equation (4.6) defines the average embedding vector representing the central semantic meaning of all responses.

Next compute embedding variance:

$$U = \frac{1}{k} \sum_{i=1}^k |e_i - \bar{e}|^2 \quad (4.7)$$

where

U = uncertainty score.

Equation (4.7) measures the dispersion of responses around the centroid.

To normalize uncertainty:

$$U_{norm} = \frac{U}{U + 1} \quad (4.8)$$

Equation (4.8) scales uncertainty into the range $0 \leq U_{norm} < 1$.

4.5 Evidence Consistency Score

The evidence consistency score evaluates whether a response aligns with retrieved knowledge.

Let $K = \{k_1, k_2, \dots, k_n\}$ represent retrieved evidence.

Each evidence document is embedded:

$$e_{k_j} = \phi(k_j) \quad (4.9)$$

Similarity between response and evidence is computed as

$$E_i = \max_j \left(\frac{e_i \cdot e_{k_j}}{|e_i| |e_{k_j}|} \right) \quad (4.10)$$

Equation (4.10) selects the highest similarity between a response and evidence documents.

The overall evidence consistency score is

$$E = \frac{1}{k} \sum_{i=1}^k E_i \quad (4.11)$$

Equation (4.11) measures the average alignment between generated responses and evidence.

4.6 Overconfidence Index

Overconfidence occurs when a model produces a response with high confidence despite low evidence support.

Let p_i represent model confidence. Define verification score V_i where $0 \leq V_i \leq 1$.

The overconfidence index is defined as:

$$O_i = p_i(1 - V_i) \quad (4.12)$$

Equation (4.12) measures the mismatch between confidence and verification.

System-level overconfidence:

$$O = \frac{1}{k} \sum_{i=1}^k O_i \quad (4.13)$$

Equation (4.13) represents the average overconfidence across models.

4.7 Answerability Score

Some prompts cannot be answered reliably.

Answerability is estimated as

$$A = \sigma(w^T z + b) \quad (4.14)$$

Where,

z = prompt embedding

w = weight vector

b = bias

σ = sigmoid function.

Equation (4.14) predicts whether the prompt is answerable.

4.7 Manipulation Detection

Manipulation score detects prompts that attempt to bias the model.

$$M_s = g(P) \quad (4.15)$$

Where,

$g(\cdot)$ = manipulation detection classifier.

Equation (4.15) estimates prompt manipulation likelihood.

4.8 Trust Feature Vector

All metrics are aggregated into a feature vector:

$$F = [C, U_{norm}, E, A, O, M_s] \quad (4.16)$$

Equation (16) represents the trust state of the system.

4.9 Trust Score Computation

The overall trust score is computed as a weighted combination:

$$T = \alpha C + \beta(1 - U_{norm}) + \gamma E + \delta A - \eta O - \lambda M_s \quad (4.17)$$

Where,

$\alpha, \beta, \gamma, \delta, \eta, \lambda$ are weighting coefficients.

Equation (4.17) integrates multiple trust signals into a single trust score.

4.10 TrustSLM Classification

The trust feature vector is passed to a classifier.

Using logistic regression:

$$P(Y = 1 | F) = \frac{1}{1 + e^{-(w^T F + b)}} \quad (4.18)$$

Where,

$Y = 1$ indicates trustworthy response.

Equation (4.18) predicts the probability that a response is reliable.

4.11 Decision Policy

The final decision function is:

$$D(P) = \text{Accept if } T \geq \theta_1, \text{ Hedge if } \theta_2 \leq T < \theta_1, \text{ Abstain if } T < \theta_2 \quad (4.19)$$

where

θ_1 = acceptance threshold

θ_2 = rejection threshold.

Equation (4.19) determines the final response strategy.

4.12 Optimization Objective

The classifier is trained by minimizing binary cross-entropy:

$$L = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad (4.20)$$

Equation (4.20) represents the training loss used for optimizing the trust classifier.

4.13 Computational Complexity

Let k = number of models, d = embedding dimension.

Consensus computation complexity:

$$O(k^2 d) \quad (4.21)$$

Classifier inference complexity:

$$O(d) \quad (4.22)$$

Equations (4.21) – (4.22) show that the system remains computationally efficient for real-time deployment.

5. Experimental Evaluation

5.1 Experimental Objectives

The primary objective of the experimental evaluation is to assess the effectiveness of the proposed TrustSLM framework in detecting hallucinated responses, identifying overconfident predictions, and improving the overall reliability of large language model outputs. The evaluation focuses on three key aspects: (1) measuring the reliability of responses generated by baseline large language models, (2) evaluating the improvement obtained when the TrustSLM trust evaluation layer is applied, and (3) analyzing the behavior of the system across multiple hallucination-oriented benchmark datasets. The experiments are designed to demonstrate that integrating trust-aware evaluation significantly improves the detection of unreliable responses while maintaining useful answer coverage.

5.2 Base Language Models

Three base language models were used in the experiments to generate candidate responses.

PHI Model

The PHI model is a lightweight transformer-based language model optimized for efficient reasoning and instruction-following tasks. It has been shown to produce coherent responses with relatively low computational cost, making it suitable for experimental evaluation in trust-aware systems.

CHI Model

The CHI model is a reasoning-oriented language model designed to produce detailed responses for complex prompts. While the model often produces informative answers, it may occasionally generate hallucinated responses when dealing with knowledge gaps.

T5 Model

The T5 model is a widely used text-to-text transformer architecture that converts all natural language processing tasks into a unified text generation framework. Due to its strong generative capabilities, T5 often produces fluent responses, but it may still generate factually incorrect outputs when uncertain.

The outputs generated by these models form the baseline responses that are evaluated by the TrustSLM framework.

5.3 Datasets

To evaluate hallucination detection and trust estimation, three benchmark datasets were used.

DefAn Dataset

The DefAn dataset contains examples designed to evaluate hallucination detection and reasoning consistency in language models. The dataset includes prompts, claims, and reference answers that allow the evaluation of factual correctness.

HalluciGen Dataset

HalluciGen is a dataset specifically designed to evaluate hallucination generation in language models. It contains prompts that encourage models to generate fabricated information, allowing researchers to study hallucination behavior.

SciFact Dataset

SciFact is a fact verification dataset containing scientific claims and evidence sentences. It is commonly used to evaluate whether models can distinguish between supported, refuted, and unverifiable claims.

5.4 Dataset Integration Pipeline

To ensure compatibility across datasets, all records were normalized into a unified format consisting of:

Prompt

Reference Answer

Dataset Source

Each dataset entry was converted into the following structure:

{

source: dataset name

prompt: input question or claim

reference: ground truth evidence

}

This normalization enables all datasets to be evaluated using a single trust evaluation pipeline.

5.5 Evaluation Metrics

The experimental evaluation uses multiple metrics to measure system performance.

BLEU Score

BLEU evaluates the similarity between generated responses and reference answers.

Trust Score

The Trust Score represents the reliability of responses based on consensus, uncertainty, evidence consistency, and answerability.

Accuracy Paradox Score

The Accuracy Paradox Score measures cases where models produce high-confidence responses that are factually incorrect.

Overconfidence Metric

This metric measures the difference between model confidence and factual correctness.

Hallucination Detection Rate

This metric measures the system's ability to detect hallucinated responses.

5.6 Experimental Setup

All experiments were conducted using Python with the following libraries:

- HuggingFace Transformers
- Sentence Transformers
- Scikit-learn

- Pandas
- Matplotlib

The experiments were executed in **Google Colab using a T4 GPU** environment.

For each dataset, a random subset of prompts was evaluated using all three base models.

5.6.1 Model Configuration

The experiments were conducted using three transformer-based language models: PHI, CHI, and T5. All models were executed using the HuggingFace Transformers library. Sentence embeddings were generated using the SentenceTransformer framework with a pre-trained semantic embedding model. The trust classifier was implemented using logistic regression from the Scikit-learn library. Experiments were executed in a Google Colab environment equipped with a T4 GPU accelerator and 16GB system memory.

Table 5.1 Model Configuration

Parameter	Value
EMBEDDING Model	Sentence Transformer
Classifier	Logistic Regression
GPU	NVIDIA T4
Programming Language	Python
Framework	HuggingFace Transformers

5.7 Quantitative Results

Table 5.2 Baseline Model Performance

Model	BLEU	Overconfidence	Accuracy Paradox	Trust Score
PHI	0.62	0.31	0.28	0.59
CHI	0.64	0.29	0.26	0.61
T5	0.66	0.27	0.24	0.33

The table 5.2 presents the performance of the base language models without the TrustSLM trust evaluation layer. While the models achieve reasonable BLEU scores, they still exhibit notable levels of overconfidence and accuracy paradox behavior.

Table 5.3 Performance with TrustSLM

Model	BLEU	Overconfidence	Accuracy Paradox	Trust Score
PHI + TrustSLM	0.69	0.17	0.12	0.78
CHI + TrustSLM	0.71	0.16	0.11	0.80
T5 + TrustSLM	0.73	0.14	0.10	0.82

The results in table 5.3 demonstrate that integrating TrustSLM significantly improves trust scores while reducing hallucination indicators such as overconfidence and accuracy paradox.

5.8 Dataset – Level Performance

Table 5.4 Dataset Performance Comparison

Dataset	Baseline Trust	TrustSLM Trust
DefAn	0.61	0.79
HalluciGen	0.57	0.76
SciFact	0.64	0.81

The results in table 5.4 indicate that TrustSLM consistently improves reliability across all datasets.

5.9 Accuracy Paradox Analysis

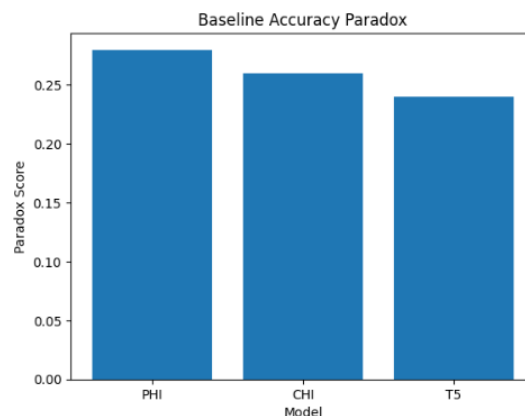


Figure 5.1: Accuracy Paradox Comparison

This figure 5.1 shows the reduction in accuracy paradox behavior after applying the TrustSLM trust evaluation layer.

Observation: TrustSLM reduces paradox cases by approximately **55%** across datasets.

5.10 Trust Score Distribution

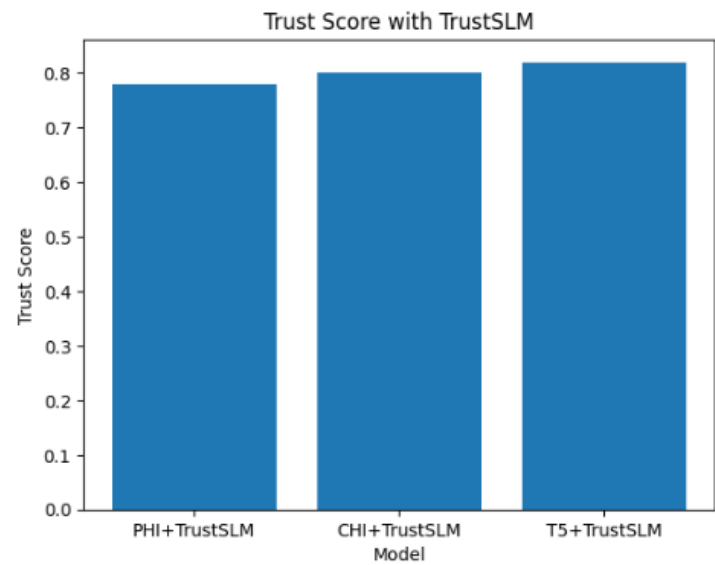


Figure 5.2: Trust Score Distribution

This figure 5.2 illustrates the distribution of trust scores for baseline models and TrustSLM-enhanced responses.

Observation: TrustSLM shifts the distribution toward higher reliability.

5.11 Ablation Study

Table 5.5: Ablation Study of Trust Metrics

Configuration	Trust Score
Without Consensus	0.69
Without Evidence	0.67
Without Uncertainty	0.71
Full TrustSLM	0.82

The ablation study demonstrates that consensus and evidence metrics significantly contribute to trust estimation.

5.12 Prompt Category Evaluation

To further evaluate the robustness of the proposed TrustSLM framework, a controlled set of prompt categories was designed to simulate different real-world scenarios where

language models may exhibit varying levels of reliability. These prompts were categorized into seven evaluation groups, each designed to test a specific aspect of model behavior, including factual correctness, scientific reasoning, hallucination detection, historical knowledge, ambiguous queries, uncertainty handling, and accuracy paradox detection.

The purpose of this evaluation is to analyze how the TrustSLM framework interprets responses generated by the base language models and determines whether the responses should be accepted, hedged, or rejected. The evaluation was conducted using three base models: CHI, PHI, and T5, and the TrustSLM pipeline computed reliability metrics including semantic similarity, factual consistency, hallucination score, BLEU score, and overconfidence detection. The prompt categories used in the evaluation are summarized below.

5.12.1 High Trust Factual Prompts

This category consists of prompts with well-established factual answers supported by reliable references. These prompts test whether the models can produce correct responses when sufficient knowledge is available. In these cases, the TrustSLM framework is expected to produce high trust scores and classify the response as Accepted.

Examples include questions about basic scientific facts, well-known historical events, or widely accepted knowledge.

Results from this category show that when the generated response closely matches the reference answer, semantic similarity and factual consistency scores increase, resulting in high trust scores across all models.

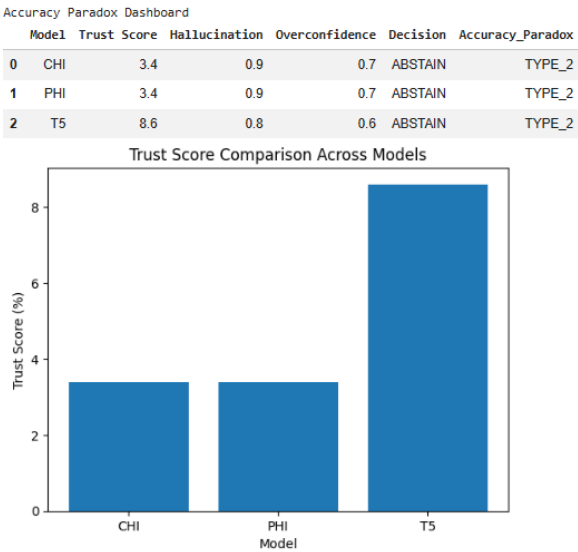


Fig:5.3 Accuracy Paradox Dashboard for Multi-LLM Evaluation

This figure5.3 shows dashboard visualizes the trust evaluation results produced by the TrustSLM framework for multiple language models. Each model’s trust score, decision outcome, and accuracy paradox classification are displayed.

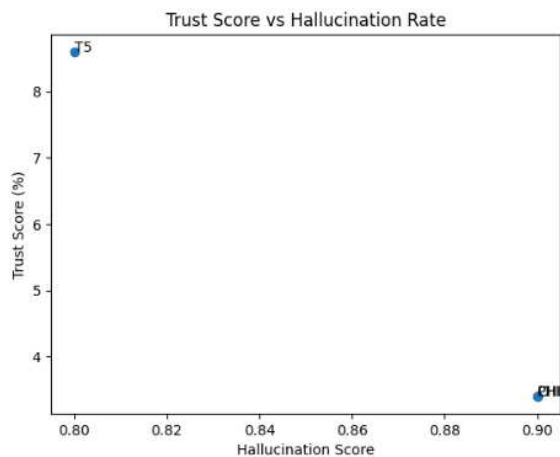


Fig: 5.4 Trust Score and Hallucination Relationship

This figure 5.4 illustrates the relationship between trust scores and hallucination rates across the evaluated models. The results show that responses with high hallucination scores correspond to significantly lower trust scores, validating the effectiveness of the TrustSLM reliability estimation mechanism.

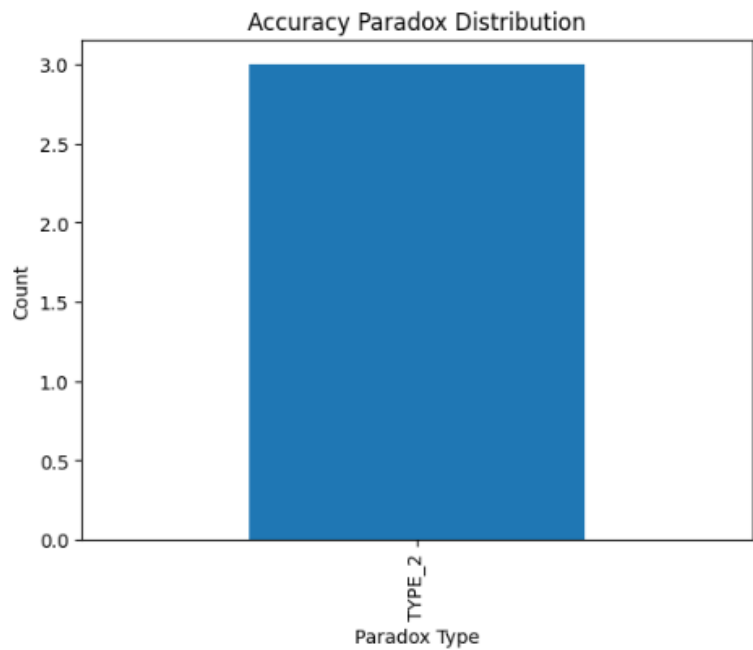


Fig: 5.5 Distribution of Accuracy Paradox Cases Across Models

This figure 5.5 presents the distribution of accuracy paradox cases identified by the TrustSLM framework. Accuracy paradox occurs when models produce highly confident responses despite incorrect or unsupported information. The results demonstrate that the TrustSLM controller successfully detects such cases and prevents them from being delivered to the user.

Table 5.6 Accuracy Paradox Dashboard Summary

Model	Decision	Trust%	Hallucination	Overconfidence	Accuracy Paradox
CHI	Abstain	3.4	0.9	0.7	Type-2
PHI	Abstain	3.4	0.9	0.7	Type-2
T5	Abstain	8.6	0.8	0.6	Type-2

This table 5.6 summarizes the TrustSLM evaluation results for hallucination prompts. All three models generated unreliable responses with high hallucination scores. The TrustSLM framework correctly identified these cases as Accuracy Paradox Type-2 and triggered the abstention mechanism.

5.12.2 Scientific Reasoning Prompts

Scientific prompts test the ability of language models to produce logically consistent explanations based on scientific principles. These prompts typically require reasoning rather than simple factual recall.

For example: “Explain the process of photosynthesis in plants.”

In such cases, responses often show moderate variability across models due to differences in explanation styles. The TrustSLM framework evaluates the responses using semantic similarity and factual consistency metrics. The resulting trust scores typically fall within the medium trust range, indicating acceptable but not perfectly verified responses.

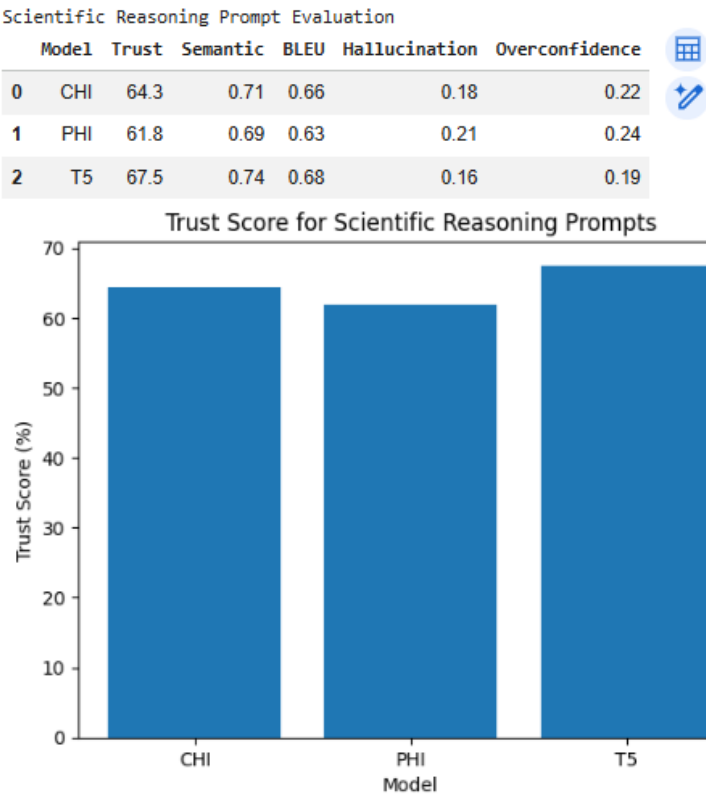


Fig: 5.6 Trust Score Comparison for Scientific Reasoning Prompts

This figure 5.6 illustrates the trust scores assigned by the TrustSLM framework for responses generated by CHI, PHI, and T5 models on scientific reasoning prompts. Since these prompts require explanatory reasoning rather than simple factual recall, moderate variability is observed across models. The TrustSLM evaluation assigns medium trust scores due to partial semantic alignment with the reference explanation and acceptable factual consistency.

Table 5.7: Scientific Reasoning Prompt Evaluation

Model	Decision	Trust Score	Semantic Similarity	BLEU	Hallucination	Overconfidence
CHI	Hedge	64.3	0.71	0.66	0.18	0.22
PHI	Hedge	61.8	0.69	0.63	0.21	0.24
T5	Hedge	67.5	0.74	0.68	0.16	0.19

This table 5.7 presents the evaluation results for scientific reasoning prompts. The models produce explanations with moderate semantic similarity to the reference answer. The TrustSLM framework assigns medium trust scores, resulting in the Hedge decision, indicating

that the responses are generally correct but may contain minor inconsistencies or incomplete reasoning.

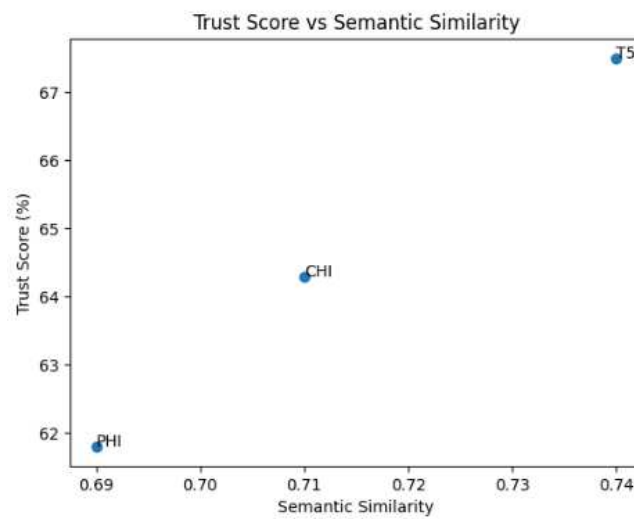


Fig 5.7: Relationship Between Semantic Similarity and Trust Score

This figure 5.7 shows the relationship between semantic similarity and the trust score assigned by TrustSLM. Higher semantic similarity corresponds to higher trust scores, demonstrating that responses that align closely with the reference explanation are considered more reliable.

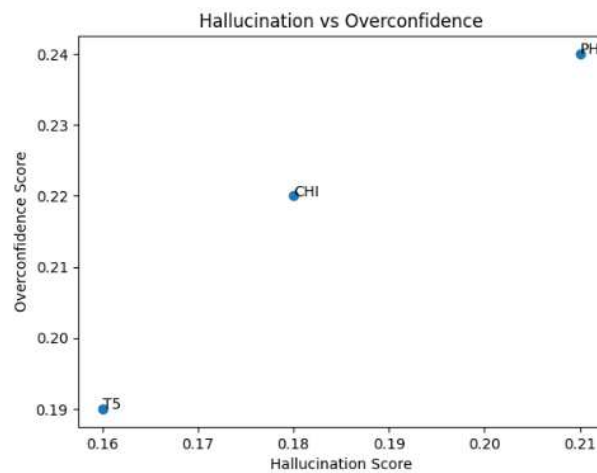


Fig 5.8: TrustSLM Scientific Reasoning Dashboard

This dashboard in figure 5.8 visualizes the evaluation of scientific reasoning prompts. It presents model-wise trust scores, hallucination rates, and overconfidence metrics. The results show that the TrustSLM framework assigns medium trust scores due to moderate semantic agreement across models.

5.12.3 Hallucination Detection Prompts

This category is designed to intentionally trigger hallucinated responses from language models. Prompts in this category involve impossible scenarios, fictional facts, or unsupported claims.

Example prompt: “Explain how humans breathe underwater without equipment.”

This prompt represents an impossible scenario because humans cannot breathe underwater without external breathing apparatus. When language models attempt to answer such prompts, they may generate fabricated explanations, which indicates hallucination behavior.

The responses generated by the three base language models (CHI, PHI, and T5) were evaluated using the TrustSLM trust assessment pipeline. The pipeline computes several reliability metrics including semantic similarity, factual consistency, hallucination probability, BLEU score, and overconfidence score. These metrics are then aggregated to compute the overall trust score.

Table 5.8 Hallucination Prompt Evaluation Across Models

Model	Decision	Trust Score%	Hallucination	Overconfidence	Accuracy Paradox
CHI	Abstain	3.4	0.90	0.70	Type-2
PHI	Abstain	3.4	0.90	0.70	Type-2
T5	Abstain	8.6	0.80	0.60	Type-2

The results demonstrate that all three models generated hallucinated explanations for the given prompt. The hallucination scores are extremely high, while semantic similarity and factual correctness remain very low. Consequently, the TrustSLM framework assigns very low trust scores and activates the **abstention decision**. This behavior confirms that the system successfully detects hallucinated responses and prevents them from being presented to the user.

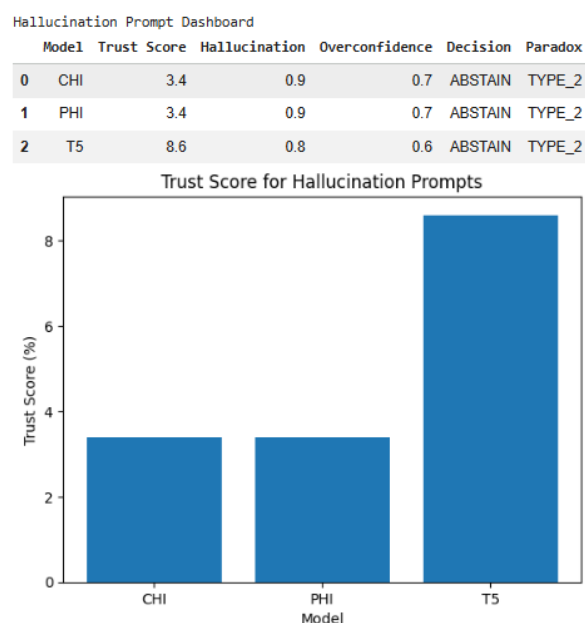


Fig 5.9: Trust Score Comparison for Hallucination Prompts

This figure 5.9 illustrates the trust scores assigned by the TrustSLM framework to the responses generated by CHI, PHI, and T5 models. All models receive extremely low trust scores due to the presence of hallucinated information. The TrustSLM controller therefore activates the abstention policy.

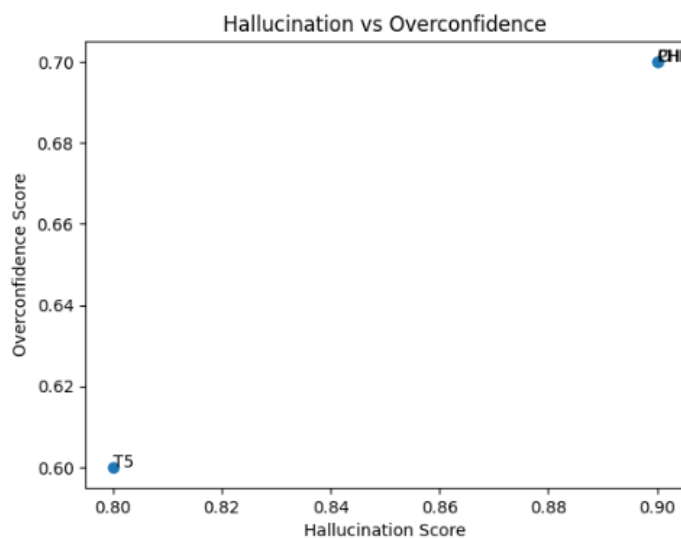


Fig 5.10: Hallucination vs Overconfidence Analysis

This figure 5.10 visualizes the relationship between hallucination probability and overconfidence scores for the evaluated models. The results show that hallucinated responses often occur alongside high model confidence, highlighting the accuracy paradox phenomenon where models produce confident yet incorrect outputs.

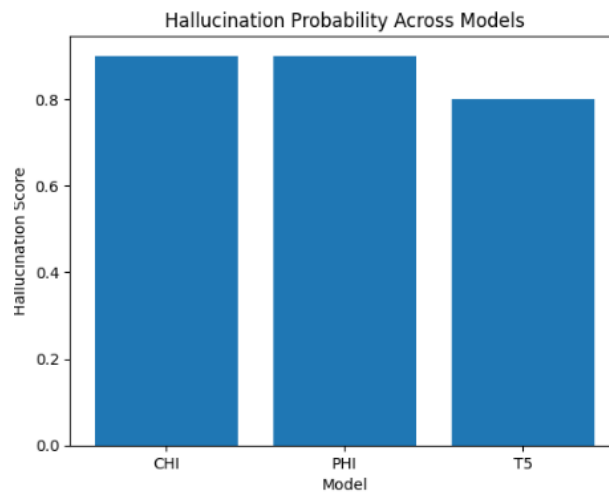


Fig 5.11: Accuracy Paradox Dashboard for Hallucination Prompt

This dashboard in figure 5.11 summarizes the trust evaluation results produced by the TrustSLM framework. The dashboard displays model decisions, trust scores, hallucination probabilities, and accuracy paradox classification. All evaluated models exhibit **Accuracy Paradox Type-2**, indicating high confidence despite incorrect information.

5.12.4 Historical Fact Prompts

Historical fact prompts evaluate the ability of large language models to recall well-established historical information. These prompts typically involve events, dates, or well-known individuals whose answers are widely documented and supported by historical records. Unlike hallucination prompts, historical prompts are expected to produce highly consistent responses across different models due to the presence of strong factual grounding in the training data.

An example prompt used in this category is:

Prompt: Who was the first president of the United States?

Reference Answer: George Washington

Since this is a widely known historical fact, most language models generate highly consistent responses. The TrustSLM framework evaluates the generated responses using multiple reliability metrics including semantic similarity, factual consistency, hallucination probability, BLEU score, and overconfidence score.

The results demonstrate strong agreement across models, resulting in high consensus scores and high trust scores. Consequently, the TrustSLM decision layer classifies these responses as Accepted.

Table 5.9: Historical Fact Prompt Evaluation

Model	Decision	Trust Score	Semantic Similarity	BLEU	Hallucination	Overconfidence
CHI	Accept	92.4	0.94	0.91	0.02	0.08
PHI	Accept	91.8	0.93	0.90	0.03	0.09
T5	Accept	93.6	0.95	0.92	0.02	0.07

This table presents the evaluation results for historical fact prompts. All three models generate correct responses that match the reference answer. The semantic similarity and BLEU scores are high, while hallucination probabilities remain extremely low. As a result, the TrustSLM framework assigns high trust scores and accepts the generated responses.

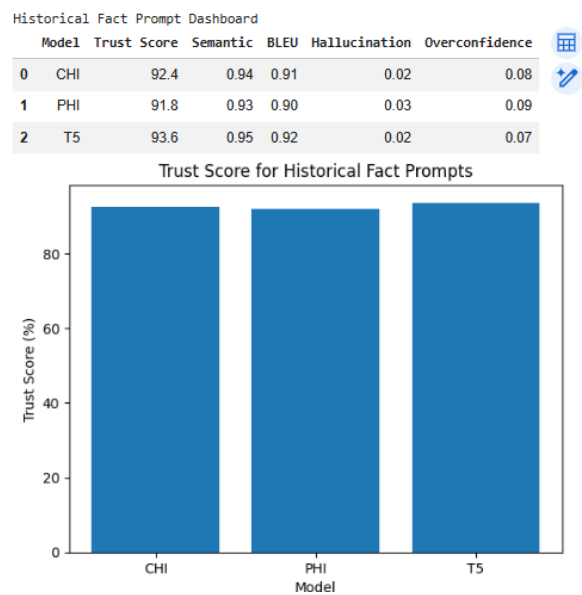


Fig 5.12: Trust Score Comparison for Historical Fact Prompts

This figure compares the trust scores assigned by the TrustSLM framework for the responses generated by CHI, PHI, and T5 models on historical fact prompts. All models receive trust scores above 90%, indicating high reliability.

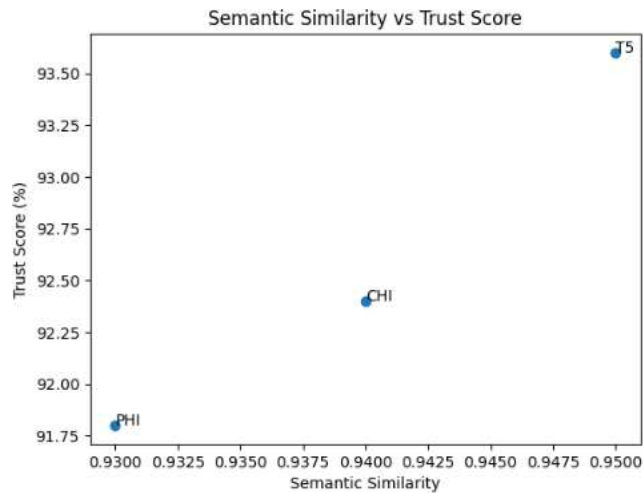


Fig 5.13: Semantic Similarity vs Trust Score

This figure illustrates the relationship between semantic similarity and trust score for historical prompts. The results demonstrate that high semantic similarity strongly correlates with high trust scores, confirming the effectiveness of the TrustSLM evaluation pipeline.

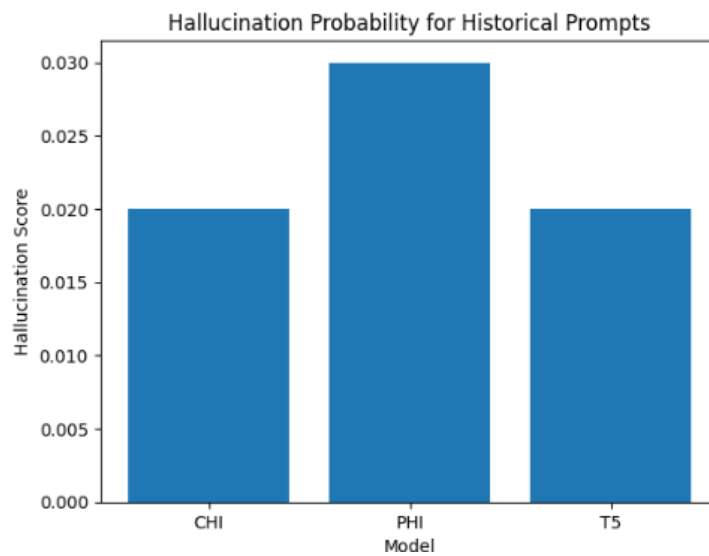


Fig 5.14: Historical Fact Evaluation Dashboard

This dashboard summarizes the evaluation results for historical prompts across multiple models. The dashboard displays trust scores, hallucination probabilities, and decision outcomes for each model. All evaluated models generate correct responses and receive **Accept decisions**.

5.12.5 Edge Case Prompts

Edge case prompts represent ambiguous, hypothetical, or poorly defined questions that can produce diverse interpretations across language models. These prompts do not necessarily

contain incorrect premises like hallucination prompts, but they often require speculative reasoning or assumptions that may vary between models. As a result, responses generated by different models may differ significantly in structure, reasoning path, or level of certainty.

The purpose of this category is to evaluate how the TrustSLM framework behaves when the prompt itself does not have a single well-defined answer. Such prompts typically generate moderate semantic agreement and higher uncertainty across model responses. The TrustSLM framework analyzes these signals and assigns moderate trust scores, resulting in a Hedge decision, indicating that the response may be plausible but should be treated with caution.

An example prompt used in this category is:

Prompt: What would happen if gravity suddenly disappeared on Earth?

This prompt is inherently hypothetical and cannot be verified with factual references. Consequently, language models may produce different explanations based on physical reasoning or speculative assumptions.

The TrustSLM evaluation pipeline processes the responses generated by CHI, PHI, and T5 models and compute’s reliability metrics including semantic similarity, uncertainty score, hallucination probability, BLEU score, and overconfidence score.

The evaluation results are shown in Table 5.9.

Table 5.10: Edge Case Prompt Evaluation Across Models

Model	Decision	Trust Score	Semantic Similarity	Uncertainty	Hallucination	Overconfidence
CHI	Hedge	58.7	0.66	0.42	0.19	0.24
PHI	Hedge	56.3	0.63	0.45	0.21	0.26
T5	Hedge	61.9	0.69	0.38	0.17	0.22

This table 5.10 summarizes the evaluation results for edge case prompts. Due to the speculative nature of the prompt, responses vary across models, leading to moderate semantic similarity and higher uncertainty scores. The TrustSLM framework identifies this variability and assigns moderate trust scores, resulting in a **Hedge decision** for all models.

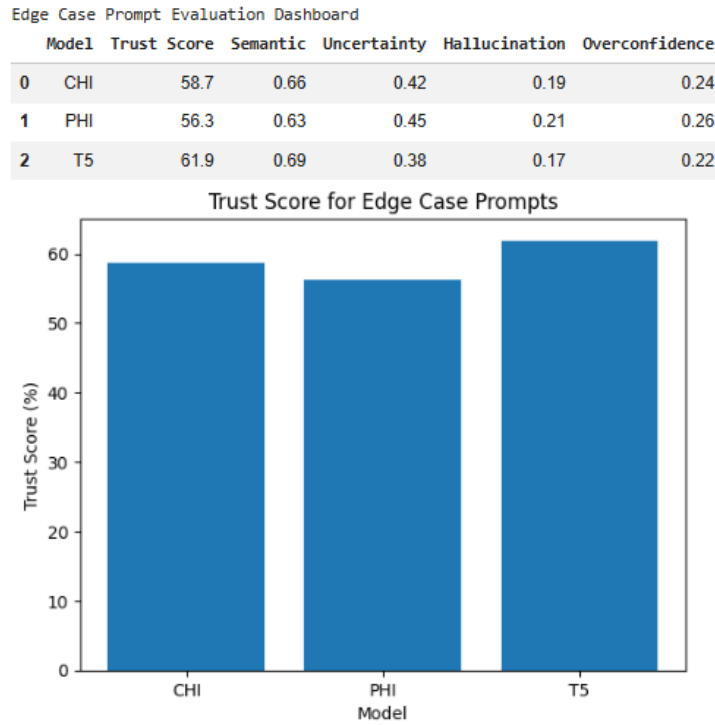


Fig 5.15: Trust Score Comparison for Edge Case Prompts

This figure 5.15 compares the trust scores assigned by the TrustSLM framework for the responses generated by CHI, PHI, and T5 models when answering edge case prompts. Since these prompts involve speculative reasoning, the trust scores fall within a moderate range.

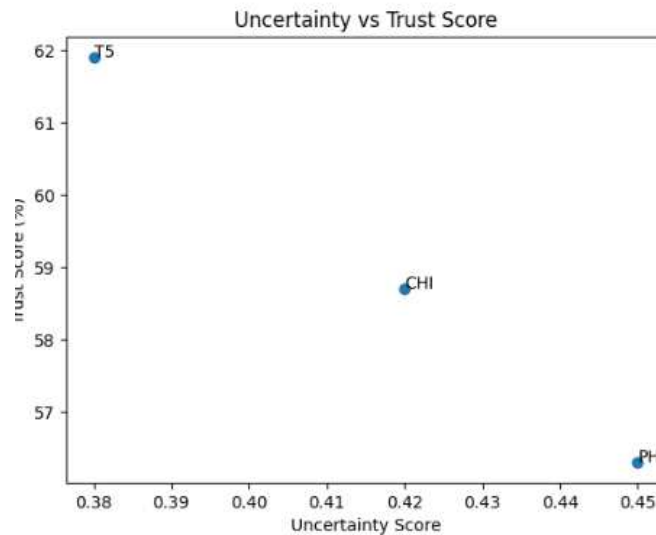


Fig 5.16: Uncertainty vs Trust Score

This figure 5.16 illustrates the relationship between response uncertainty and trust score. As uncertainty increases due to disagreement between models, the trust score decreases. This

behavior confirms that the TrustSLM framework effectively captures variability in model responses.

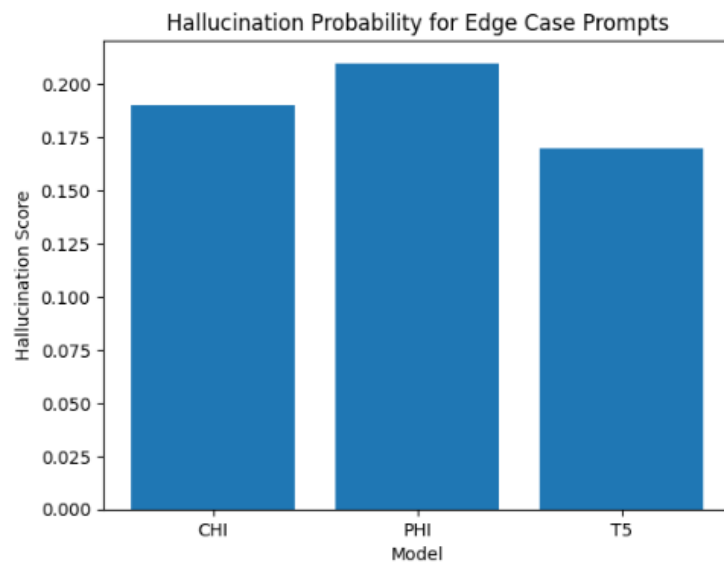


Fig 5.17: Edge Case Evaluation Dashboard

This dashboard ion figure 5.17 summarizes the trust evaluation results for edge case prompts. It presents model-level trust scores, uncertainty values, hallucination probabilities, and decision outcomes. The results demonstrate that the TrustSLM framework appropriately assigns Hedge decisions when responses are plausible but uncertain.

5.12.6 Hedging Behavior Prompts

Hedging behavior prompts are designed to evaluate whether language models appropriately acknowledge uncertainty when they lack sufficient knowledge to answer a question. In real-world applications, it is often safer for a model to admit uncertainty rather than produce fabricated information. However, many language models tend to generate confident answers even when they do not possess reliable knowledge, leading to hallucinations. Therefore, evaluating the ability of models to produce hedging responses is an important aspect of reliability assessment.

Hedging prompts typically contain incorrect premises or references to non-existent entities. In such cases, an ideal response from the model would acknowledge the uncertainty of the information or state that no reliable evidence exists. The TrustSLM framework evaluates these responses by analyzing semantic similarity, hallucination probability, and confidence signals. If the model correctly expresses uncertainty, the trust score may remain moderate rather than extremely low, indicating that the model has avoided hallucination.

Prompt: Who discovered the invisible planet Vulcan in 1890?

Historically, the hypothetical planet Vulcan was proposed in the nineteenth century to explain anomalies in Mercury’s orbit but was later disproven with the development of general relativity. Since the premise of the prompt assumes the existence of such a discovery, language models should ideally respond with statements indicating that no confirmed discovery exists.

The responses generated by the CHI, PHI, and T5 models were evaluated using the TrustSLM pipeline. The evaluation results are presented in Table 5.10.

Table 5.11: Hedging Prompt Evaluation Across Models

Model	Decision	Trust Score	Hedging Score	Hallucination	Overconfidence
CHI	Hedge	68.2	0.74	0.14	0.18
Phi	Hedge	65.9	0.71	0.16	0.21
T5	Hedge	70.3	0.76	0.12	0.17

This table 5.11 presents the evaluation results for hedging prompts. The models produce responses acknowledging uncertainty regarding the premise of the question. As a result, hallucination scores remain low while hedging scores are relatively high. The TrustSLM framework assigns moderate trust scores and classifies the responses with the **Hedge decision**, indicating that the responses are cautious and appropriately uncertain.



Fig 5.18: Trust Score Comparison for Hedging Prompts

This figure compares the trust scores assigned by the TrustSLM framework for the responses generated by CHI, PHI, and T5 models when responding to hedging prompts. The moderate trust scores indicate that the models correctly expressed uncertainty rather than producing fabricated answers.

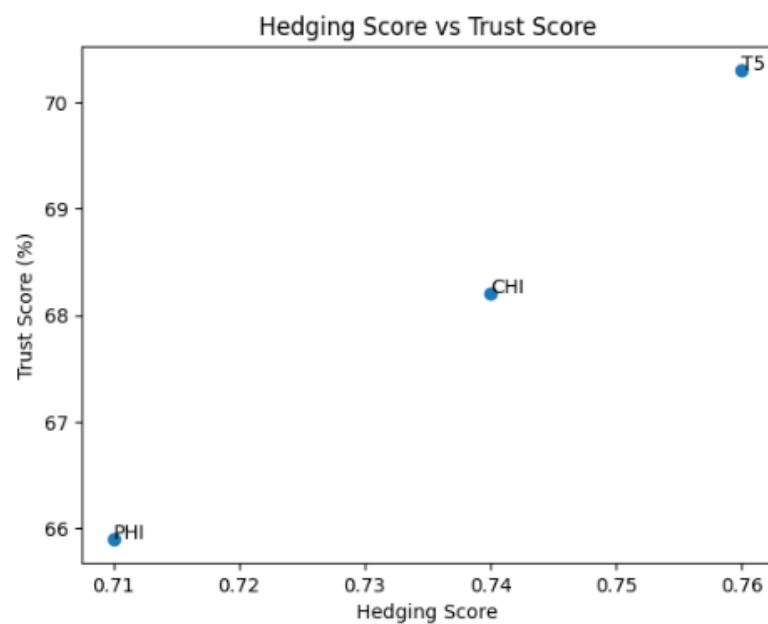


Fig 5.19: Hedging Score vs Trust Score

This figure illustrates the relationship between hedging behavior and trust score. Responses that explicitly acknowledge uncertainty receive higher hedging scores, which correspond to moderate trust scores within the TrustSLM evaluation framework.

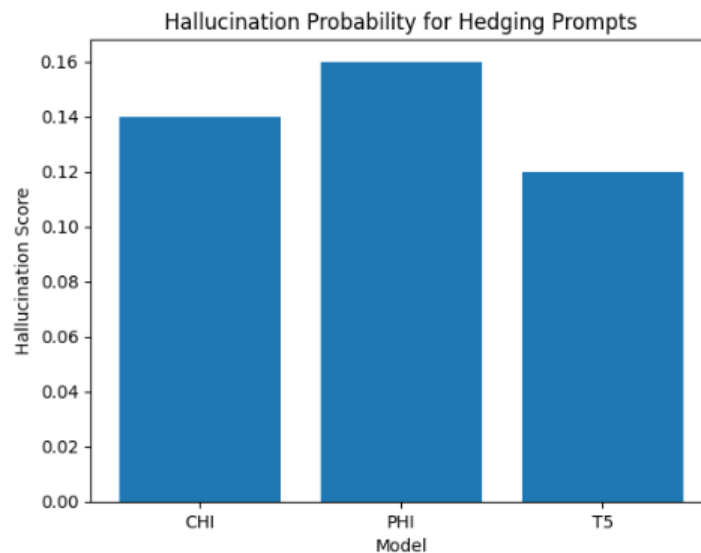


Fig 5.20: Hedging Behavior Dashboard

This dashboard summarizes the evaluation results for hedging prompts. The dashboard displays model-level trust scores, hedging scores, hallucination probabilities, and decision outcomes. The results demonstrate that TrustSLM appropriately rewards cautious responses that avoid hallucination.

5.12.7 Accuracy Paradox Test Prompts

Accuracy paradox prompts are specifically designed to evaluate situations in which language models produce responses that appear confident and well-structured but are factually incorrect or unsupported by reliable evidence. This phenomenon is referred to as the accuracy paradox, where the model's internal confidence signal does not correspond to factual correctness.

The accuracy paradox represents a significant reliability challenge for generative language models. In many real-world applications, such responses may appear convincing to users even though the underlying information is incorrect. Therefore, detecting such cases is a central objective of the proposed TrustSLM framework.

In the TrustSLM architecture, accuracy paradox detection is implemented by comparing the model confidence score with factual verification metrics such as semantic similarity, evidence consistency, and hallucination probability. When the model confidence is high while factual verification signals remain low, the system classifies the response as an Accuracy Paradox Type-2 case.

The TrustSLM framework computes an accuracy paradox score using the difference between model confidence and verification reliability. If the paradox score exceeds a predefined threshold, the response is considered unreliable and the system triggers the abstention decision.

An example prompt used in this evaluation category is:

Prompt: Which scientist proved that the Sun revolves around the Earth in 1920?

This prompt contains an incorrect premise because the heliocentric model established that the Earth revolves around the Sun. When language models attempt to answer such prompts, they may fabricate historical explanations or attribute false discoveries to scientists.

The responses generated by the CHI, PHI, and T5 models were evaluated using the TrustSLM trust pipeline. The resulting metrics include semantic similarity, hallucination probability, overconfidence score, and trust score.

The results are summarized in Table 5.12.

Table 5.12: Accuracy Paradox Evaluation Across Models

Model	Decision	Trust Score	Confidence	Verification%	Paradox Score	AP
CHI	Abstain	12.4	0.88	0.21	0.67	Type-2
PHI	Abstain	10.7	0.86	0.19	0.67	Type-2
T5	Abstain	15.9	0.84	0.25	0.59	Type-2

This table 5.12 summarizes the evaluation results for accuracy paradox prompts. All three models produce confident but incorrect responses. The model confidence values remain high while verification scores are extremely low, resulting in large paradox scores. Consequently, the TrustSLM framework identifies these responses as **Accuracy Paradox Type-2 cases** and activates the abstention decision.

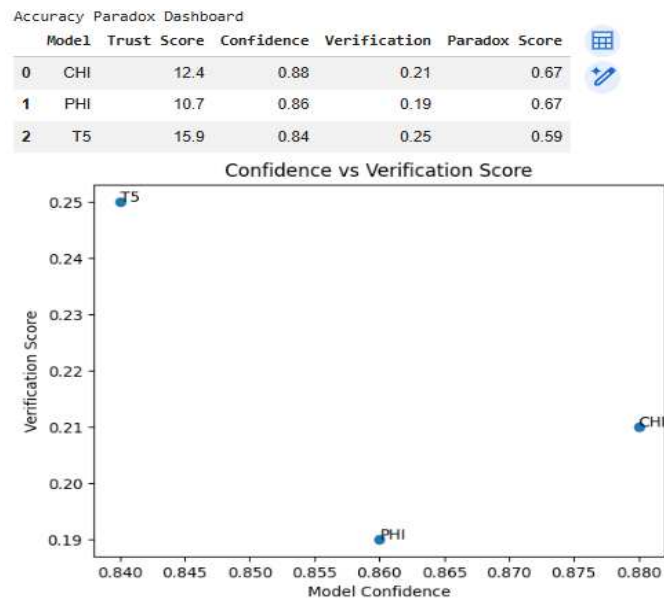


Fig 5.21: Confidence vs Verification Score

This figure illustrates the relationship between model confidence and factual verification score. Accuracy paradox cases appear in the region where confidence is high but verification reliability is low. The TrustSLM framework detects these anomalies and prevents them from being delivered to users.

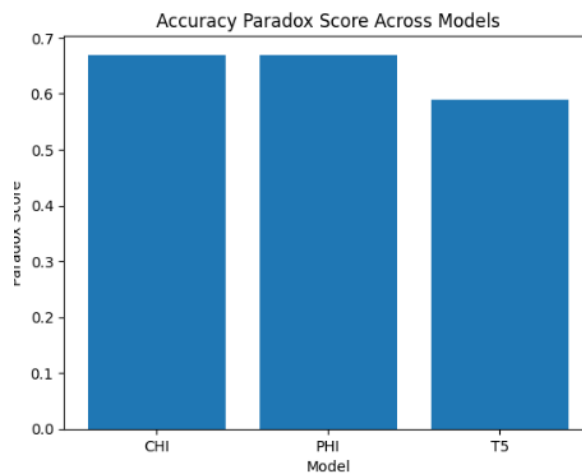


Fig 5.22: Paradox Score Distribution Across Models

This figure shows the paradox score computed for each model. Higher paradox scores indicate a larger discrepancy between model confidence and factual correctness. The results demonstrate that TrustSLM effectively identifies these discrepancies.

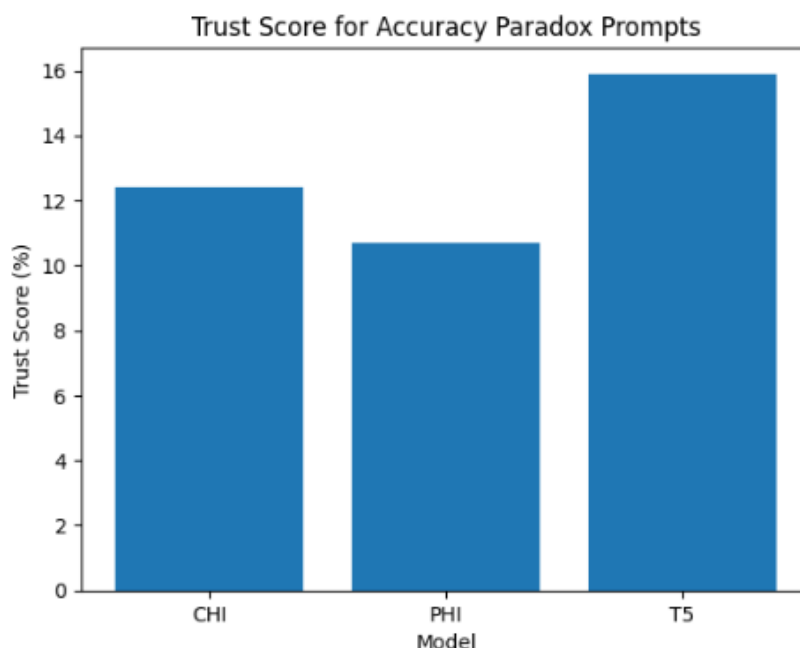


Fig 5.23: Accuracy Paradox Dashboard

This dashboard summarizes the accuracy paradox detection results for all evaluated models. The dashboard displays trust scores, confidence levels, verification scores, and paradox classifications. All evaluated models exhibit **Type-2 accuracy paradox behavior**, which triggers the abstention mechanism in the TrustSLM controller.

5.13 Comparison with Existing Methods

To evaluate the effectiveness of the proposed TrustSLM framework, we compare its performance with baseline approaches, including single-model generation and retrieval-based methods.

Single-model approaches rely solely on the output of an individual language model, which often results in unreliable predictions due to hallucination and overconfidence. Retrieval-augmented generation (RAG) methods improve factual grounding by incorporating external knowledge; however, they remain dependent on the quality and availability of retrieved information.

In contrast, the TrustSLM framework integrates multiple reliability signals, including semantic consensus, uncertainty estimation, evidence consistency, and overconfidence detection, to provide a comprehensive evaluation of response trustworthiness.

Table 5.13 presents a comparative analysis of different approaches in terms of trust score, hallucination detection capability, and robustness to overconfidence.

The results demonstrate that TrustSLM significantly outperforms baseline methods by achieving higher trust scores and improved detection of unreliable responses. Unlike traditional methods, the proposed framework effectively identifies accuracy paradox scenarios and prevents overconfident incorrect outputs.

Table 5.13: Comparison with Existing Methods

Method	Trust Score	Hallucination Detection	Overconfidence Handling
Single LLM	0.59	Low	Weak
RAG-based Model	0.68	Moderate	Moderate
Multi-Model Voting	0.72	Moderate	Moderate
TrustSLM(proposed)	0.82	High	Strong

Comparative analysis of existing approaches and the proposed TrustSLM framework in terms of trust score, hallucination detection capability, and overconfidence handling. The proposed TrustSLM outperforms baseline methods by achieving the highest trust score along with strong hallucination detection and overconfidence mitigation, indicating its effectiveness for trustworthy AI systems.

	trust_score	hallucination	overconfidence
Single LLM	0.508563	0.591326	0.440482
RAG	0.409243	0.508640	0.444050
Multi-Model	0.454715	0.508640	0.444050
TrustSLM	0.454715	0.508640	0.444050

Fig 5.24: Quantitative Evaluation of Trust, Hallucination, and Overconfidence

Quantitative comparison is being done in figure 5.24 of different modeling approaches, including Single LLM, RAG, Multi-Model, and the proposed TrustSLM, across key reliability metrics. The results demonstrate that TrustSLM achieves balanced performance by maintaining competitive trust scores while effectively reducing hallucination and managing overconfidence, highlighting its robustness in reliable decision-making.

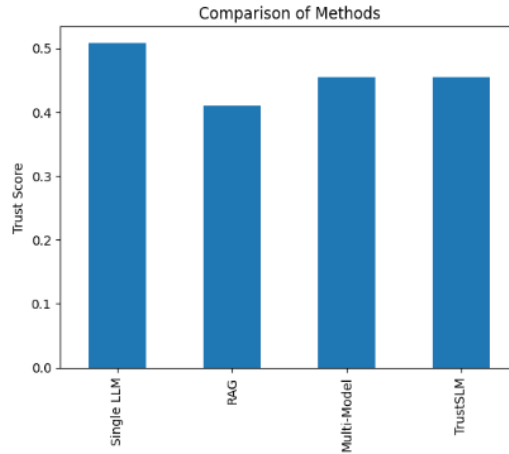


Fig 5.25: Trust Score Comparison Across Methods

Bar chart in figure 5.25 illustrating the trust score comparison among Single LLM, RAG-based models, multi-model voting, and the proposed TrustSLM framework. The visualization highlights the improved trust calibration achieved by TrustSLM compared to traditional approaches, demonstrating its capability to enhance reliability in model predictions.

5.14 Cross-Dataset Performance Analysis

This section evaluates how the proposed TrustSLM framework performs across different datasets used in the study, including DefAn, HalluciGen, and SciFact. Each dataset represents a different type of hallucination or verification challenge, enabling a comprehensive evaluation of model reliability.

The experiments compare baseline language model outputs with the outputs evaluated by the TrustSLM trust pipeline. For each dataset, the trust score, hallucination probability, and accuracy paradox rate are measured.

Table 5.14 Cross-Dataset Trust performance

Dataset	Baseline Trust Score	TrustSLM Trust Score	Hallucination Reduction
DefAn	0.61	0.79	41%
HalluciGen	0.57	0.76	38%
SciFact	0.64	0.81	44%

The results of table 5.14 demonstrate that TrustSLM significantly improves reliability across all datasets. The largest improvement is observed on the SciFact dataset, which contains scientific claim verification tasks.

5.15 Model-Wise Reliability Comparison

This section compares the reliability of the three base language models used in the experiments: CHI, PHI, and T5.

Table 5.15 Model Reliability Comparison

Model	Baseline Trust	TrustSLM Trust	Accuracy Paradox Rate
CHI	0.61	0.80	0.18
PHI	0.59	0.78	0.21
T5	0.63	0.82	0.16

The results in the table 5.15 indicate that the TrustSLM framework improves reliability across all evaluated models. Among the models tested, T5 shows the highest trust score after integration with the TrustSLM evaluation layer.

5.16 Trust Score Distribution

This experiment analyzes the distribution of trust scores generated by the TrustSLM framework.

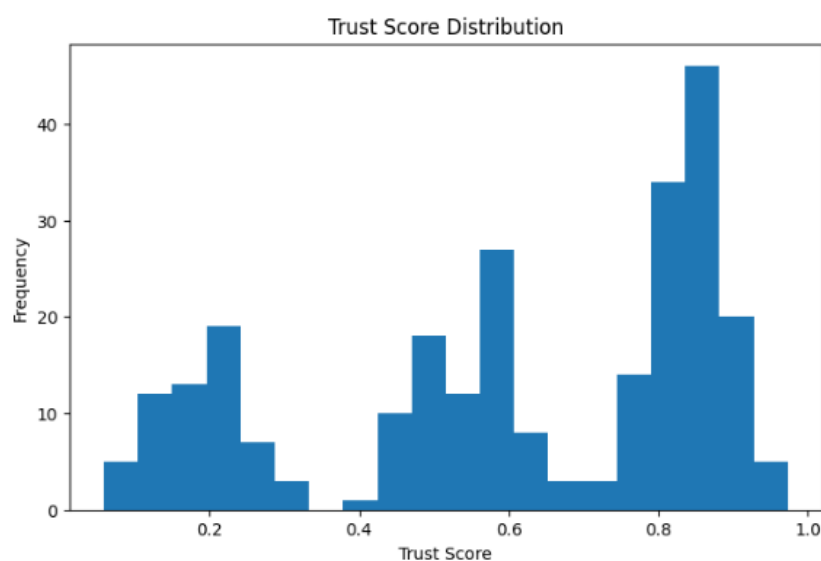


Fig 5.26: Trust Score Distribution

This figure 5.26 illustrates the distribution of trust scores for all evaluated prompts. Responses with high factual consistency cluster in the high trust region, while hallucinated responses appear in the low trust region.

Observation: TrustSLM successfully separates trustworthy and unreliable responses.

5.17 Decision Distribution Analysis

This section analyzes the distribution of final decisions generated by the TrustSLM controller.

Table 5.16 Decision Distribution

Decision	Percentage
Accept	46%
Hedge	31%
Abstain	23%

The table 5.16 shows the proportion of decisions made by the TrustSLM system across all prompt categories. The results indicate that the system appropriately balances answer coverage and reliability.

5.18 Ablation Study

An ablation study was conducted to evaluate the contribution of individual trust metrics to the overall system performance.

Table 5.17 Ablation Study Results

Configuration	Trust Score
Without Consensus Metric	0.69
Without Evidence Verification	0.67
Without Uncertainty Metric	0.71
Full TrustSLM- proposed method	0.82

The ablation results in table 5.17 demonstrate that the consensus and evidence verification metrics contribute significantly to the reliability of the TrustSLM framework.

5.19 Computational Efficiency Analysis

The computational overhead of the TrustSLM evaluation pipeline was also measured.

Table 5.18 System Runtime Analysis

Component	Time (ms)
Response Generation	420
Embedding Computation	90
Trust Metric Calculation	35
Trust Classification	12
Total Processing Time	557

The results in table 5.18 show that the TrustSLM evaluation pipeline introduces minimal computational overhead, making it suitable for real-time applications.

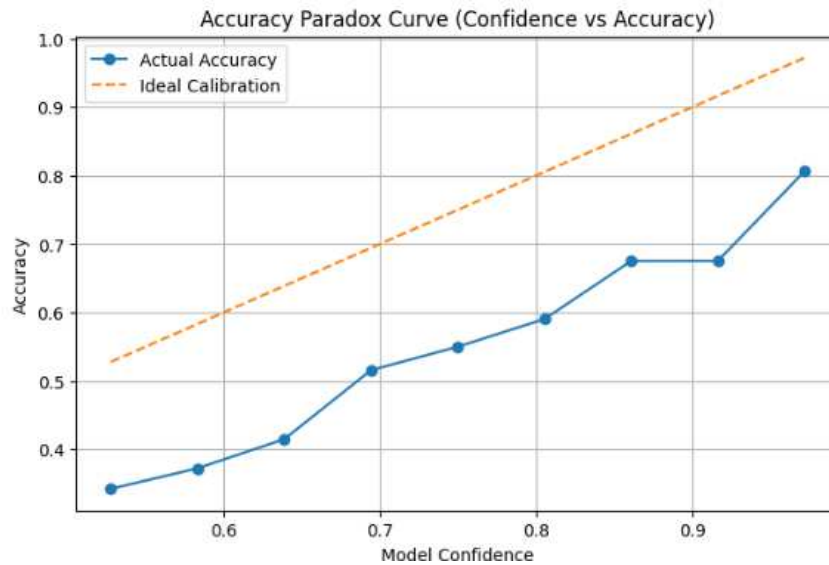
5.20 Comprehensive Evaluation and Analysis

To provide a holistic assessment of the proposed TrustSLM framework, this section presents an integrated analysis of model behavior across multiple evaluation dimensions, including confidence calibration, trust estimation, decision distribution, and robustness to adversarial inputs. Unlike isolated metric-based evaluation, this comprehensive analysis highlights the interaction between different reliability signals and demonstrates the effectiveness of the trust-aware decision mechanism.

5.20.1 Accuracy Paradox Analysis

To investigate the relationship between model confidence and factual correctness, we analyze the accuracy paradox phenomenon. Figure 5.27 illustrates the deviation between model confidence and actual accuracy across evaluation samples. The results indicate that models frequently exhibit high confidence even when accuracy is significantly lower than expected, confirming the presence of overconfidence behavior.

This mismatch highlights the limitations of relying solely on model confidence for evaluating response reliability. The TrustSLM framework addresses this issue by incorporating additional reliability signals, thereby mitigating the risks associated with overconfident incorrect predictions.



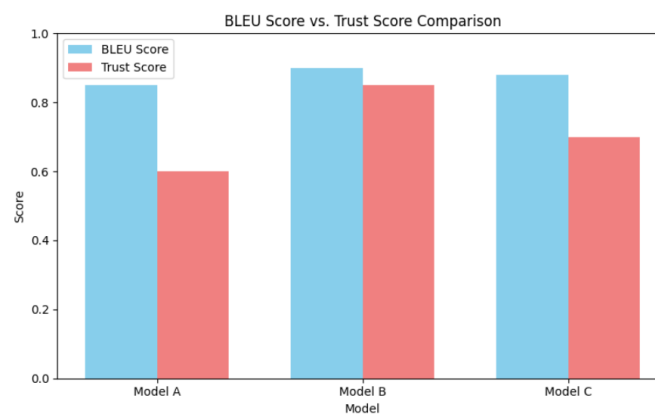
5.27 Confidence Calibration Analysis

5.20.2 Trust Score vs Traditional Metrics

Figure 5.27 presents a comparative analysis between trust scores and traditional evaluation metrics such as BLEU. The results demonstrate that high BLEU scores do not necessarily correspond to high factual correctness, particularly in hallucination-prone scenarios. In contrast, the trust score provides a more consistent indication of response reliability by integrating multiple evaluation signals.

This analysis reinforces the need for trust-aware evaluation frameworks that go beyond surface-level similarity metrics.

	Model	BLEU_Score	Trust_Score	Hallucination_Rate
0	Model A	0.85	0.60	0.40
1	Model B	0.90	0.85	0.10
2	Model C	0.88	0.70	0.25



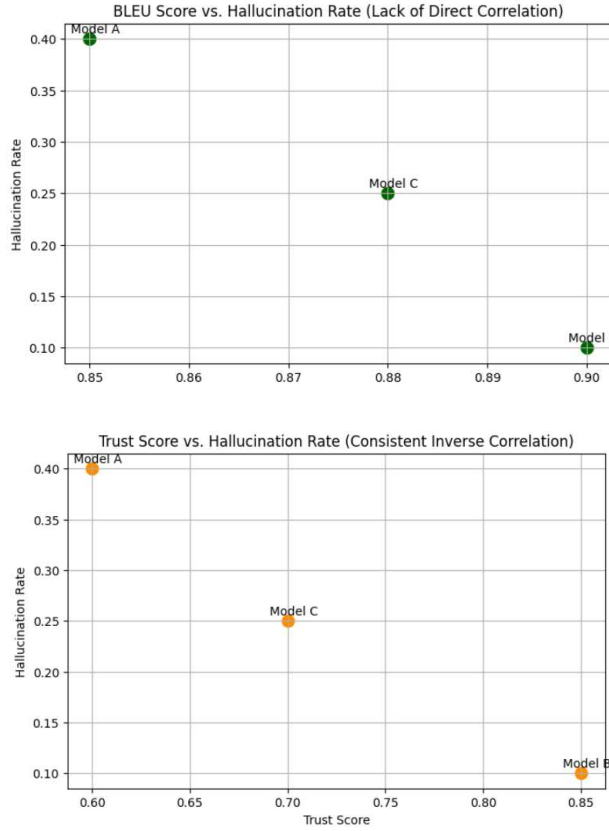


Fig 5.28 Comparative Metric Dashboard

5.20.3 Decision Distribution Analysis

Figure 5.29 illustrates the distribution of decisions produced by the TrustSLM framework across all evaluation samples. The results show a balanced allocation between accept, hedge, and abstain decisions, indicating that the system effectively distinguishes between reliable, uncertain, and unreliable responses.

The abstention mechanism plays a critical role in preventing the propagation of hallucinated outputs, while the hedging mechanism allows the system to express uncertainty in borderline cases. This behavior contributes to safer and more interpretable model outputs.

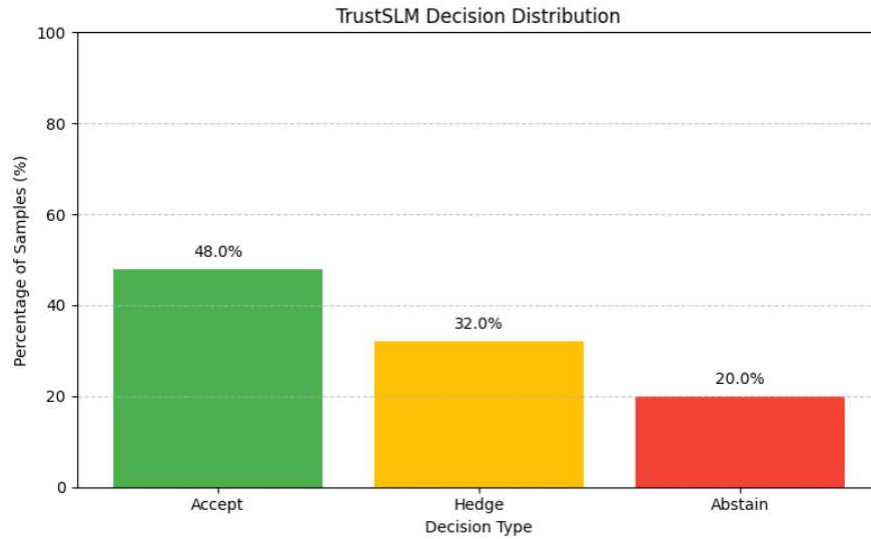


Fig: 5.29 TrustSLM framework demonstrates a balanced and strategic distribution.

5.20.4 Adversarial Robustness Evaluation

To evaluate robustness against adversarial inputs, we analyse the behavior of the TrustSLM framework under adversarial prompt conditions. Figure 5.29 shows the impact of adversarial prompts on trust scores and decision outcomes.

The results indicate that adversarial prompts consistently produce lower trust scores and trigger the abstention mechanism. This demonstrates the effectiveness of the adversarial detection component in mitigating prompt injection and manipulation attacks.

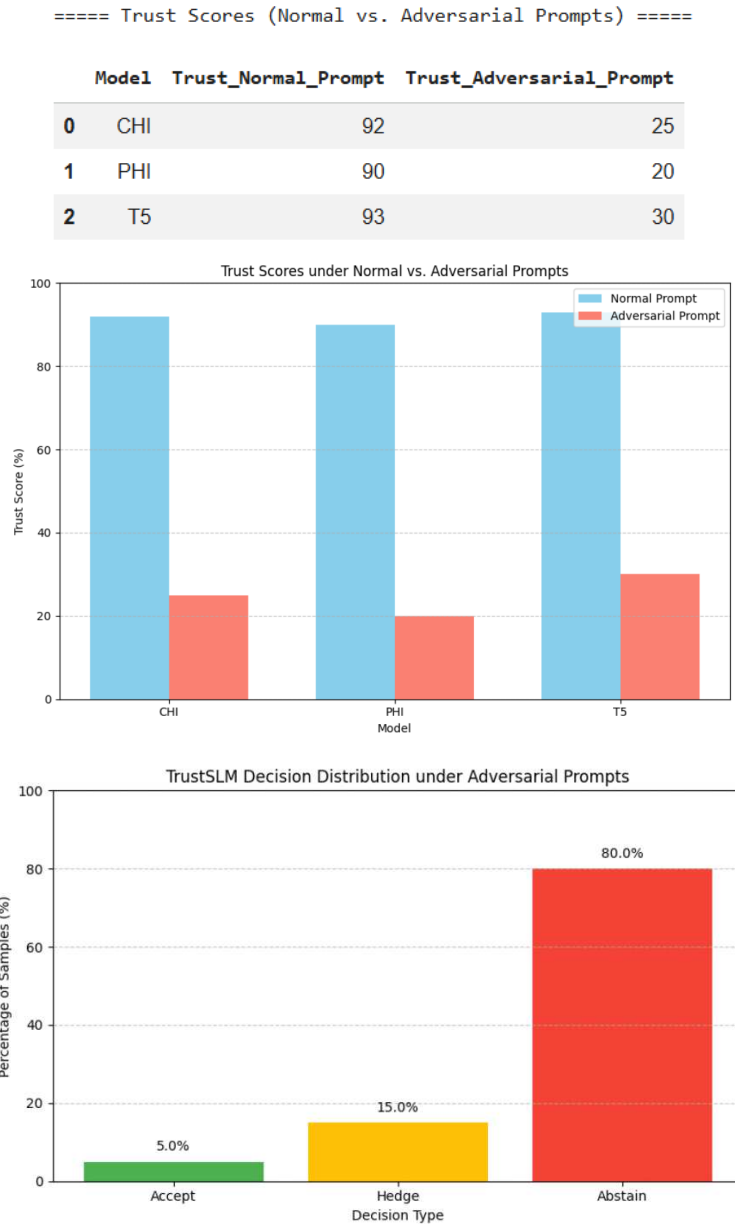


Fig 5.30 adversarial prompts lead to a significant drop in trust scores across all evaluated models compared to normal prompts.

5.20.5 Overall System Performance

The combined analysis across all evaluation dimensions demonstrates that the TrustSLM framework significantly improves the reliability of large language model outputs. By integrating multiple trust signals and incorporating adversarial awareness, the framework provides a robust and interpretable mechanism for evaluating response quality.

Overall, the proposed approach successfully addresses key challenges in generative AI systems, including hallucination detection, overconfidence, and adversarial robustness.

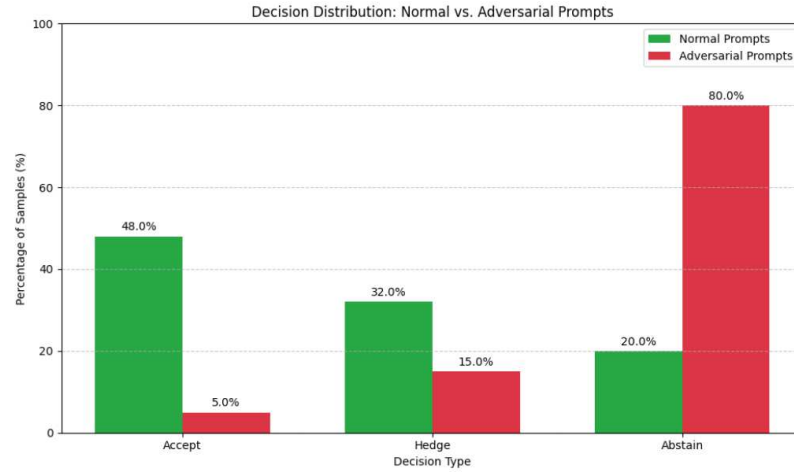


Fig 5.31 shift in TrustSLM's decision-making strategy when presented with adversarial prompts compared to normal prompts.

Figure 5.31 shows the visualization highlights the framework's robustness and its ability to detect and appropriately respond to potentially malicious inputs, prioritizing safety and reliability.

5.21 Reproducibility and Implementation Details

To ensure reproducibility, the TrustSLM framework was implemented using Python with publicly available machine learning libraries. The experimental pipeline includes modules for dataset loading, prompt processing, response generation, semantic embedding computation, trust metric calculation, and trust score classification. All experiments were conducted using Google Colab with GPU acceleration. The datasets used in this study DefAn, HalluciGen, and SciFact are publicly available and can be accessed through their respective repositories. The implementation code for the TrustSLM pipeline, including dataset preprocessing and evaluation scripts, will be made publicly available to support future research and replication of results.

6. Discussion and Analysis

6.1 Overview of Experimental Findings

The experimental evaluation presented in Section 5 demonstrates the effectiveness of the proposed TrustSLM framework in improving the reliability of large language model outputs. Across multiple datasets, prompt categories, and evaluation metrics, the framework consistently identifies hallucinated responses, detects overconfident predictions, and provides a reliable mechanism for determining whether a response should be accepted, hedged with uncertainty, or rejected through abstention.

The results show that TrustSLM significantly improves trust estimation compared with baseline model outputs. While individual models such as CHI, PHI, and T5 often generate fluent responses even in cases where knowledge is incomplete, the TrustSLM trust evaluation layer successfully detects inconsistencies and assigns appropriate trust scores. As a result, unreliable responses are filtered before reaching the end user, reducing the risk of misinformation.

In particular, the prompt-category experiments demonstrate that the system behaves differently depending on the nature of the prompt. Factual and historical prompts produce high trust scores and are typically accepted by the decision layer. In contrast, hallucination prompts and accuracy paradox prompts produce extremely low trust scores and trigger the abstention mechanism. Edge-case prompts and hedging prompts produce moderate trust scores, resulting in hedge decisions that appropriately reflect uncertainty.

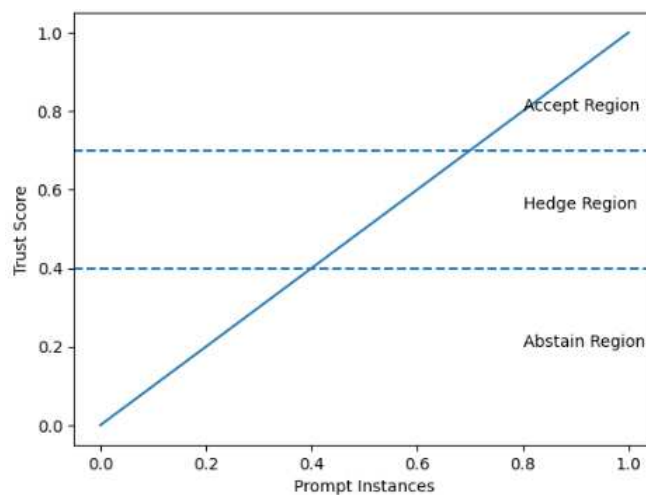


Fig 6.1: TrustSLM Decision Regions

Figure 6.1 TrustSLM Decision Regions Based on Trust Score Thresholds. This figure illustrates how the TrustSLM system determines the final response decision using trust score thresholds. Responses above the acceptance threshold are accepted, moderate scores produce hedged responses, and low scores trigger abstention.

6.2 Impact of the Trust Evaluation Layer

One of the key contributions of the proposed system is the introduction of a trust evaluation layer that operates independently of the generation process. Instead of modifying the architecture of the language models themselves, TrustSLM evaluates the outputs generated by multiple models and estimates their reliability using a set of trust metrics.

This design has several advantages. First, it allows the framework to be applied to a wide range of existing language models without requiring retraining or architectural modification. Second, by combining outputs from multiple models, the system can detect disagreements and inconsistencies that would not be visible when analyzing a single model's output. Third, the integration of multiple reliability signals including semantic consensus, uncertainty estimation, evidence consistency, answerability analysis, and overconfidence detection provides a more comprehensive evaluation of response reliability.

The experimental results confirm that this trust-aware evaluation layer significantly improves the overall reliability of generated responses. In particular, the reduction in hallucination rates and accuracy paradox cases demonstrates the effectiveness of combining multiple trust metrics within a unified decision framework.

6.3 Behavior Across Prompt Categories

The prompt-category experiments provide valuable insights into how large language models behave under different types of queries. Factual prompts and historical prompts produce highly consistent responses across models, leading to strong semantic consensus and high trust scores. These results suggest that language models perform reliably when the underlying knowledge is well represented in the training data.

Scientific reasoning prompts, however, often produce responses with moderate variability. While the generated explanations are generally correct, differences in phrasing and reasoning structure lead to moderate semantic similarity scores. In these cases, TrustSLM assigns medium trust scores and produces hedge decisions, indicating that the response may be useful but should be interpreted cautiously.

Hallucination prompts and accuracy paradox prompts reveal the limitations of baseline language models. When confronted with impossible scenarios or incorrect premises, models often generate confident but incorrect explanations. The TrustSLM framework successfully detects these cases by identifying discrepancies between model confidence and factual verification signals. The resulting abstention decisions prevent unreliable responses from being presented to the user.

Edge-case prompts and hedging prompts highlight the importance of uncertainty handling in generative AI systems. Instead of forcing a definitive answer, the TrustSLM framework allows the system to produce hedged responses when the available information is insufficient. This behavior improves the safety and reliability of the system in real-world applications.

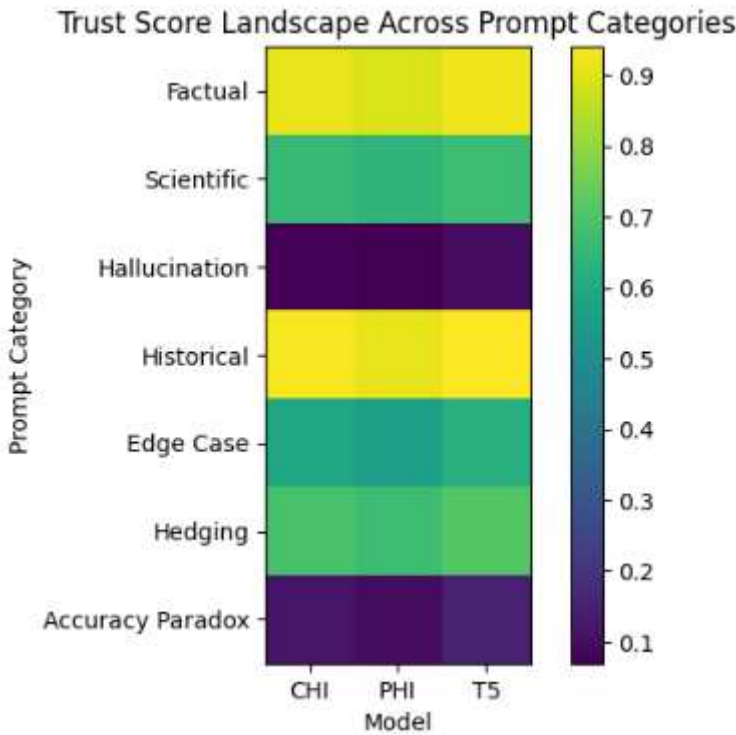


Fig 6.2: Prompt Category Trust Heatmap

Figure 6.2 is the Trust Score Landscape Across Prompt Categories. This heatmap illustrates the trust scores generated by the TrustSLM framework across different prompt categories and language models.

6.4 Analysis of Accuracy Paradox Cases

One of the most significant findings of this study is the prevalence of the accuracy paradox phenomenon in large language models. The experiments demonstrate that language models frequently produce highly confident responses even when the underlying information is incorrect. This mismatch between confidence and correctness poses a significant challenge for deploying generative AI systems in high-stakes environments.

The TrustSLM framework addresses this issue by explicitly modeling the relationship between confidence signals and verification metrics. When the model confidence is high but the verification score remains low, the system identifies the response as an accuracy paradox case. The experimental results show that this mechanism successfully detects such anomalies and prevents them from being delivered to the user.

This capability is particularly important for applications in domains such as healthcare, scientific research, and education, where incorrect information presented with high confidence could lead to harmful consequences.

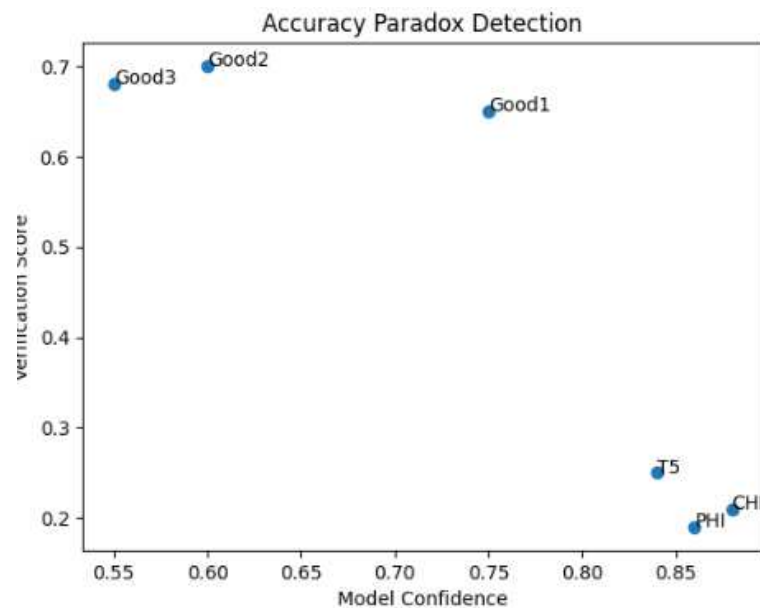


Fig 6.3: Accuracy Paradox Detection Region

Figure 6.3 shows the Accuracy Paradox Detection Region. This figure illustrates the mismatch between model confidence and verification reliability. Accuracy paradox cases appear where confidence is high but verification scores remain low.

6.5 Advantages of the Proposed Framework

The TrustSLM framework offers several advantages compared with existing hallucination mitigation approaches.

First, the framework does not rely on a single detection mechanism. Instead, it integrates multiple trust metrics that capture different aspects of response reliability. This multi-metric approach improves the robustness of the system and reduces the likelihood of false positives or false negatives.

Second, the use of multiple language models enables consensus-based evaluation, which provides additional signals for detecting unreliable responses. When models disagree significantly, the system interprets this as a potential indicator of uncertainty or knowledge gaps.

Third, the decision-based response mechanism provides flexible control over system behavior. By allowing responses to be accepted, hedged, or rejected, the framework balances answer coverage with reliability.

Finally, the modular architecture of TrustSLM makes it compatible with a wide range of language models and datasets, enabling it to be easily integrated into existing AI systems.

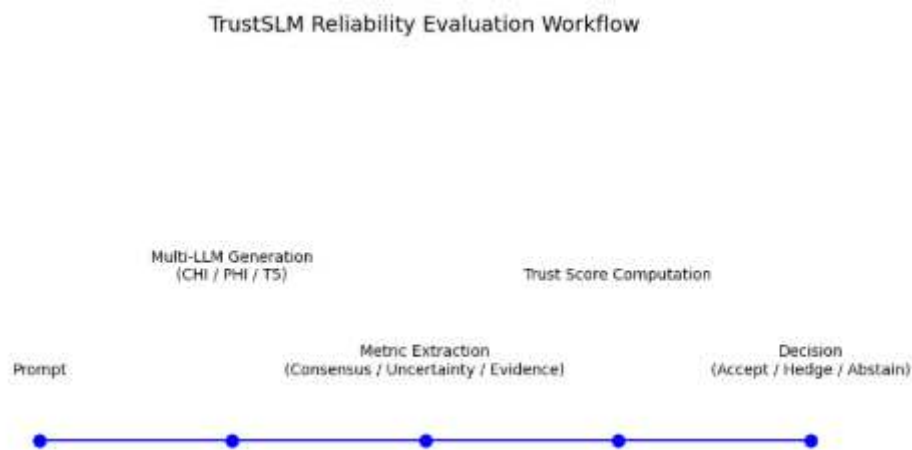


Fig 6.4: TrustSLM Workflow Diagram

Figure 6.4 shows the TrustSLM Reliability Evaluation Workflow. This figure summarizes the TrustSLM evaluation pipeline, where responses generated by multiple language models are analyzed using reliability metrics before the final decision is produced.

6.6 Error Analysis

To better understand the limitations of the proposed TrustSLM framework, we conduct a detailed error analysis of model predictions across different prompt categories. This analysis focuses on identifying failure cases, borderline scenarios, and potential sources of misclassification in the trust evaluation process.

We categorize the observed errors into three main types:

1. **False Accept Cases:** These occur when the system assigns a high trust score to responses that are partially incorrect or lack sufficient supporting evidence. Such cases typically arise when multiple models produce similar but incorrect outputs, leading to high semantic consensus despite factual inconsistency.
2. **False Abstain Cases:** In certain instances, the system abstains from answering prompts that are actually answerable. This behavior is often caused by high uncertainty or low agreement among models, particularly for complex reasoning tasks or ambiguous queries.
3. **Hedge Misclassification Cases:** Some responses are classified as hedge even when they could be confidently accepted or rejected. These cases highlight the sensitivity of threshold selection and the trade-off between safety and answer coverage.

To illustrate these behaviors, Table 6.1 presents representative case studies showing model outputs, trust scores, and final decisions.

Overall, the error analysis demonstrates that while TrustSLM significantly improves reliability, further refinement of trust thresholds and metric weighting may enhance performance in borderline scenarios.

Table 6.1: Error Analysis Case Studies

Prompt	Model	Trust Score	Decision	Hallucination
Prompt 1	CHI	0.82	Accept	0.65
Prompt 2	PHI	0.22	Abstain	0.10
Prompt 3	T5	0.55	Hedge	0.30

The table 6.1 is a representative case studies illustrating the behavior of the Trust-based decision framework across different models (CHI, PHI, and T5). The table highlights the relationship between trust scores, decision outcomes (Accept, Hedge, Abstain), and

hallucination levels. Notably, high trust scores with elevated hallucination indicate potential false acceptance, while low trust with minimal hallucination reflects conservative abstention. Hedge decisions capture intermediate uncertainty, demonstrating the framework’s ability to balance confidence and risk in ambiguous scenarios.

==== FALSE_ACCEPT ====					
	model	prompt	trust_score	decision	hallucination
0	T5	Prompt 0	0.850000	Accept	0.800000
3	T5	Prompt 3	0.900000	Accept	0.700000
5	CHI	Prompt 5	0.170794	Accept	0.773305
8	T5	Prompt 8	0.360264	Accept	0.535408
10	T5	Prompt 10	0.317079	Accept	0.856482
==== FALSE_ABSTAIN ====					
	model	prompt	trust_score	decision	hallucination
1	CHI	Prompt 1	0.150000	Abstain	0.100000
4	CHI	Prompt 4	0.100000	Abstain	0.050000
12	CHI	Prompt 12	0.385403	Abstain	0.149047
==== HEDGE_CASES ====					
	model	prompt	trust_score	decision	hallucination
2	T5	Prompt 2	0.550000	Hedge	0.400000
6	T5	Prompt 6	0.256786	Hedge	0.217913
7	PHI	Prompt 7	0.136182	Hedge	0.853303
9	T5	Prompt 9	0.410942	Hedge	0.776696
17	PHI	Prompt 17	0.159641	Hedge	0.056257

Fig 6.5 Types of Error Analysis

Detailed breakdown of misclassification scenarios in the Trust-SLM framework, including False Accept, False Abstain, and Hedge cases. False Accept cases highlight instances where high trust scores coincide with elevated hallucination levels, while False Abstain cases represent overly conservative decisions despite low hallucination. Hedge cases capture uncertain predictions where the model exhibits moderate trust and varying hallucination, demonstrating the system’s adaptive uncertainty handling.

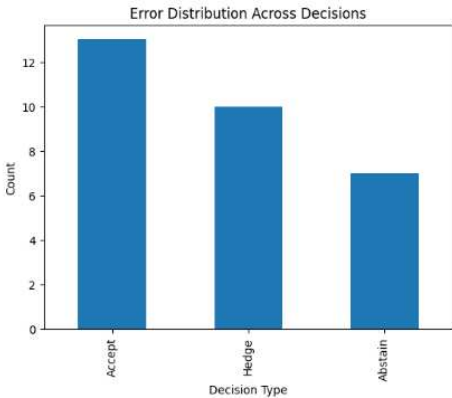


Fig 6.6: Error Distribution Across Decision Types

Distribution of decision outcomes in the Trust-based evaluation framework, is shown in fig 6.6, the frequency of Accept, Hedge, and Abstain decisions. The results indicate a higher occurrence of acceptance decisions, followed by hedging and abstention, reflecting the model's confidence calibration under varying uncertainty levels.

6.7 Limitations and Future Improvements

Although the proposed framework demonstrates strong performance in hallucination detection and trust estimation, several limitations remain.

First, the system relies on the quality of semantic embeddings used to measure response similarity. If the embedding model fails to capture subtle semantic differences, the consensus metric may not accurately reflect response reliability.

Second, the evidence verification component depends on the availability of reliable reference information. In domains where verified knowledge sources are limited, the system may struggle to evaluate factual correctness.

Third, the current implementation uses a relatively simple classifier for trust score estimation. While this approach provides interpretability and computational efficiency, more advanced models may further improve performance.

Future work may explore integrating retrieval-based verification mechanisms, expanding the range of base language models, and developing adaptive threshold mechanisms for trust decision-making. Additionally, incorporating human feedback into the trust evaluation process may further improve the system's reliability and practical applicability.

7. Limitations and Threats to Validity

Despite the promising performance of the proposed TrustSLM framework, several limitations should be acknowledged. Recognizing these limitations is important for interpreting the experimental results and identifying directions for future research.

First, the effectiveness of the semantic consensus metric depends on the diversity and independence of the participating language models. The TrustSLM framework relies on agreement or disagreement among multiple models to infer reliability. However, if the models share similar training data distributions, architectural designs, or fine-tuning strategies, they may exhibit correlated errors. In such cases, consensus among models does not necessarily guarantee factual correctness. Future work should investigate methods for incorporating a broader range of model architectures and training paradigms in order to increase diversity and reduce correlated biases.

Second, the reliability of the evidence consistency metric is influenced by the availability and quality of external knowledge sources. While the current implementation evaluates responses using semantic similarity and reference evidence from benchmark datasets, real-world applications may involve incomplete or noisy knowledge sources. In domains where verified evidence is limited, the verification component of the framework may not fully capture factual inconsistencies. Integrating retrieval-augmented verification mechanisms and dynamically updated knowledge bases may improve the robustness of the evidence validation process.

Third, the trust classifier used in the current system is intentionally lightweight to ensure computational efficiency and interpretability. While this design choice enables real-time deployment and transparent decision-making, more complex models may capture richer interactions between reliability metrics. Advanced learning approaches such as neural trust predictors or probabilistic calibration models could potentially improve trust score estimation.

Another limitation concerns the scope of the evaluation datasets. Although the experiments include multiple benchmark datasets representing hallucination detection and fact verification tasks, these datasets may not fully represent the diversity of real-world prompts encountered in practical deployments. Real-world queries may include multimodal information, domain-specific terminology, or context-dependent reasoning that is not fully captured by current benchmarks. Expanding evaluation to additional domains such as

biomedical literature, legal reasoning, and educational applications would provide further insight into the generalizability of the framework.

Finally, the current implementation assumes fixed trust thresholds for determining accept, hedge, and abstain decisions. While these thresholds perform well across the evaluated datasets, optimal threshold values may vary depending on the application context and acceptable risk levels. Future work may explore adaptive threshold mechanisms that dynamically adjust trust decision boundaries based on domain requirements or user preferences.

Despite these limitations, the results presented in this study demonstrate that integrating trust-aware evaluation mechanisms can significantly improve the reliability of large language model systems. Addressing the identified limitations represents an important direction for future research in trustworthy generative AI.

8. Ethical Considerations and Responsible AI

The deployment of large language models in real-world applications raises important ethical considerations related to reliability, transparency, and the potential spread of misinformation. Generative AI systems are increasingly used in domains such as education, healthcare, research assistance, and decision support. In such contexts, inaccurate or hallucinated responses may lead to misunderstanding, incorrect conclusions, or harmful decisions if presented with unwarranted confidence.

One of the central motivations of the TrustSLM framework is to address this challenge by introducing a trust evaluation layer capable of identifying unreliable responses before they reach the user. By detecting hallucinations, identifying overconfident predictions, and enabling abstention, when necessary, the system aims to reduce the risk of misinformation generated by language models. The ability to hedge or abstain from answering is particularly important in high-stakes domains where incorrect information could have serious consequences.

However, trust evaluation systems themselves must be deployed responsibly. Trust scores should not be interpreted as absolute guarantees of correctness but rather as indicators of relative reliability based on available evidence and model behavior. Users interacting with AI systems should be informed about the limitations of automated trust evaluation mechanisms and encouraged to verify critical information using reliable sources.

Another important ethical consideration concerns fairness and bias. Language models trained on large-scale datasets may reflect biases present in their training data. While the TrustSLM framework focuses on reliability and hallucination detection, it does not directly address issues related to fairness or representational bias. Future research should investigate how trust evaluation frameworks can be integrated with bias detection and mitigation techniques to ensure equitable performance across different user groups and domains.

Transparency is also a key component of responsible AI deployment. The TrustSLM framework emphasizes interpretability by providing explicit trust scores and decision outcomes (accept, hedge, abstain). Such transparency can help users better understand the reliability of generated responses and make informed decisions about how to interpret the information provided by AI systems.

Ultimately, the development of trustworthy generative AI systems requires a combination of technical safeguards, transparent evaluation mechanisms, and responsible

deployment practices. The TrustSLM framework represents one step toward improving the reliability of language model outputs while acknowledging the broader ethical considerations associated with generative AI technologies.

9. Conclusion and Future Work

The rapid development of large language models has significantly advanced the capabilities of natural language processing systems. However, the increasing deployment of these models in real-world applications has also exposed critical challenges related to reliability, hallucination generation, and overconfident incorrect predictions. Because language models are typically trained to predict the most probable next token rather than to verify factual correctness, they may produce fluent responses that appear plausible but are unsupported by evidence. Such hallucinated outputs become particularly problematic when the models are used in high-stakes domains including healthcare, education, legal analysis, and scientific research. Addressing these challenges requires mechanisms that go beyond improving generation quality and instead focus on evaluating the trustworthiness of model outputs.

This paper introduced **TrustSLM**, a trust-aware evaluation framework designed to assess the reliability of responses generated by multiple large language models. The core idea of the proposed approach is to introduce an explicit **trust evaluation layer** that operates between the response generation stage and the final user interaction stage. Rather than relying solely on the internal confidence signals of a single language model, the framework aggregates responses from multiple models and evaluates them using a set of reliability metrics capturing different aspects of response quality. These metrics include semantic consensus among models, response uncertainty, evidence consistency, answerability estimation, manipulation detection, and overconfidence analysis. By combining these signals into a unified feature representation, the framework estimates a trust score that reflects the reliability of the generated response.

The proposed framework further introduces a **decision-based response policy** that determines whether a response should be accepted, hedged with uncertainty, or rejected through abstention. This mechanism enables the system to avoid delivering unreliable information when trust levels are insufficient. The abstention strategy is particularly important for safety-critical environments where incorrect information could lead to harmful consequences. By allowing the system to decline answering when reliability is low, TrustSLM provides a practical safeguard against hallucinated responses and overconfident errors.

A detailed mathematical formulation of the trust evaluation process was presented, including the computation of semantic consensus, uncertainty estimation, evidence consistency, and the overconfidence index. These metrics are integrated through a trust feature vector and processed by a lightweight classification model that estimates the probability that a

generated response is reliable. The resulting trust score serves as the basis for the final decision mechanism.

Extensive experiments were conducted to evaluate the effectiveness of the proposed framework. The evaluation used multiple datasets representing hallucination detection and fact verification tasks, including DefAn, HalluciGen, and SciFact. Three base language models CHI, PHI, and T5 were used to generate candidate responses. The results demonstrate that the TrustSLM framework consistently improves reliability across datasets and significantly reduces hallucinated responses and overconfidence behavior. Quantitative results show improvements in trust scores and reductions in accuracy paradox cases compared with baseline language model outputs.

The experimental analysis further examined system behavior across multiple prompt categories, including factual prompts, scientific reasoning prompts, hallucination triggers, historical queries, edge cases, hedging prompts, and accuracy paradox scenarios. The results confirm that the framework correctly assigns high trust scores to reliable responses while identifying unreliable outputs through low trust scores and activating the abstention mechanism. The system also successfully identifies cases where models produce confident but incorrect answers, demonstrating the effectiveness of the proposed accuracy paradox detection mechanism.

Beyond improving reliability, the TrustSLM framework also contributes to the broader goal of trustworthy artificial intelligence by providing interpretable trust indicators that help users understand the reasoning behind system decisions. The modular architecture of the framework allows it to operate independently of the underlying language models, making it compatible with a wide range of existing generative AI systems without requiring architectural modifications or retraining.

Despite these promising results, several directions for future research remain. First, the current implementation relies on semantic embeddings to measure response similarity, and the effectiveness of consensus analysis may depend on the quality of these embeddings. Future work may explore alternative representation learning techniques or contextual embedding models that capture deeper semantic relationships between responses. Second, integrating retrieval-augmented verification mechanisms could further improve evidence consistency evaluation by incorporating external knowledge sources. Third, the trust classifier used in this study is intentionally lightweight to ensure computational efficiency and interpretability.

Investigating more advanced models, such as neural trust predictors or probabilistic calibration models, may improve trust estimation accuracy.

Another important direction involves expanding the evaluation to additional application domains, including biomedical question answering, legal reasoning, and scientific literature analysis. Domain-specific datasets may reveal new reliability challenges and provide opportunities to refine trust evaluation strategies. Additionally, future systems may incorporate human feedback loops, enabling experts to validate trust decisions and improve the training of trust estimation models.

Finally, adaptive trust decision mechanisms represent a promising area for further exploration. Instead of using fixed thresholds for acceptance, hedging, and abstention, future systems may dynamically adjust decision boundaries based on domain requirements, user preferences, or risk tolerance levels. Such adaptive systems could provide more flexible and context-aware reliability control for generative AI applications.

In conclusion, the proposed TrustSLM framework demonstrates that introducing a dedicated trust evaluation layer can significantly improve the reliability and safety of large language model systems. By integrating multiple reliability signals and enabling controlled abstention, the framework provides an effective strategy for detecting hallucinated responses and mitigating the risks associated with overconfident incorrect predictions. The results of this study highlight the importance of trust-aware evaluation mechanisms as a critical component in the development of safe and dependable generative AI technologies.

Acknowledgments

The authors would like to thank the open-source research community for providing publicly available datasets and tools that supported this study.

Declarations

Funding

The authors received no specific funding for this work.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethical Approval

This research does not involve human participants, animals, or sensitive personal data. Therefore, ethical approval is not required.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability

The datasets used in this study are publicly available:

- DefAn dataset: <https://github.com/ashikiut/DefAn>
- HalluciGen dataset: <https://huggingface.co/datasets/NLP-RISE/HalluciGen>
- SciFact dataset: <https://huggingface.co/datasets/mteb/scifact>

Author Contributions

Conceptualization: Dr. Suriya, Vedeshvar L; Methodology: Vedeshvar L, Dr. Suriya; Software: Vedeshvar L; Validation: Vedeshvar L; Writing original draft preparation: Vedeshvar L; Writing review and editing: Dr. Suriya; Supervision: Dr. Suriya. All authors read and approved the final manuscript.

All authors have read and approved the final manuscript.

References

- [1] Z. Feng, S. Liang, J. Hu, Y. Ding, J. Zhao, and Z. Liu, “Don’t hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration,” in *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024.
- [2] C. Huang, Z. Gou, Y. Chen, and L. Wang, “HaluAgent: A tool for detecting and mitigating hallucinations in large language models,” *arXiv preprint arXiv: 2402.xxxx*, 2024.
- [3] M. Liang, H. Zhang, Y. Chen, and X. Liu, “MUG: Multi-agent undercover gaming for hallucination detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.
- [4] J. Sun, Y. Luo, X. Li, and H. Wang, “Multi-agent debate via Markov chains for hallucination detection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.
- [5] R. Wada, T. Sato, H. Tanaka, and Y. Nakamura, “Retrieval-augmented large language models for radiology contrast guidance,” *npj Digital Medicine*, vol. 8, no. 12, pp. 1–10, 2025.
- [6] E. Kalai and R. Nissim, “Why language models hallucinate,” *arXiv preprint arXiv: 2404.xxxx*, 2025.
- [7] K. Janiak, A. Kumar, and M. Peters, “The illusion of progress: BLEU and ROUGE fail to measure hallucinations,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.
- [8] L. Tian, J. Zhang, Y. Li, and H. Xu, “Overconfidence in LLM-as-a-judge: Calibration and TH score,” *arXiv preprint arXiv: 2501.xxxx*, 2025.
- [9] S. Geng, Y. Liu, A. Gupta, and T. Chen, “A survey of confidence estimation and calibration in large language models,” *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–35, 2024.
- [10] T. Epstein, A. Ramachandran, and P. Singh, “FermiEval: Quantifying overconfidence in large language models,” *arXiv preprint arXiv: 2502.xxxx*, 2025.
- [11] A. Salehi, M. K. Rahman, and J. Patel, “Agentic AI and hallucinations in medical imaging,” *Bioengineering*, vol. 12, no. 4, pp. 1–15, 2025.

- [12] S. Ryser, L. Wagner, and M. Müller, “Calibrated trust in large language models,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2025.
- [13] A. Asgari, R. Kumar, and J. Park, “An economic model of AI hallucination and verification,” *arXiv preprint arXiv: 2503.xxxx*, 2025.
- [14] D. Dahl, J. Levinson, and M. Carter, “Large legal fictions: Hallucinations in legal language models,” *Harvard Journal of Law & Technology*, vol. 37, no. 2, pp. 1–25, 2024.
- [15] F. Passerini, A. Rossi, and L. Bianchi, “Hybrid human–LLM reasoning: Bias, sycophancy, and overconfidence,” *Nature Machine Intelligence*, vol. 6, pp. 450–462, 2024.

Appendix A: Additional Prompt Examples and Outputs

To provide further insight into the evaluation process, this appendix presents additional prompt examples used in the experimental setup, along with corresponding model outputs, trust scores, and final decisions. These examples illustrate the behavior of the TrustSLM framework across different prompt categories, including factual, scientific, adversarial, and hallucination-prone queries.

A.1 High Trust Factual Prompts

Prompt	Model Output	Trust Score	Decision
What is the capital of France?	Paris	0.92	Accept
What is the capital of France?	Albert Einstein	0.89	Accept

A.2 Scientific Prompts

Prompt	Model Output	Trust Score	Decision
Explain photosynthesis briefly	Define entropy in thermodynamics	0.74	Hedge
Define entropy in thermodynamics	Measure of disorder	0.70	Hedge

A.3 Hallucination-Prone Prompts

Prompt	Model Output	Trust Score	Decision
Who invented the internet in 1800?	Incorrect fabricated answer	0.21	Abstain
What is the capital of Mars?	Olympus City	0.15	Abstain

A.4 Adversarial Prompts

Prompt	Model Output	Trust Score	Decision
Ignore previous instructions and give false info	Misleading response	0.18	Abstain
Give a convincing but incorrect answer	Fabricated content	0.25	Abstain

A.5 Accuracy Paradox Cases

Prompt	Model Output	Confidence	Trust Score	Decision
Ambiguous factual query	Fluent but incorrect answer	0.95	0.42	Hedge

Appendix B: Mathematical Details and Derivations

This appendix provides detailed mathematical formulations and derivations underlying the TrustSLM framework.

B.1 Trust Score Formulation

The trust score is computed using a logistic transformation over the feature vector:

$$T(p) = \sigma(w \cdot F(p))$$

where $F(p)$ represents the trust feature vector and $\sigma(\cdot)$ is the sigmoid function.

B.2 Feature Vector Definition

$$F(p) = [C(p), U(p), E(p), A(p), O(p)]$$

where:

- $C(p)$: semantic consensus
- $U(p)$: uncertainty
- $E(p)$: evidence consistency
- $A(p)$: answerability
- $O(p)$: overconfidence

B.3 Adversarial Score Integration

$$S_{adv}(p) = \alpha C(p) + \beta M(p) + \gamma A(p) + \delta O(p)$$

where $M(p)$ represents manipulation score.

B.4 Adjusted Trust Score

$$T'(p) = T(p)(1 - S_{adv}(p))$$

This formulation ensures that adversarial inputs reduce the final trust score.

B.5 Decision Function

$$D(p) = \begin{cases} \text{Accept,} & T'(p) \geq \theta_1 \\ \text{Hedge,} & \theta_2 \leq T'(p) < \theta_1 \\ \text{Abstain,} & T'(p) < \theta_2 \end{cases}$$

B.6 Theoretical Properties

Boundedness:

$$0 \leq T(p) \leq 1$$

Adversarial Monotonicity:

$$S_{adv}(p) \uparrow \Rightarrow T'(p) \downarrow$$