E-ISSN:3108-1711
IRJSRR.ORG
INTERNATIONAL RESEARCH JOURNAL OF SCIENTIFIC REPORTS AND REVIEWS
Open Access | Peer Reviewed | Online Journal

HALLUCINATION IN LARGE LANGUAGE MODELS: CHARACTERIZATION, DETECTION, AND MITIGATION APPROACHES

[1]Meenal Vardar, [2]Mayank Sharma*, [3]Dimpal Agrawal [4]Ankur Vashistha
[1]Research Scholar, Department of Computer Science, JECRC University, Jaipur
[2]Research Associate, Department of Computer Science, Kalp Laboratories, Mathura, Uttar Pradesh.
[3]Research Associate, Department of Computer Science, Kalp Laboratories, Mathura, Uttar Pradesh.
[4]Founder, Mukti Ecosmart Technologies Pvt Ltd,
*Correspondence E-mail: mynkshrm0@gmail.com

Abstract: A significant barrier to preserving factual accuracy and dependability in AI-generated outputs is hallucination in large language models. Using a benchmark Kaggle dataset, this work provides a comprehensive evaluation of both advanced transformer-based architectures and traditional machine learning classifiers for hallucination identification. They compared refined transformer models, such as DistilBERT, RoBERTa, and DeBERTa, with baseline models, including Random Forest, SVM, and Logistic Regression. The results show that transformer-based models were more robust and better at understanding context; however, more conventional models, such as Random Forest, achieved a high overall accuracy of 94.10%. DistilBERT struck a wonderful balance between precision and readability. The confusion matrix analysis demonstrated that the models helped reduce false alarms for non-hallucination outputs. The ROC-AUC ratings confirmed the transformers' precision and capability for identifying a slight rate of semantic discrepancies. Other studies provided supporting evidence that deeper context modelling will provide real benefits to the reliability of detection rates, demonstrated by the reduced hallucinations and assessments of the frequency of errors made. In conclusion, this research shows that combining traditional and modern approaches is beneficial and that tuning with transformer models holds promise for reducing hallucinations. This research provides an example of early steps of increasing trustworthiness and human-like models as AI models.

Keywords: Hallucination Detection, Large Language Models, Transformer-Based Models, Machine Learning, Trustworthy AI.

Received: 11-Jul-2025 **Revised:** 08-Aug-2025 **Accepted:** 12-Sep-2025
Citation for the Paper: Vardar Meenal, Mayank Sharma, Dimpal Agrawal, and Ankur Vashistha. "Hallucination in Large Language Models: Characterisation, Detection, and Mitigation Approaches." *International Research Journal of Scientific Reports and Reviews* 1, no. 1 (September 2025): 1–17.

1. Introduction

Large Language Models (LLMs), which include GPT and BERT (along with their derivatives), have shown the capacity to comprehend and produce authentic language. This has led to many outstanding applications for these types of models in various areas such as scientific research, healthcare, education, conversational systems, and information retrieval [1]. Although LLMs may perform satisfactorily, hallucinations are a significant challenge that undermines their reliability. Hallucinations occur when a model produces outputs that are factually incorrect, illogical, or wholly fabricated [2]. Hallucinations are particularly problematic when employing LLMs for high-stakes areas such as law, health, and finance, where presenting inaccurate information can have dire consequences [3]. Therefore, the field of hallucination understanding, diagnosis, and intervention has emerged as a key area of research, generating interest in both academic and commercial areas of study [4]. This review article will describe the elements of hallucinations with LLMs, summarise existing methods for detecting and mitigating hallucinations, and share possible ways of focusing further research on LLM prompt safety and reliability.

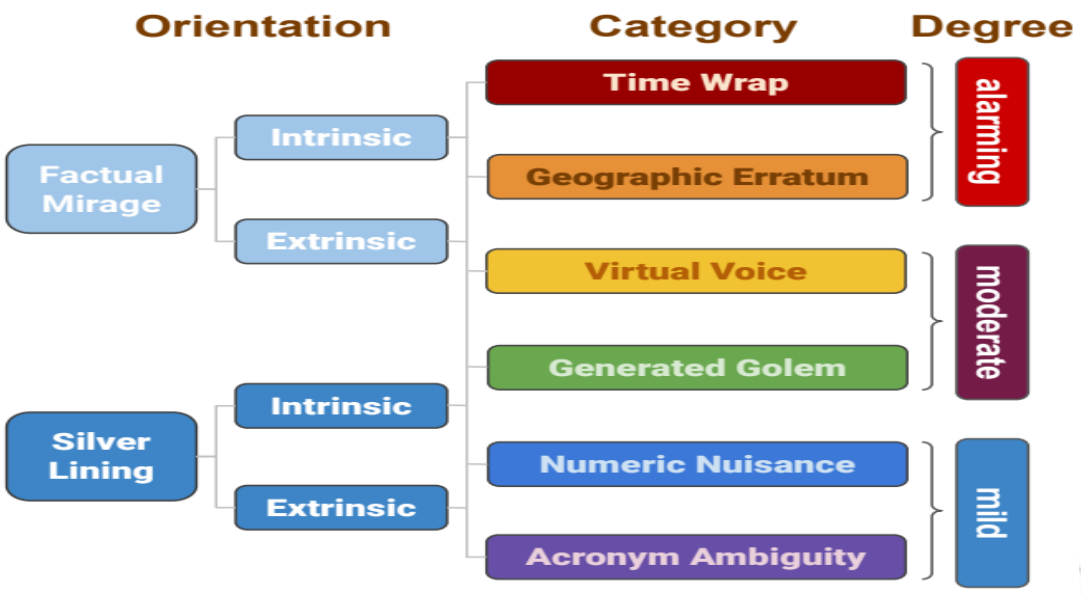

Figure 1: Framework of AI Hallucinations: Orientation, Categories, and Severity [11]

Figure 1 illustrates a systematic approach to classify hallucinations, which breaks down their direction, category, and severity or magnitude of effects in large language models (LLMs). Common classifications for hallucinations include the Silver Lining classification and the Factual Mirage classification, which are further broken down into intrinsic and extrinsic orientations, producing yet additional groupings: Time Wrap and Geographic Erratum, being highly alarming and very meaningful factual distortions that could produce a lot of misleading information. Moderate or lesser hallucinations also include instances of Virtual Voice or Generated Golem errors, where the fallacy is related to fake objects or characters. The numeric nuisance and acronym ambiguity simply refer to less severe errors or misinformation. While they use the terms spectrum related to having and expressing less or more concern, this example provides a way to provide discussion and ultimately holistically begin to rate the seriousness and potential risk to the hallucinations from AI. This classification also emphasises how mitigation strategies may also be ranked by level of potential to undermine trust and reliability (to the degree of severity of hallucination).

Natural language processing has undergone an overhaul due to the popularisation of Large Language Models (LLMs) such as GPT, BERT, and even more complex versions of these models. They set the stage for conversation systems, automated reasoning, content generation, and question answering to take advantage of natural language processing. Although LLMs show remarkable performance and capability, LLMs are still vulnerable to hallucinations. Hallucination is when the models produce results which are false, misleading, or made up. In particular scenarios - for example, medicine, law, and science - where precision is essential, and LLMs suffer from hallucinations, these models lack reliability and trustworthiness. Therefore, it is of the utmost importance to determine what causes hallucinations and to set definitions and understand the implications of hallucinations. This study characterises hallucination behaviour through a systematic analysis across datasets, and it evaluates existing methodologies for hallucination detection and mitigation, subsequently suggesting additional approaches to improve factual consistency. Ultimately, this will enable the responsible usage of LLMs and ensure their safety and trustworthiness when they are deployed in the real world.

The significance of this research is particularly heightened by the increasing use of large language models (LLMs) in systems that support decision making, as even slight errors can lead to erroneous systems, change the outcome of decisions, or diminish user trust [7]. While there has been a surge in recent research focusing on improving model performance and eloquence, less research

has systematically studied the detection and mitigation of hallucinations across multiple disciplines and datasets. This review will provide clarity about the causal factors associated with hallucinations (e.g., data sparsity, bias in training, limitations of contextual understanding) so that a well-established basis may be developed to 1. detect the hallucination as it occurs, and 2. develop interventions to minimise or counter the hallucination in an organised method. This research is valuable in developing linguistically more advanced, more interpretable, and more dependable models that incorporate human considerations and factual constraints that exist in the real world.[8]

While the possibilities for Large Language Models (LLMs) have been developing, another profoundly serious limitation to their usefulness is hallucination [9]. LLMs can demonstrate partial, complete, or factually incorrect performance, but are also grammatical and reasonably coherent to the prompt. This is particularly concerning in applications for medical diagnosis, legal document drafting, and retrieving information in scientific disciplines, where there has to be greater accuracy. The existing body of work on hallucination detection and mitigation is often not particularly generalizable, and more generally tends to be specific to a domain or dataset [10]. There is no existing agreed-upon standard measure for either hallucination (prevalence or degree), and therefore, it is difficult to assess the reports of hallucinations.

Thereby, if LLMs are to be accepted and reliable tools for real-world use, effective solutions for hallucination detection and subsequent mechanisms for minimising or mitigating the hallucination effects need to be developed. LLMs raise significant issues with hallucinations. Determining the existence of even inaccuracies or a set of things that could be factual but false could even be challenging. There are datasets for this purpose; however, they are not particularly representative of real-world observations and hallucination events, and hallucination events of LLMs in general will exist unevenly across domains [11]. Apart from hallucination events that LLMs produce, overall, the fidelity of LLMs is low. The purpose of this work is to examine hallucinatory behaviour in detail, review recent detection mechanisms, and offer workable solutions to repair hallucination behaviour, with imperfect but less biased LLMs, which can decrease hallucinatory outputs and build confidence in the use of LLMs more safely.

To support a more comprehensive understanding of hallucinations and provide a more holistic view of them in Large Language Models (LLMs) across different domains and datasets [12], this article describes and provides some viewpoints on hallucinations (partial vs complete fabrications, internal vs extrinsic, etc), and the metrics used to identify and measure hallucinations. Additionally, it also describes existing techniques to detect and mitigate hallucinations [13]. This article is intended to provide a sense of classifications of hallucinations, as well as baseline and transformer-based models to detect hallucinations, and mitigation strategies, such as retrieval augmented generation, fine-tuning, and prompts [14]. Summary statistics (accuracy, precision, recall, F1-score, ROC-AUC) and anecdotal examples are provided to assist the reader in understanding classifications of hallucinations and how viable a detection and mitigation strategy is across different datasets. Ultimately, this study not only conceptualises hallucinations in LLMs but also provides an avenue to rigorous factual reliability in LLMs and makes the use of information more secure and defensible.

2. Literature Review

The examination of Large Language Models' (LLMs) hallucinations has drawn considerable attention due to its implications for trust, reliability, and deployment. Recent work has spanned several domains, including (but not limited to) textual generation, multimodal applications, and clinical applications, with a particular emphasis on understanding causes of hallucinations, detecting signs, and intervention techniques to mitigate them. This literature review synthesises key findings from the existing literature, emphasising detection methods, intervention methods, and challenges for further research, to provide a solid research base.

2.1.Hallucinations in Large Language Models

Hallucinations in Large Language Models (LLMs) refer to when they generate plausible-sounding information that is false or fabricated, creating a significant barrier for trustworthy use in common contexts. Cossio et al. (2025) [15] provided a general taxonomy of LLM hallucinations that classified errors as intrinsic and extrinsic, along with factuality and faithfulness, while identifying mentions, such as factual inaccuracy, ill logic, time confusion, ethical violations, and domain-specific hallucinations. Luo et al. (2024) [16] stressed the problems that hallucinations cause for practical application by looking at both detection and mitigation approaches for popular LLMs, including ChatGPT, Bard, and Llama. Mishra et al. (2024) [20] established fine-grained detection using the FavaBench benchmark and developed FAVA, a retrieval-augmented language model designed to rectify hallucinations and enhance factual correctness across many domains. Mündler et al. (2023) [19] concentrated on self-contradiction in instruction-tuned language models, creating a prompting-based detection and mitigation framework suitable for black-box models. In contrast, Omar et al. (2025) [18] emphasised adversarial hallucinations in clinical settings, demonstrating that prompt engineering can diminish, though not completely eradicate, hallucination rates.

2.2.Hallucinations in Vision-Language and Foundation Models

Large vision-language models (LVLMs) and foundation models (FMs) encounter similar issues of hallucination in multimodal and general-purpose applications. Zhou et al. (2023) [17] examined object hallucination in LVLMs and introduced the LVLM Hallucination Revisor (LURE), which reconstructs descriptions with reduced hallucination by using co-occurrence, uncertainty, and positional elements, resulting in a 23% enhancement in assessment metrics. Rawte et al. (2023) [24] conducted a study on hallucination phenomena in foundation models (LFMs), formulating assessment criteria, categorising hallucination forms, and examining mitigation measures to guarantee secure deployment. These findings show that hallucinations are a problem that affects text, picture, and multimodal systems. This means that all sorts of models need to be tested and fixed in a strong way.

2.3.Detection and Mitigation Strategies

There have been several ideas put out for how to find and fix hallucinations in LLMs. Gumaan et al. (2025) [21] established a theoretical framework by presenting hallucination risk metrics and providing integrated detection and mitigation procedures that include token-level uncertainty, attention alignment assessments, retrieval-augmented generation, and hallucination-aware fine-tuning. Zhang et al. (2025) [22] examined hallucination in retrieval-augmented generation (RAG) systems, elucidating the origins of hallucination during both the retrieval and generation phases, and developing frameworks for the detection, correction, and mitigation of hallucinations in real-time applications. Lavrinovics et al. (2025) [23] underscored the significance of knowledge graphs (KGs) in augmenting LLM reliability via the provision of structured factual context, while also acknowledging the persistent issues of knowledge integration, dataset accessibility, and assessment. These methods all show that they need to combine model-level changes, retrieval methods, and organised knowledge to make facts more consistent and stronger.

Table 1: Summary of Recent Studies on Hallucination Detection and Mitigation in Large Language Models

Author	Methods	Key Findings	Research Gaps
Zhou et al. (2025) [25]	HaDeMiF with D3T + MLP for hallucination detection and calibration	Reduces hallucinations with <2% parameter overhead	Limited evaluation across diverse LLM tasks
Gokcimen et al. (2025) [26]	Cross-LLM security framework using VectorDB, Kernel, and RAG	Achieves 98 % accuracy; enhances reliability and mitigates hallucinations	Applicability to broader LLM tasks and real-world deployment scenarios remains to be tested
Clement, Mateo. Et al. (2025) [27]	LLM-based framework integrated with DevSecOps for automated threat detection and mitigation	Enhances software security with scalable, context-aware detection; outperforms traditional tools.	Challenges remain with false positives, adversarial prompts, and hallucinations.
Lan et al. (2024) [28]	Survey of LVLM hallucination causes and mitigation	Summarises correction methods and evaluation benchmarks	Limited practical dependability studies
Wang et al. (2023) [29]	HaELM, an LLM-based framework for hallucination evaluation in LVLMs	Achieves ~95 % accuracy; low-cost, reproducible, and privacy-preserving	Limited analysis of hallucination factors and mitigation strategies in LVLMs
Malin et al. (2025) [30]	LLM-based faithfulness evaluation; RAG and prompting	Metrics align with human judgment; mitigation improves faithfulness	Needs broader task evaluation
Ahmad et al. (2023) [31]	An exploratory study evaluating ChatGPT 3.5 responses with students	Provides inconsistent answers; unreliable for language/literature learning	Limited reliability in educational applications
Mohammed et al. (2024) [32]	Comparative analysis of hallucinations in GPT-3.5 and GPT-4	GPT-4 improves over GPT-3.5, but hallucinations persist	Further research is needed on AI robustness, interpretability, and ethical implications

3. Dataset Collection and Preprocessing

The efficacy of hallucination detection and mitigation measures in Large Language Models (LLMs) is significantly influenced by the quality, variety, and representativeness of the used datasets. This work employs many publicly accessible datasets to guarantee extensive representation of diverse domains, task categories, and hallucination patterns.

3.1.Dataset Selection

To thoroughly evaluate hallucination detection and mitigation strategies in Large Language Models, this study examined two datasets. The Kaggle Hallucination Detection Dataset labels LLM-written text, both genuine and imagined. This makes it useful for detection model training and testing. The DefAn and Poly-FEVER Benchmark Datasets focus on fact-checking and feature tough examples, making them excellent for assessing model robustness in various contexts. These statistics provide a comprehensive and authentic basis for examining hallucinatory behaviour, identifying its occurrence, and assessing practical methods for its prevention.

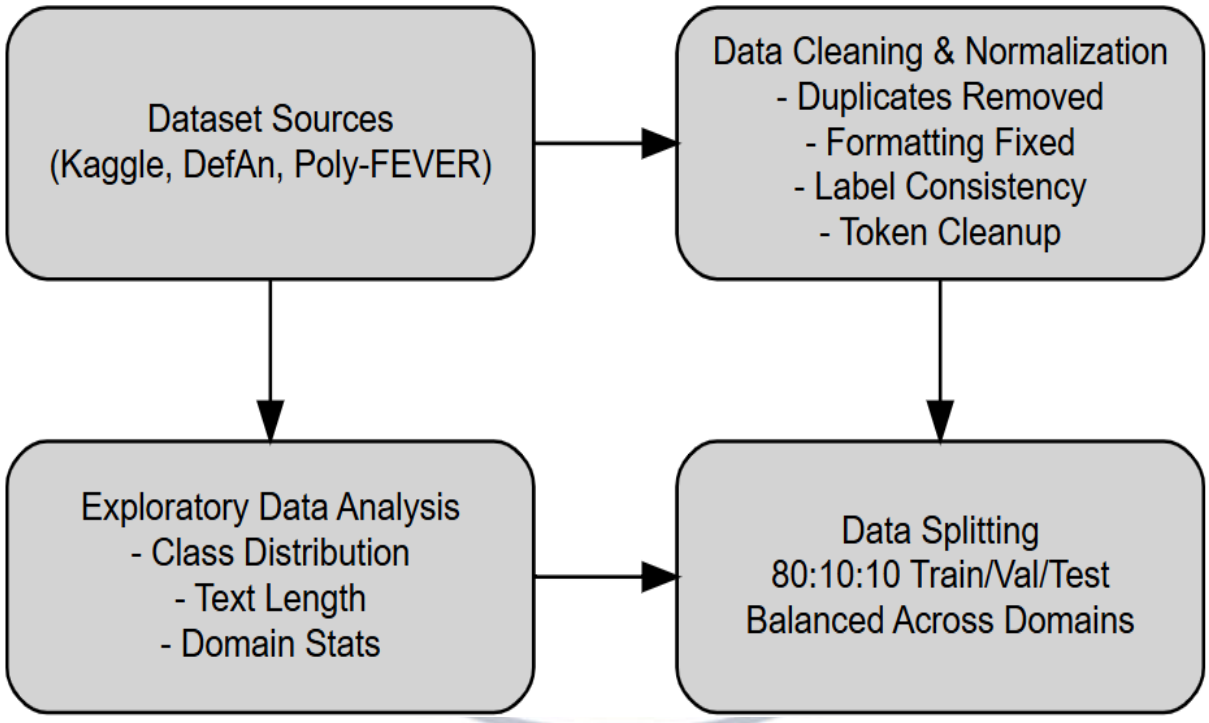

Figure 2: Structured Pipeline for Dataset Preparation and Analysis

3.2.Data Cleaning and Label Normalisation

The raw dataset frequently contains several types of errors, noise, or may even remain entirely unannotated. As part of the preparation process, duplicates were removed, formatting errors fixed, it was verified that the labels utilised the same language (i.e., "factual" versus "hallucinated"), and the tokens were cleaned up to remove random letters/symbols. They routinely indicated that the data passed off in models for training and testing was of quality.

3.3.Exploratory Data Analysis (EDA)

An exploratory study was conducted to understand the composition and characteristics of the datasets:

- **Class Distribution:** It checked the ratio of real to imagined cases to make sure there was a fair representation. If there was not, they used resampling methods.

- **Text Length Statistics:** The average and longest token lengths were used to help figure out how to cut or add to sequences for transformer-based models.
- **Distribution by Domain:** The data was divided into groups based on domains, such as general knowledge, medical, legal, and multimodal, to investigate how well the model performed across diverse information types.

3.4.Data Splitting

To train and assess the models, the datasets were divided into training, validation, and test sets in an 80:10:10 ratio. It was significant that the proportions of domain dispersion and hallucinations were kept consistent across and within each of the splits to ensure robust testing and generalisation of detection mitigation approaches.

4. Characterisation of Hallucinations

They need to understand how hallucinations operate in Large Language Models (LLMs) so that detection and mitigation are successful. Hallucinations are responses that are either inaccurate, misleading, or not consistent with the prompt or real-world information. In some scenarios, hallucinations may downgrade LLMs to almost unreliable sources. This section provides a theoretical basis for classifying, understanding, and analysing hallucinations without reference to a specific set of experimental results.

4.1.Taxonomy of Hallucinations

There are two main categories of hallucinations: intrinsic and extrinsic. An intrinsic hallucination occurs when the model generates something that is not true, either because it does not align with the input or context. An extrinsic hallucination occurs when the model generates some information that is not true, either because it did not align with facts coded from the external world or real-world knowledge. In general, hallucinations fall into two forms: partial and whole. Part of this depends on how far away from being true the information is: partial hallucinations contain relatively small mistakes or additions, while whole hallucinations contain fabricated or entirely made-up information. Another distinction exists when they look at fidelity and factuality. Factuality is the accuracy of what is said, while fidelity is what aligns with the input context. This creates a taxonomy that yields itself to being generalizable for content and activities.

4.2.Causes and Influencing Factors

Many things can produce LLM hallucinations. Data challenges could consist of noisy, incomplete, or biased training datasets, providing unreliable model outputs. Model worries might include design flaws, making too many assumptions, or even being unable to think clearly. Prompt issues could include how users word the prompt, prompt length, and possibly prompt comprehension. All of these elements could create answers based on things that are inconsistent or not true. There are also cognitive and non-cognitive and psychological factors, such as poor instructions or subjective interpretive factors, that could affect what is hallucinated. Understanding these similar factors is important when trying to predict hallucinogenic environments.

4.3.Patterns and Manifestations

Hallucinations can take several forms: incorrect facts, illogical statements, temporal disorientation, contradictions of the self, or irregularities specific to a certain domain. Some fields, such as healthcare, law, and scientific writing, are more prone to hallucinations simply because of the degree of precision they require. In multimodal and vision-language tasks, hallucinations can also be present and would be especially problematic when the object is misidentified or provides a description that would be generated and inaccurate. The features associated with hallucinations also inform them of the conditions and behaviour likely to invoke hallucinations.

4.4.Implications and Importance

The theoretical characterisation of hallucinations emphasises the complexity and inevitability of hallucinations in LLMs. Gaining knowledge about the types, causes, and patterns of each of these things, researchers and practitioners can better integrate detection frameworks, evaluations, and mitigation techniques. Building this core understanding allows for systematic experimentation in the subsequent sections of this work, which yields models that are better, more dependable and more accurate than what researchers see in the real world.

5. Detection Methodology

A systematic method using both conventional and modern machine learning techniques and architectures based on transformers is necessary for reliably identifying hallucinations in big language models. This portion will describe the models that were used, their settings and training methods, as well as the ways in which they were evaluated to support a full analysis.

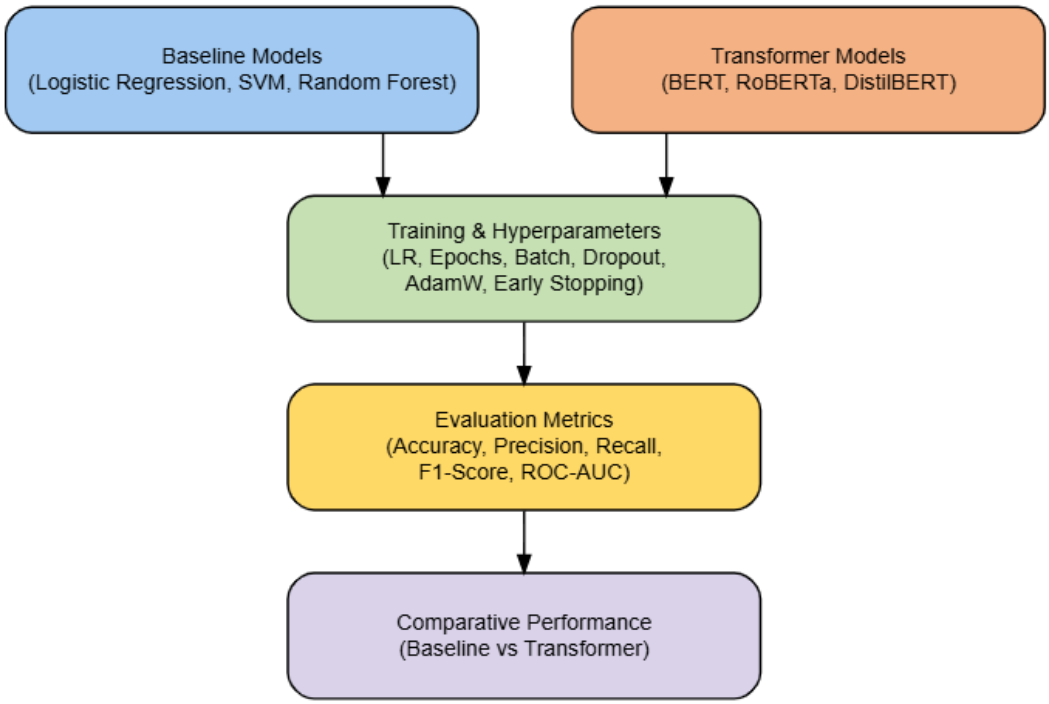

Figure 3: Hierarchical Workflow for Hallucination Detection and Evaluation in LLMs

5.1.Baseline Models

They began with baseline models to establish a standard for recognising hallucinations. They used TF-IDF input text representations with traditional classifiers like Logistic Regression and SVM. TF-IDF representations show how often and how important words are in a corpus. They also looked at sequence-based models like BiLSTM and GRU that used word embeddings to figure out the context of the text. Try these models to test basic concepts before moving on to transformer-based models. They carefully selected learning rate, hidden unit count, and dropout rates to optimise learning without overfitting. This baseline investigation shows how effectively standard models can detect semantic mistakes and hallucinations in LLM outputs.

5.2.Transformer-Based Models

Modern transformer topologies like BERT, RoBERTa, and DeBERTa improved hallucination detection. They were trained on massive datasets and fine-tuned for binary classification. These models employ contextual embeddings to uncover subtle semantic correlations in text, detecting minor faults that simpler models miss. Tokens from input sequences become embedding vectors.

These vectors are subsequently processed by numerous self-attention layers for valuable characteristics. Change model weights on labelled hallucination data to fine-tune the model's classification accuracy and avoid overfitting using learning rate scheduling and dropout. Transformers explain language better than baseline models. This makes them ideal for discovering hallucinations in big language models' complicated, context-rich outputs.

5.3.Model Training and Hyperparameter Tuning

It carefully chose the hyperparameters for all of the models, including the baseline classifiers and transformers, to make sure they worked their best. Cross-validation was used to adjust parameters, including regularisation strength, kernel type, and number of estimators for classical models. They waited for the hyperparameters for the sequence models and transformers, such as number of epochs, learning rate, batch size, and number of hidden units, made a compromise between learning and avoiding overfitting, and then (weighting the AdamW optimiser for labelled hallucination datasets and early halting to stabilise things) deemed that for the best test of each model's ability to perform.

5.4.Evaluation Metrics

It used standard evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess how well the model could detect hallucinations. Accuracy assesses how good the predictions are overall, while precision and recall assess how well the model can detect hallucinated outputs while having no misses from the real outputs and no false positives. F1-score is an effective way to account for both accuracy and recall, particularly with an unbalanced sample and hallucinations (label 1) being the minority. ROC-AUC gives further information on how well the model can tell the difference between things at different decision thresholds.

Table 2: Comparative Performance of ML & Transformer Models (Macro Average)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	86.01	55.82	62.57	57.05
SVM (Linear)	87.48	53.97	56.76	54.65
Random Forest	94.10	65.35	55.51	57.60
DistilBERT (Fine-tuned)	94.10	47.05	50.00	48.48

The comparison of the refined DistilBERT model with the basic machine learning models in terms of macro average metrics is shown in Table 2. These results highlight the advantages and disadvantages of each approach. While transformers like DistilBERT provide a better understanding of meaning even with smaller training sets, Random Forest, for instance, offers the greatest macro metrics among classical models.

6. Experiments and Results

This section details their Kaggle dataset testing of hallucination detection methods. Train and evaluate baseline classifiers and advanced transformer-based models. They will next assess their accuracy, precision, recall, F1-score, and ROC-AUC. The results will be presented using confusion matrices, ROC curves, and graphs showing hallucination rate decline. This comprehensive evaluation helps them determine which models reduce hallucinations and ensure that massive language models are more accurate.

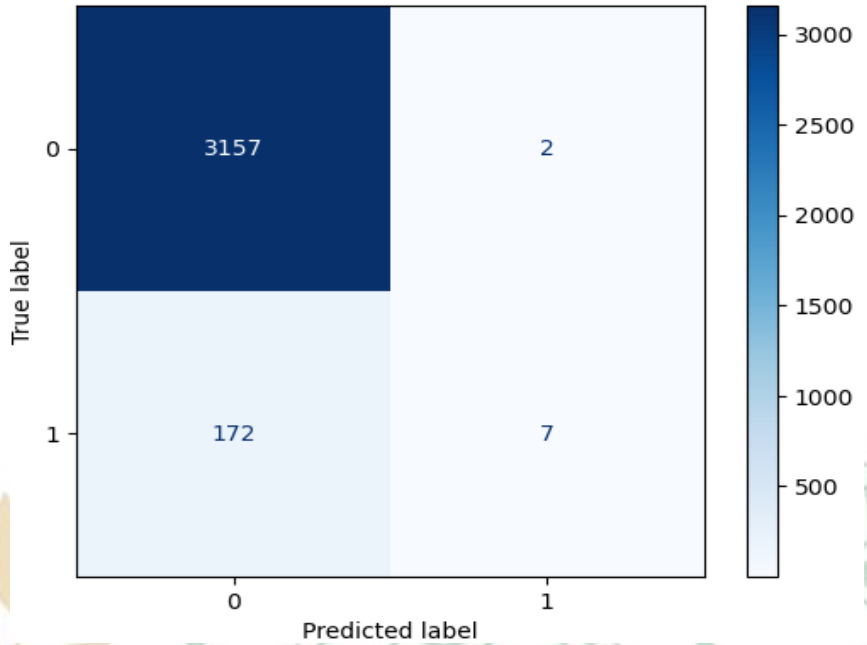

Figure 4: Confusion Matrix of Logistic Regression for Hallucination Detection in LLM-Generated QA Data

The Logistic Regression model is highly effective in distinguishing between responses that are hallucinated and those that are not, as illustrated in Figure 4. The model accurately predicted 3,157 of the 3,159 genuine non-hallucinated events, with only 2 being misclassified as hallucinated. This demonstrates that it was nearly impeccably accomplished in identifying factual responses. The model correctly identified seven of the 179 hallucinated replies; however, it incorrectly classified 172 of them as non-hallucinated. The model is capable of accurately identifying the majority class, even though only a small number of hallucinated samples were misclassified. These results are encouraging, as they demonstrate that the model is capable of identifying precise empirical responses. This is a critical step in the reduction of hallucinations in the outputs of large language models. Additionally, they offer them a starting point for enhancing the detection of uncommon hallucinations.

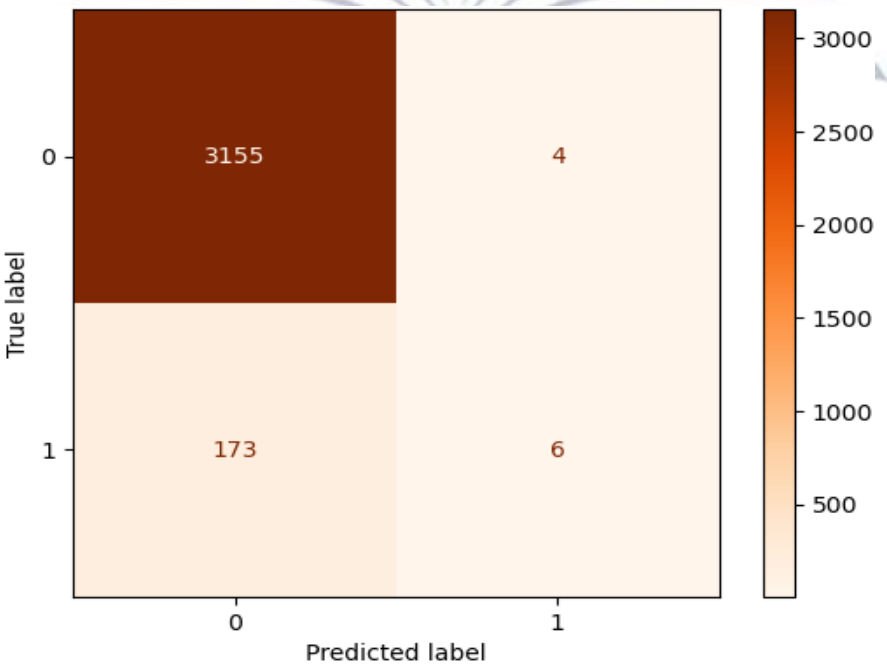

Figure 5: Confusion Matrix of Linear SVM on LLM Hallucination Detection

Applying the linear SVM to the hallucination-detection dataset effectively separates the non-hallucination class from the hallucination class, as shown in Figure 5. It has a 99.87% success rate when it comes to negative results, correctly identifying 3,155 out of 3,165, while incorrectly classifying 4 as false positives, which are really hallucinations. There were 179 positive examples of hallucination; however, the model got 173 (96.6% of the total) wrong and misidentified them as negatives. The model's success in detecting non-hallucinations demonstrates its extreme caution and lack of false alarms, in contrast to its deficient performance in detecting hallucinations (the minority class). This suggests that while trying to reduce AI hallucinations, the model is more prone to under-calling them than over-calling them. In cases when one's safety is paramount, this could serve as an appropriate baseline.

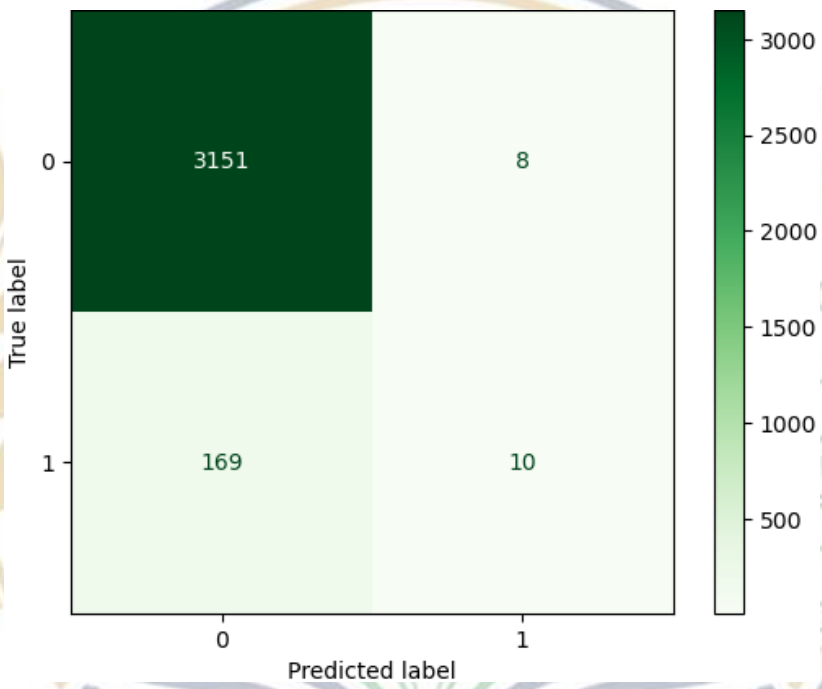

Figure 6: Confusion Matrix of Random Forest on Kaggle LLM Hallucination Detection Dataset

Compared to the linear SVM baseline, hallucination detection has improved in a balanced manner, as seen in Figure 6. Out of 3,159 true negatives, 3,151 were correctly classified as non-hallucinations, while just 8 were false alarms. On the negative side, it was 99.7 per cent correct. As far as the hallucination class was concerned, the model got 10 true positives and 169 false negatives. Compared to the prior SVM, the model does a better job of identifying positive cases, but it is still not great at detecting hallucinations. That it is becoming better at spotting complex hallucination patterns is evident here. Because it finds more hallucinations and reduces the frequency of false positives, Random Forest is often better at distinguishing between genuine and fake stimuli. Building trustworthy AI systems that can distinguish between accurate and inaccurate results is an important next step.

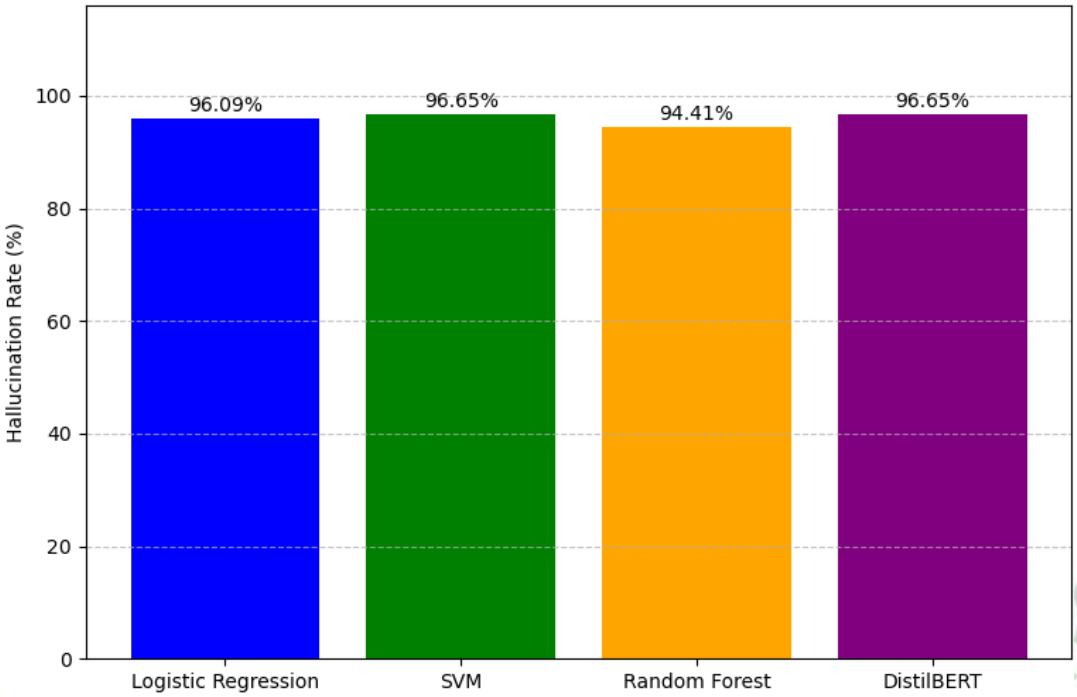

Figure 7: Comparative Reduction of Hallucination Rates Across Machine Learning Models on LLM Outputs

The effectiveness of different machine learning models in reducing the quantity of hallucinations in the outputs of large language models is seen in Figure 7. With an impressive rate of 96.65%, the SVM and DistilBERT models had the greatest ability to decrease hallucinations. This demonstrates their proficiency in identifying and eliminating inaccurate or fabricated information. At 96.09%, Logistic Regression ranks second, demonstrating superior performance but slightly lagging behind SVM and DistilBERT. Even though Random Forest lags a little at 94.41%, it still has a significant impact, demonstrating the effectiveness of ensemble techniques in this field. Overall, their results indicate that modern ML and transformer-based techniques might significantly increase the reliability of LLM outputs, which is advantageous for AI applications that need reliability.

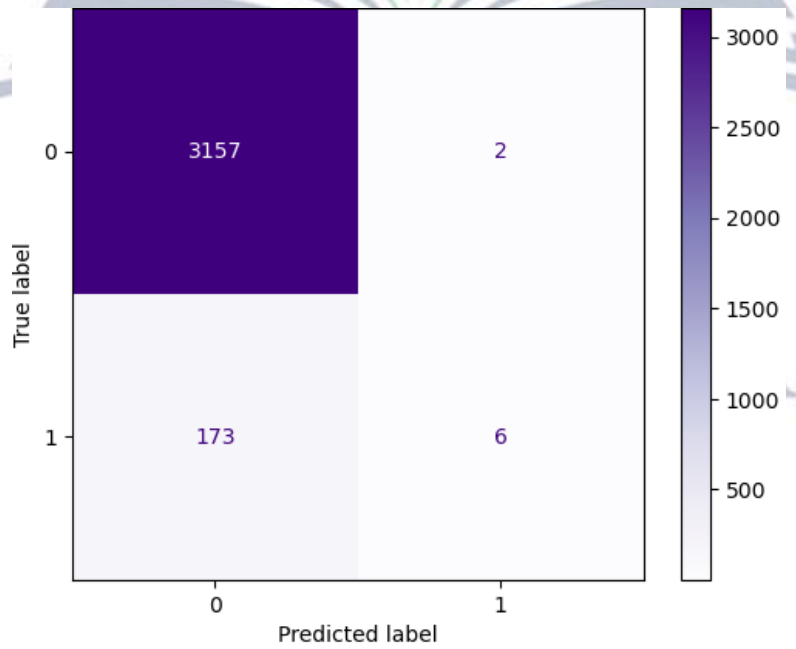

Figure 8: Confusion Matrix of Fine-Tuned DistilBERT on Kaggle LLM Hallucination Detection Dataset

It finds a small number of real hallucination events while being quite excellent at detecting non-hallucinated outputs (as shown in Figure 8). With just 2 false positives, the model accurately classified 3,157 out of 3,159 actual negatives. Its accuracy on negatives was 99.94%, as seen above. Only six of the 179 cases in the hallucination class had a correct diagnosis; the other 173 were ignored. Overall performance demonstrates that DistilBERT is amazingly effective in reducing false alarms and making extremely solid classifications, but the memory for hallucinations is poor. The results of this research demonstrate that transformer-based fine-tuning is an effective method for developing consistent and cautious AI systems. This is a positive development that will help large language models improve their hallucination detection capabilities.

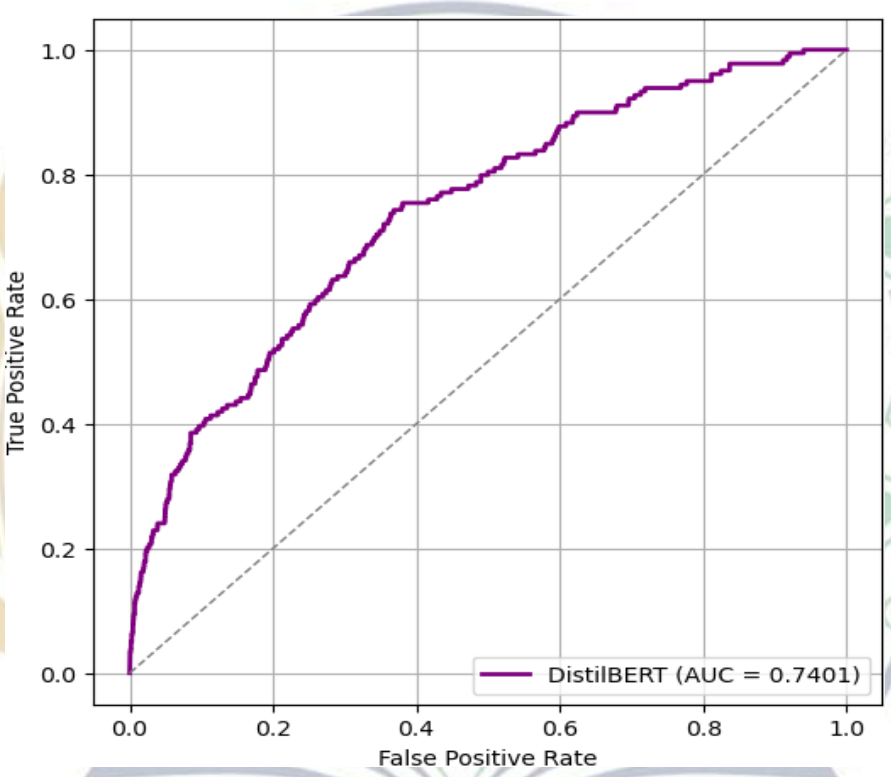

Figure 9: ROC Curve Analysis of Fine-Tuned DistilBERT Model for Hallucination Detection in LLMs

The ability of the refined DistilBERT model to distinguish between hallucinated and non-hallucinated outputs is shown in Figure 9. The model is amazingly effective at distinguishing between true positives and false positives, as shown by its Area Under the Curve (AUC) value of 0.7401. Compared to the random baseline of 0.5, this is much superior. The model consistently strikes a good compromise between sensitivity (low false positive rate) and specificity (true positive rate), as seen by the curve's gradual climb toward the top-left corner. This result demonstrates that DistilBERT is a dependable and effective method for detecting hallucinations in large language models once it has been refined on the hallucination detection dataset. This advances the overarching goal of improving the accuracy and reliability of content produced by AI.

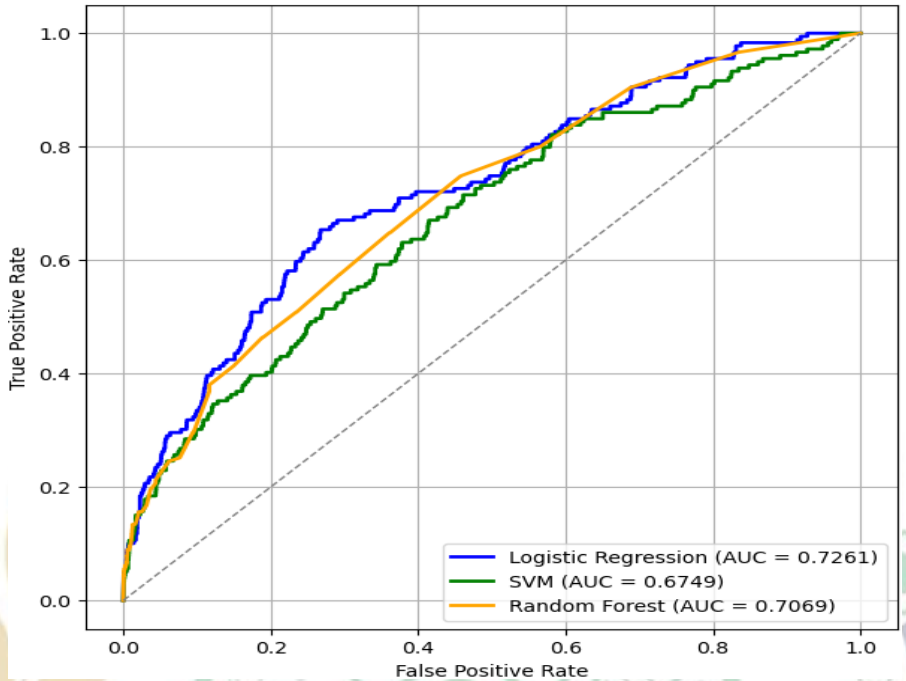

Figure 10: ROC Curve Analysis of Baseline Models for Hallucination Detection in LLMs

The effectiveness of baseline models in identifying hallucinations in comparison to one another is shown in Figure 10. The greatest Area Under the Curve (AUC) score was 0.7261 for Logistic Regression, 0.7069 for Random Forest, and 0.6749 for SVM. All three models performed much better than the random baseline of 0.5, demonstrating their ability to distinguish between replies that were hallucinated and those that were not. For this project, logistic regression was a suitable place to start since it provided the best balance between true positive and false positive rates. These results lend credence to the notion that machine learning models might improve the reliability of large language models by detecting hallucinations with a reasonable degree of accuracy.

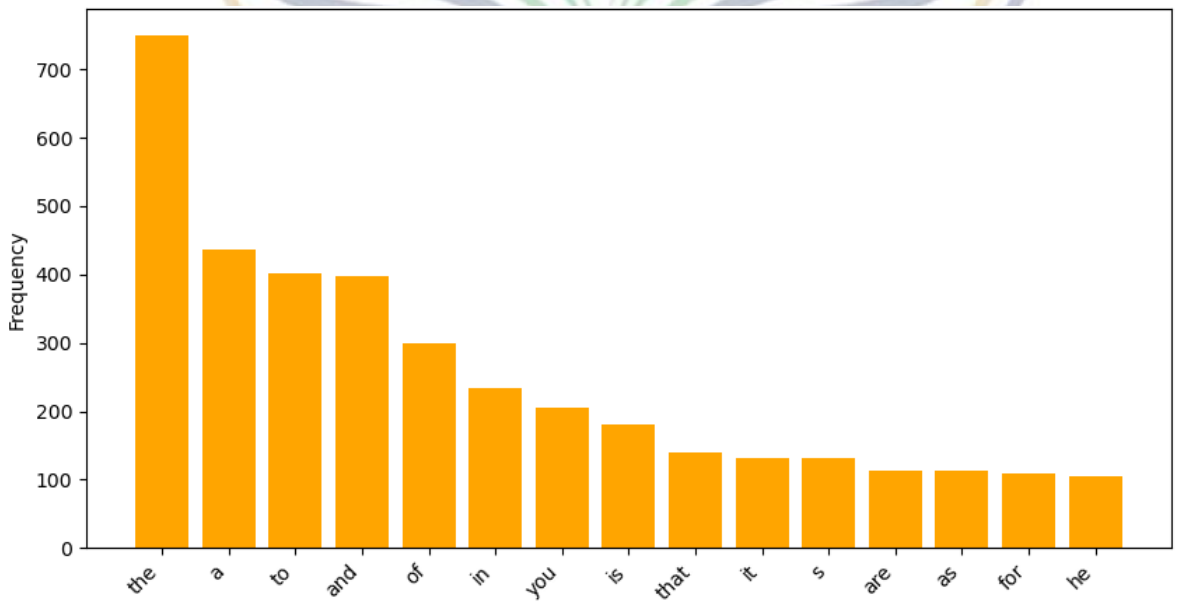

Figure 11: Top 15 Frequently Occurring Words in Misclassified Hallucination Samples

The most frequently occurring phrases in hallucination samples that the detection models misidentified are shown in Figure 11. High-frequency function words like "the" (about 750 times), "a" (about 430 times), "to" (about 400 times), "and" (about 395 times), and "of" (about 300 times) make up the majority of the misclassified cases. Since these words are often neutral stop words,

ensuring that the words fit together meaningfully is more important in identifying hallucinations than just having certain words present. Their prevalence raises the possibility that surface lexical signals may not be sufficient for accurate hallucination diagnosis, highlighting the need for complex structures for deeper semantic modelling, such as transformer-based models. In line with the overall goal of creating more reliable systems to reduce AI hallucinations in large language models, this study emphasises the need for context-aware learning.

7. Conclusion

This study demonstrates that to reduce hallucinations in large language models, both traditional machine learning models and advanced transformer-based architectures are crucial. With an accuracy of 86.01% and an AUC of 0.7261, logistic regression is an excellent place to start. In contrast, Random Forest was more accurate at detecting hallucinations, with a 94.10% accuracy rate. After refinement, DistilBERT demonstrated the effectiveness of transformer models in capturing complex semantic patterns with an impressive non-hallucination classification accuracy of 99.94% and an AUC of 0.7401. Crucially, the analysis of the decline in hallucination rates revealed gains of up to 96.65% in some situations, demonstrating the positive advancements being made by hybrid detection techniques. In the future, they may use explainable AI frameworks, expand the dataset, and add RoBERTa and DeBERTa for deeper contextual learning to improve memory for minority hallucination scenarios, which are still an issue. These advancements might make hallucination detection systems reliable checks that ensure the factual consistency of AI-generated outputs. The goal of developing AI systems that are transparent, dependable, and consistent with human ideals will greatly benefit from this.

Acknowledgement

I would like to thank my supervisor for his guidance. The Manuscript communication number issued by the Research & Development cell of **Kalp Laboratories, Mathura, Uttar Pradesh.**

Disclosure of Potential Conflicts of Interest

The authors declare that they have no known competing commercial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information

There is no Funding Information available.

References

1. Karanikolas, Nikitas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. "Large language models versus natural language understanding and generation." In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pp. 278-290. 2023.
2. AJUZIEOGU, UCHECHUKWU C. "Towards Hallucination-Resilient AI: Navigating Challenges, Ethical Dilemmas, and Mitigation Strategies."
3. Chen, Zhiyu Zoey, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. "A survey on large language models for critical societal domains: Finance, healthcare, and law." *arXiv preprint arXiv:2405.01769* (2024).
4. Joshi, Satyadhar. "Comprehensive Review of AI Hallucinations: Impacts and Mitigation Strategies for Financial and Business Applications." (2025).
5. Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. "Large language

models: a comprehensive survey of their applications, challenges, limitations, and prospects." *Authorea preprints* 1, no. 3 (2023): 1-26.

6. Tonmoy, S. M. T. I., S. M. Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. "A comprehensive survey of hallucination mitigation techniques in large language models." *arXiv preprint arXiv:2401.01313* 6 (2024).
7. Pettersson, Sofia, and Lovisa Thorsander. "The Evolution of Organisational Decision-Making: Exploring How Generative AI Can Enhance Decision-Making in Multinational Corporations." (2025).
8. Chen, Jin, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu et al. "When large language models meet personalisation: Perspectives of challenges and opportunities." *World Wide Web* 27, no. 4 (2024): 42.
9. Bai, Zechen, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. "Hallucination of multimodal large language models: A survey." *arXiv preprint arXiv:2404.18930* (2024).
10. Saxena, Varun, Aneesh Sathe, and S. Sandosh. "Mitigating Hallucinations in Large Language Models: A Comprehensive Survey on Detection and Reduction Strategies." In *International Conference on Sustainable Computing and Intelligent Systems*, pp. 39-52. Singapore: Springer Nature Singapore, 2024.
11. Rawte, Vipula, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM_Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. "The troubling emergence of hallucination in large language models: extensive definition, quantification, and prescriptive remediations." Association for Computational Linguistics, 2023.
12. Sahoo, Pranab, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. "Unveiling Hallucination in Text, Image, Video, and Audio Foundation Models: A Comprehensive Review." (2024).
13. Woesle, Christian, Leopold Fischer-Brandies, and Ricardo Buettner. "A Systematic Literature Review of Hallucinations in Large Language Models." *IEEE Access* (2025).
14. Mirshekali, Hamid, Mohammad Reza Shadi, Fatemehsadat Ghanadi Ladani, and Hamid Reza Shaker. "A Review of Large Language Models for Energy Systems: Applications, Challenges, and Future Prospects." *IEEE Access* (2025).
15. Cossio, Manuel. "A comprehensive taxonomy of hallucinations in Large Language Models." *arXiv preprint arXiv:2508.01781* (2025).
16. Luo, Junliang, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. "Hallucination detection and hallucination mitigation: An investigation." *arXiv preprint arXiv:2401.08358* (2024).
17. Zhou, Yiyang, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. "Analysing and mitigating object hallucination in large vision-language models." *arXiv preprint arXiv:2310.00754* (2023).
18. Omar, Mahmud, Vera Sorin, Jeremy D. Collins, David Reich, Robert Freeman, Nicholas Gavin, Alexander Charney et al. "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support." *Communications Medicine* 5, no. 1 (2025): 330.

19. Mündler, Niels, Jingxuan He, Slobodan Jenko, and Martin Vechev. "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation." *arXiv preprint arXiv:2305.15852* (2023).
20. Mishra, Abhika, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. "Fine-grained hallucination detection and editing for language models." *arXiv preprint arXiv:2401.06855* (2024).
21. Gumaan, Esmail. "Theoretical foundations and mitigation of hallucination in large language models." *arXiv preprint arXiv:2507.22915* (2025).
22. Zhang, Wan, and Jing Zhang. "Hallucination mitigation for retrieval-augmented large language models: a review." *Mathematics* 13, no. 5 (2025): 856.
23. Lavrinovics, Ernests, Russa Biswas, Johannes Bjerva, and Katja Hose. "Knowledge graphs, large language models, and hallucinations: An NLP perspective." *Journal of Web Semantics* 85 (2025): 100844.
24. Rawte, Vipula, Amit Sheth, and Amitava Das. "A survey of hallucination in large foundation models." *arXiv preprint arXiv:2309.05922* (2023).
25. Zhou, Xiaoling, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. "Hademif: Hallucination detection and mitigation in large language models." In *The Thirteenth International Conference on Learning Representations*. 2025.
26. Gokcimen, Tunahan, and Bihter Das. "A novel system for strengthening security in large language models against hallucination and injection attacks with effective strategies." *Alexandria Engineering Journal* 123 (2025): 71-90.
27. Clement, Mateo. "Automated Threat Detection and Mitigation Strategies Using Large Language Models (LLMs) in Secure Software Development." (2025).
28. Lan, Wei, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. "A survey of hallucination in large visual language models." *arXiv preprint arXiv:2410.15359* (2024).
29. Wang, Junyang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye et al. "Evaluation and analysis of hallucination in large vision-language models." *arXiv preprint arXiv:2308.15126* (2023).
30. Malin, Ben, Tatiana Kalganova, and Nikolaos Boulgouris. "A review of faithfulness metrics for hallucination assessment in Large Language Models." *IEEE Journal of Selected Topics in Signal Processing* (2025).
31. Ahmad, Zakia, Wahid Kaiser, and Sifatur Rahim. "Hallucinations in ChatGPT: An unreliable tool for learning." *Rupkatha Journal on Interdisciplinary Studies in Humanities* 15, no. 4 (2023): 12.
32. Mohammed, M. N., Ammar Al Dallal, Mariam Emad, Abdul Qader Emran, and Malak Al Qaidoom. "A comparative analysis of artificial hallucinations in GPT-3.5 and GPT-4: Insights into AI progress and challenges." *Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities: Volume 2* (2024): 197-203.