Detecting misleading information from Large Language Models responses

by

Muntasir Ahmed Ador
20101259
Fahim Hasan
20201144
Syed Ashik Mahamud
20301124
Iffat Ara Nazmin
21101106
Md. Mahim Muntasir Arin
24141109

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

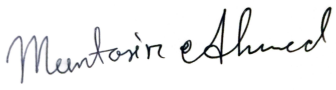
Department of Computer Science and Engineering
Brac University
February 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing the degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Muntasir Ahmed Ador
20101259

Fahim Hasan
20201144

Syed Ashik Mahamud
20301124

Iffat Ara Nazmin
21101106

Md. Mahim Muntasir Arin
24141109

Approval

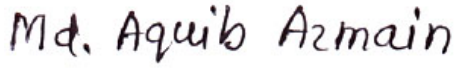
The thesis titled "Detecting misleading information from Large Language Models responses" submitted by

1. Muntasir Ahmed Ador (20101259)
2. Fahim Hasan (20201144)
3. Syed Ashik Mahamud (20301124)
4. Iffat Ara Nazmin (21101106)
5. Md. Mahim Muntasir Arin (24141109)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on February, 2025.

Examining Committee:

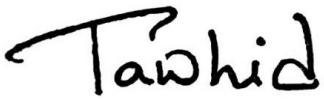
Supervisor:
(Member)



Md. Aquib Azmain

Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Md. Tawhid Anwar

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Table of Contents

Declaration i

Approval ii

Table of Contents iv

Nomenclature v

Abstract 1

1 Introduction 2
- 1.1 Background . . . 2
- 1.2 Reseacrch Problem . . . 3
  - 1.2.1 Large Language Models . . . 3
  - 1.2.2 What is Hallucination? . . . 4
  - 1.2.3 Cause of Hallucination . . . 4
  - 1.2.4 Classification of Hallucination . . . 5
  - 1.2.5 Research Objectives . . . 6
  - 1.2.6 Related Works . . . 6

2 Dataset and Processing 11
- 2.1 Dataset . . . 11
- 2.2 Dataset (Bhrantio Satya) Analysis . . . 11
- 2.3 Scoring Rubric . . . 14
  - 2.3.1 Annotation Class . . . 14
- 2.4 Data Preprocessing . . . 17
- 2.5 Dataset Report . . . 19

3 Methodology 22
- 3.1 Model Selection and Training . . . 22
- 3.2 Model Optimization . . . 23

4 Results And Discussion 25
- 4.1 Experimental Analysis . . . 25
- 4.2 Model Performance Overview . . . 26
- 4.3 Fine-Tuned DistilBERT Performance . . . 26
- 4.4 Discussion . . . 28

5 Conclusion 30

Bibliography 34

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

LLM =Large language Model

NLG =Natural Language Generation

HADES =HAllucination detection dataSet

RLKF = Real-Time Learning and Knowledge Fusion

NLP = Natural Language Processing

CNN = Convolutional Neural Network

RELQA = Relevant Question Answering

REID = Re-Identification

PLMs = Pre-trained language models

AI = Artificial Intelligence

Abstract

The arrival of large language models (LLMs) have been a game-changer in natural language processing (NLP). It revolutionized the way we comprehend and generate content.However, LLMs can hallucinate-that is, contradict the reality or input provided by the user. This is a big problem because these models are being used these days in many diverse industries, including for medical and legal purposes where accuracy is paramount. Hallucinations can damage user trust and lead to the spread of incorrect facts. Although it is not a major issue in ordinary situations, it raises serious concerns in sensitive areas like healthcare and legal advice. Also it can inadvertently become part of the training corpus for future models if not carefully filtered. This creates a feedback loop where errors in one generation of models can propagate and potentially amplify in subsequent iterations. To solve this problem, we have created an all-rounded dataset with questions from SQuAD (Stanford Question Answering Dataset), HotpotQA, and TriviaQA, among other datasets. We will use the state-of-the-art LLM GPT-4o mini to generate answers. Finally, to this end, the paper describes several rules for the annotation of a corresponding dataset, its resulting characteristic properties and the classification quality that can be achieved when using the dataset for fine-tuning different models.

Keywords: Machine Learning, Natural Language Processing, Large Language Models, AI hallucination, GPT, Transformer, ALBERT, T5

Chapter 1

Introduction

1.1 Background

Large Language Models (LLMs) like GPT-3, GPT-3.5, GPT-4, GPT-4o, Gemini, Phi-3 etc. have been serving a great purpose in human life. These are generally developed to understand human command from the input text given. These models have been trained by using a large amount of training data. One can do work like text generation, summarizing large or small paragraphs or answering MCQ type questions with short or long answers and so on. It has a substantial contribution in the field of NLP and AI. This model has the coherent power of generating texts, ideas, and other necessary AI based answers which are playing a vast role in human life.

However, in the process of all the amazing things it can do, LLM sometimes faces problems like Bias, Hallucinations or Limited length of contents etc. In this paper we will be focusing on Hallucination in LLM. Hallucinations are patterns of models that can produce coherent concepts, albeit completely fictitious. This issue has raised growing safety and ethics concerns since LLMs are being used more broadly and commonly. Prior to LLM, the term "hallucination" had a more specific meaning in the NLP context: producing illogical or untrustworthy results out of an input.[28].

Within the field of NLP, the term "hallucination" describes how a system will produce content illogical or not trustworthy from an available source. [14]. User inputs often vary between tasks that require summarization and those involving other forms of text generation.In NLG activities like machine translation, hallucinations are generated when information produced is contradictory to the input of the task[5] and summarization [11]. Hallucination often relates to faithfulness, focusing on factual accuracy rather than grammatical correctness or fluency. Factual errors, newly invented topics, or logical errors can indicate AI hallucinations in LLMs. Misinterpretations may occur, leading to incorrect answers, dates, or times, often due to model overfitting or lack of relevant external data.

Generally, in LLM user inputs vary between two types such as users asking for summarization and input for a task to summarize. We often refer to hallucination as faithfulness, mainly based on factual information or accuracy, not grammatical correctness, fluency or authenticity. Factual errors, totally newly invented topics, or logical errors can be detected as AI Hallucinations in LLM. Sometimes AI can

misinterpret situations like giving wrong answers of a situation, giving wrong dates or times about a legend for hallucination. This generally happens because of model overfitting or lack of external related data. We will use Fine-Tuning in this paper to mitigate hallucination, but other approaches can be keeping a human in the process, introducing the integration of external data, or improving the dataset used for training.

The research in LLM hallucinations is an open topic in the field of medical, financial, and other research-related fields due to the accuracy rates which are very low from input to output. As, Hallucination, a logical or factual error of LLM, in the real world is still an open problem to all people, it's very difficult to give a real formal definition to it [35][9]. Hallucination happens whenever any model fails to give a result as predicted, be it due to its design, learning algorithms, prompting tactics, or training data. The bridge between the formal and real world leads to LLMs performing well in the latter. The research in LLM hallucinations is an open topic in the field of medical, financial, and other research-related fields due to the accuracy rates which are very low from input to output. As, Hallucination, a logical or factual error of LLM, in the real world is still an open problem to all people, it's very difficult to give a real formal definition to it [35][8]. Hallucination mainly occurs when the model fails to give the output as expected regardless of model architecture, learning algorithms, prompting techniques, or training data.

1.2 Reseacrch Problem

In this era of technology, the emergence of large language models (LLMs) has been a breakthrough in the field of artificial intelligence especially in Natural language programming (NLP). This large language model with their vast knowledge and capability of generating contextually relevant text has left a great impact not only in research or study but also in industry and workplace. However, one critical problem has arisen in this case which is hallucination. It is when LLM models generate factually inaccurate information. For this reason, in some cases where information is very sensitive, it cannot be used.

1.2.1 Large Language Models

An LLM is a type of AI that is trained on large volumes of data to comprehend and generate text. Such models as GPT-3 [10] and PaLM [13] have shown the best performance in terms of generating a response to diverse user requests with human-like characteristics. Autoregressive models are employed in applications like report automation, virtual assistants, and summarization systems. They have been used in many applications such as automation tools to draft reports, virtual assistants and sum-marization systems by using autoregressive models. Autoregressive models are an important category of Large language models [7] [13] [26].Considering Transforms of autoregressive model is the backbone of its architecture which contains encoder and decoder that helps the model to understand and generate text like a human and predict next token from previous token [4] . Adaptation of transformers and widespread use of autoregressive models for generative AI were build on foundation of n-grams [1] [3] and recurrent neural network [2] .Transformer-based LLMs have shown so much exceptional performance among tasks that NLP has shifted from a

paradigm centered on task-specific solutions to general-purpose pretraining [6] [7] .This widespread use of transformer base LLM brings a new horizon in the neural language generation task like summarizing dialogue generating paraphrasing and question and answering. Understanding of a language and context from sentence , reasoning that sets LLM apart from other existing models.

1.2.2 What is Hallucination?

In recent years large language models have improved their architecture and performance. Despite its performance, it is fully immune from mistakes in practical application. Large language models encounter problems hallucination is the most challenging one. The term hallucination is well known in the Natural language programming community; after the emergence of LLM, they adopted the term. The term hallucination refers to a generated output by a large language model that is factually incorrect or non-existent in real life. Hallucinations have been observed in various text generation activities, including discussion and summarization, and other natural language generating tasks[31]. For this LLMs lost credibility as one needs to verify if the LLM is hallucinating or not[32]. Hallucinating sometimes won't be much of a big deal for most cases but where the credibility of every information is important LLM cannot be implied with supervision. There are many hallucination detection schemes to detect for any prompt the LLM is hallucinating or not. In This paper, we will talk about some of the detection schemes and how this problem can be mitigated with our outlook and solution of this matter[33].

1.2.3 Cause of Hallucination

As a powerful artificial intelligence system Large language models can understand and do dialogue like humans. While LLM-generated writings are so realistic and compelling, a growing concern is the tendency of LLMs to hallucinate facts. It is common knowledge that models can confidently output utterances that are not true[34]. It has been especially difficult to formally define what a real-world LLM hallucination is - that is, a factually or logically incorrect mistake. This is very much an open problem as a lot of work and research is happening to understand the causes of hallucinations and ways to mitigate this problem[21]. From this study, the researcher found a few reasons why hallucinations can occur. Biased data, small datasets, model architecture, model tune issues, and other things are identified as reasons. They are:

- Model Architecture: Transformers of LLM contain encoder and decoder which help them to predict the next token from the previous token can be underfitted or be overfitted which will cause the LLM to generate a fictional answer [21].

- Model Fine tune issue: In most cases a pre-trained model which can predict next word from a previous text is used as base for a large language model that is targeted to NLG tasks. This kind of pre-trained model is used as a base because this kind of model has already built neurons that understand language so it has a good accuracy. But if the used pretrained model is not tuned perfectly it can be a cause of hallucination [40].

- Lack of Context: If the given prompt doesn't have clear context LLM can

struggle to understand what its user wants and will generate information that is not relevant to the given prompt[19].

- Bias Data and Small Dataset: LLMs are trained on large datasets. If the data is biased in some way, the trained model will be prone to hallucination and may generate biased answers [40]. If the text dataset used to train the LLM is not sufficiently large or lacks diverse examples, the model may struggle to learn or understand patterns, leading to hallucinations due to a lack of data [30].

1.2.4 Classification of Hallucination

Hallucinations found in large language models are primarily classified either phenomenally or mechanically. Also, hallucinations can be classified depending on generated outcomes or even training and deployment methodology. Here we will discuss a few of them:

- Factual hallucination: In this case of hallucination, when Large language model generates output that is Irrelevant frictional or factually incorrect. It occurs mostly for biased dataset or if LLM is trained with a small dataset.

- Contextual Hallucination: Large language model and even fine-tune Large language model GPT which has a pre-trained transformer given show contextual hallucination. When a large language model like gpt, bard, co-pilot gives a response to a prompt which is completely out of context. An example can be if the given prompt was "There is London" and the response was "Neil Armstrong was the first man to on moon landing" which was totally out of context. Also, incomplete response or creating alternate context is also considered as contextual hallucination.

- Logical Hallucination: Logical hallucination is when an LLM model jumps to wrong or illogical conclusion. It can if an LLM model got a conclusion which is wrong or as in some cases it can happen is, through the response seems logical but was based on incorrect information which is also a form of logical hallucination. An example can be an LLM model that said ice cream sales are the cause of sunburn as both increase in summer which is a logical hallucination as ice cream sales and sunburns are not relatable.

- Incoherent Hallucination: If the response of the LLM model is confusing, lack clarity or logical structure is considered as Incoherent hallucination. Sometimes response of a given prompt contains unnecessary information which hinders to find the real info one is looking for is also considered as incoherent hallucination.

- Grammatical Hallucination: If a response that contain grammatical error, punctuation error or the error that violate the rules of grammar is told a grammatical hallucination. It made contain incorrect verb tense, subject-verb agreement issues, or misplaced modifiers or just hard to read and understand.

1.2.5 Research Objectives

The following work pursues the formulation of a method for misleading information detection. in the output produced by large language models. They are proficient in a wide range of tasks related to language but, in most cases, tend to present hallucinations—content that deviates from factual correctness or source input. The aims of this research include:

- To come up with a comprehensive dataset through the incorporation of questions from existing datasets, which are SQuAD, DuReader, HotpotQA, and others.
- Utilize GPT-4o Mini to generate responses for all questions within the dataset.
- Development of evaluation criteria and metrics including human-machine methods to systematically identify the instances of hallucination in the generated responses.
- Train multiple transformer models on the merged dataset and then fine-tune the best-performing with the objective of enhancing the capability to detect hallucinations.
- Fine-tune the trained transformer model for achieving optimum performance in the detection and reduction of hallucinations.
- Conduct an in-depth analysis of the kinds of hallucination present in the LLM-generated content.
- Provide useful insights from the analysis that can be applied in potential strategies for effectively detecting hallucinations from LLM-generated content. We will annotate and preprocess the responses, then develop and assess transformer models for hallucination detection.

1.2.6 Related Works

Large Language Models (LLMs), while capable of generating coherent text, often produce hallucinations—factually inaccurate or fabricated content—due to their reliance on probabilistic predictions rather than understanding. This phenomenon poses challenges for generative AI applications, prompting extensive research on detection, mitigation, and evaluation techniques. This section reviews prior and relevant work in the field of hallucination in LLMs, encompassing a variety of techniques, approaches, and collaborative efforts between models and datasets to address this issue.

HaluEval is a large benchmark dataset for evaluating the ability of Large language models to detect hallucinations which contains 35,000 samples where 5,000 is done through human annotations in relevance of ChatGPT response and the rest is automated[22]. A two-stage "Sampling-then-filtering " approach is proposed using ChatGPT where many hallucinated samples are produced and then filtered to select the most plausible and challenging queries .This experiment shows ChatGPT to be 62.59% accurate on QA. This benchmark hopes to evaluate which type of content

makes the LLMs hallucinate. This can also be used to develop more reliable LLMs in future studies.

It is observed that exploring the artifacts associated with generation of answers do provide hints that the generation by the LLM will contain hallucination or not [25]. The LLM is observed through analyzing inputs, outputs. By building on this insight binary tree classifiers are trained that use these artifacts for detecting hallucinations. These classifiers get up to 0.80 AUROC.

The study of factuality hallucination in [36] implicates the tendency of Large Language Models to generate content that is factually incorrect but can not be detected through human interaction. This study presents a new hallucination benchmark HaluEval 2.0 and designs a simple but effective detection method to detect hallucination in a Large language Model. This new benchmark probes hallucination through many stages of LLMs , like pre-training, SFT, RLHF and inference. By using a variety of techniques like RLHF, retrieval augmentation, prompt improvement this study tries to mitigate factual hallucination.

A fact -checking system called Truth-O-Meter identifies and corrects inaccurate information generated by Large Language Models through text matching algorithms [20]. The system compares the LLM-generated text through search engines, web portals and other sources of valid information by using text mining and web mining techniques we can correctly identify the hallucinated answers or predictions.

Model introspection has been used to diagnose issues in large language models with many hypotheses proposed [28]. Low Source Contribution Hypothesis occurs when models generally rely more on the target context than the source context during translation. Tested with pseudo-hallucinations using random target prefixes. Static Source Contribution Hypothesis prefers that hallucinations are linked to a model's static attention to certain source tokens throughout inference. To mitigate these they used Layerwise Relevance Propagation (LRP) to identify hallucinations by analyzing contributions of different source tokens during inference. Also they used Cross-Lingual Embeddings and Classifiers which uses tools like LASER and XLM-R, especially when combined with LRP in detection of hallucination with high precision.

A study proposed a method for detecting and assessing the facility of hallucinated entities by using language models [12]. By probing these entities prior and posterior probabilities the method can distinguish between hallucinated and non hallucinated samples where it is evaluated on the XSUM dataset and compared against five baseline models.The results indicate that this method significantly outperformed most studies. The study employs a K-Nearest Neighbors (KNN) classifier trained on these probabilistic features.

A review suggests that Hallucination poses a significant challenge for these models and has been predominantly examined within the realm of large language models (LLMs), as evidenced by surveys such as those conducted by Ji et al. (2023) and Zhang et al. [24]. This paper seeks to present the first comprehensive survey of hallucination phenomena across all major modalities of foundation models.

RelD is a robust discriminator designed to effectively identify hallucinations in responses generated by large language models (LLMs) [17]. RelD is trained on RelQA, a bilingual question-answering dialogue dataset, and utilizes a comprehensive set of metrics. RelD accurately detects hallucinations in outputs from various LLMs and discerns between hallucinated answers from both in-distribution and out-of-distribution datasets. Moreover, the RelQA dataset, comprising nine diverse sub-datasets encompassing extractive reading comprehension, multiple-choice questions, and multi-turn dialogues, was utilized. To ensure compatibility, we formatted and integrated these datasets. Texts were segmented into 4,000-token windows for improved clarity and effectiveness. Several LLMs, including LLaMA, BLOOM, GPT-J, GPT-3, and GPT-3.5, were employed to generate answers, which were then evaluated using a combination of LLM-assessment, human, machine, and composite metrics. The RelD discriminator was trained using a regression approach, but it exhibited poor performance. Subsequently, we transitioned to a classification task, normalizing final scores into multiple classes. ELECTRA was selected as the backbone for RelD, surpassing other pre-trained language models (PLMs) such as BERT and RoBERTa.

Hallucinations in large language models (LLMs) occur when they generate inaccurate or fabricated information. To detect these hallucinations efficiently, a paper [16] developed AutoHall, an automated method using existing fact-checking datasets. This approach bypasses the need for manual annotation.This paper introduces AutoHall, a novel automated approach for creating hallucination detection datasets using pre-existing fact-checking datasets. Their experimental results demonstrate the effectiveness of this approach in detecting hallucinations in models such as ChatGPT and Llama-2, with estimated hallucination rates ranging from 20 percent to 30 percent.

From [18] The Wizard of Wikipedia dataset, while abstractive summarization relies solely on reference documents from the CNN/DM dataset. Token-based metrics such as BLEU-4, Rouge-L, ChrF, and METEOR measure text generation quality, while knowledge-based metrics like Knowledge-F1 (KF1) and K-Copy evaluate knowledge overlap and copying. Multi-faceted metrics consider aspects like naturalness, coherence, groundedness, and relevance. Popular large language models (LLMs) like ChatGPT, GPT-3.5, Flan-T5 (FT5), and T0++ are used, along with decoding baselines such as FUDGE, NADO, and MCTS. The methodology encompasses task selection, evaluation metric categorization, baseline model selection, implementation details, and main evaluation procedures, providing a comprehensive framework for assessing method effectiveness in knowledge-grounded natural language generation.

A method was proposed for token level reference-free hallucination detection process that can detect hallucinations at a token level allowing for a more granular and precise analysis of the generated text, enabling the model to identify and address specific errors or inaccuracies in the output for every prompt [15]. To replicate as hallucinated text produced by NLG, they first develop a HADES dataset by extracting a huge volume of text from Wikipedia. he model-in-the-loop annotation strategy, along with the use of feature-based and transformer-based models, effectively addresses label imbalance and improves the detection of inconsistencies in generated text and save a lot of manual labor.

Another mitigation method talks about "SelfCheckGPT" a simple sampling-based approach where they leverage the stochastic sample fact check responses [23]. If the sample responses diverge it is inaccurate and if they don't they are accurate.

From [29] we get a detailed review of various types of hallucinations categorized meticulously. The paper involves a comprehensive review and theoretical analysis of previous and current work. They analyze the mechanism behind these anomalies by dividing them in three perspectives: line data collection, knowledge gaps and optimization process.The paper also reviews many techniques and methods for detection and mitigation for hallucination.

An approach to both detect and mitigate hallucinations in large language models is to take a series of baseline methods based on pretrained transformer models such as BERT, GPT-2, XLNet and RoBERTa [38]. The study introduces two settings for detection: online and offline. The effectiveness of these models is to evaluate metrics like accuracy, precision, recall and F1 score. The study also explores the previously explained HADES dataset to mitigate hallucination.

The Med-HALT benchmark and dataset is used to evaluate and mitigate hallucinations in LLMs within the medical domain [27]. They introduced Med-HALT a benchmark and dataset that includes reasoning based and memory based hallucination tests. Several state-of-the-art LLMs were evaluated, including Text Davinci, GPT-3.5, LlaMa-2, MPT, and Falcon. Performance was measured using particular measures, including accuracy and pointwise scores.

In order to lessen factual hallucinations, the study [37] examines how well large language models (LLMs) are able to identify and communicate their internal knowledgestates . Furthermore, although the RLKF framework has potential, the quality of the factual preference data utilized for training has a significant impact on how effective the framework is. The study also points out that only a small number of models—such as Llama2-chat and Gwen-chat—are used to analyze the suggested approaches.

The prevalent problem of hallucinations in large language models (LLMs), which produce factually false assertions, is addressed in the study [39]. The paper develops a thorough taxonomy with six hierarchically defined forms of hallucinations and presents a novel challenge of fine-grained hallucination detection. The analysis

shows that a considerable percentage of the outputs from ChatGPT and Llama2-Chat show hallucinations. The research suggests FAVA, a retrieval-augmented LLM intended to identify and rectify fine-grained hallucinations, as a solution to these hallucinations. Based on experimental results, FAVA greatly improves the factuality of LM-generated text by 5-10 % in FActScore, outperforming ChatGPT in fine-grained hallucination identification.

In conclusion, addressing hallucinations in Large Language Models (LLMs) remains a critical challenge, as these models often generate factually inaccurate or fabricated content. Recent advancements have introduced innovative benchmarks like HaluEval and HaluEval 2.0, alongside automated detection methods such as SelfCheckGPT and AutoHall, to identify and mitigate hallucinations. Techniques like Layerwise Relevance Propagation (LRP), Cross-Lingual Embeddings, and retrieval-augmented generation (RAG) have shown promise in diagnosing and reducing hallucinations. Domain-specific datasets like Med-HALT and RelQA further enable targeted evaluation and improvement. However, challenges such as label imbalance, reliance on high-quality training data, and the need for granular detection persist. Enhancing the correctness and dependability of LLMs requires ongoing research and improvement, especially in delicate fields where factual accuracy is crucial.

Chapter 2

Dataset and Processing

2.1 Dataset

The data set consists of questions and answers where questions were created for the purpose to make a LLM hallucinate. Furthermore, The answers were generated by ChatGPT-4o mini and annotated by my team through a rubric which we created by sampling hallucinated data.(Bhrantio Satya) . The dataset (Bhrantio Satya) was built with the sole purpose of finding irregularities in an LLM so that we can detect misleading information by studying the answers generated by our questions. The research starts from various question answering datasets where we take questions that are focused on different aspects like factual, contextual, logical, and grammatical answers. These questions are asked to a certain LLM in our case ChatGPT-4o mini which generates answers and by using the rubric we have built we use human annotation to detect hallucination by giving the answer a score out of 10. We initially started by giving all the answers a score of 10 then we minus if we detect certain hallucination levels. ChatGPT-4o mini was used because it is the latest version of ChatGPT and it is widely used. ChatGPT-4 premium was not used due to the lack of user base compared to ChaGPT-4o mini. The dataset is a collection of questions and answers designed to test the capabilities of language models in terms of factual accuracy and multi-hop reasoning. It consists of four primary components: SQuAD, which makes up 39.46% of the dataset and focuses on single-document question answering; HotpotQA, which constitutes 34.77% and tests multi-hop reasoning by requiring models to combine information from multiple documents; TriviaQA, which accounts for 18.79% and involves trivia- based questions with multiple possible correct answers; and handpicked data, which makes up 6.40% of the dataset and is a custom dataset designed to focus on specific cases where hallucinations are more likely. To generate the final answers, questions were collected and then processed through ChatGPT-4o mini. The output from this model was combined with data from an external API to produce the final results. This approach aimed to improve the overall accuracy and quality of the generated answer.

2.2 Dataset (Bhrantio Satya) Analysis

Collecting 14857 questions from 3 datasets and also added our handpicked questions. There are different types of questions in our dataset as follows.

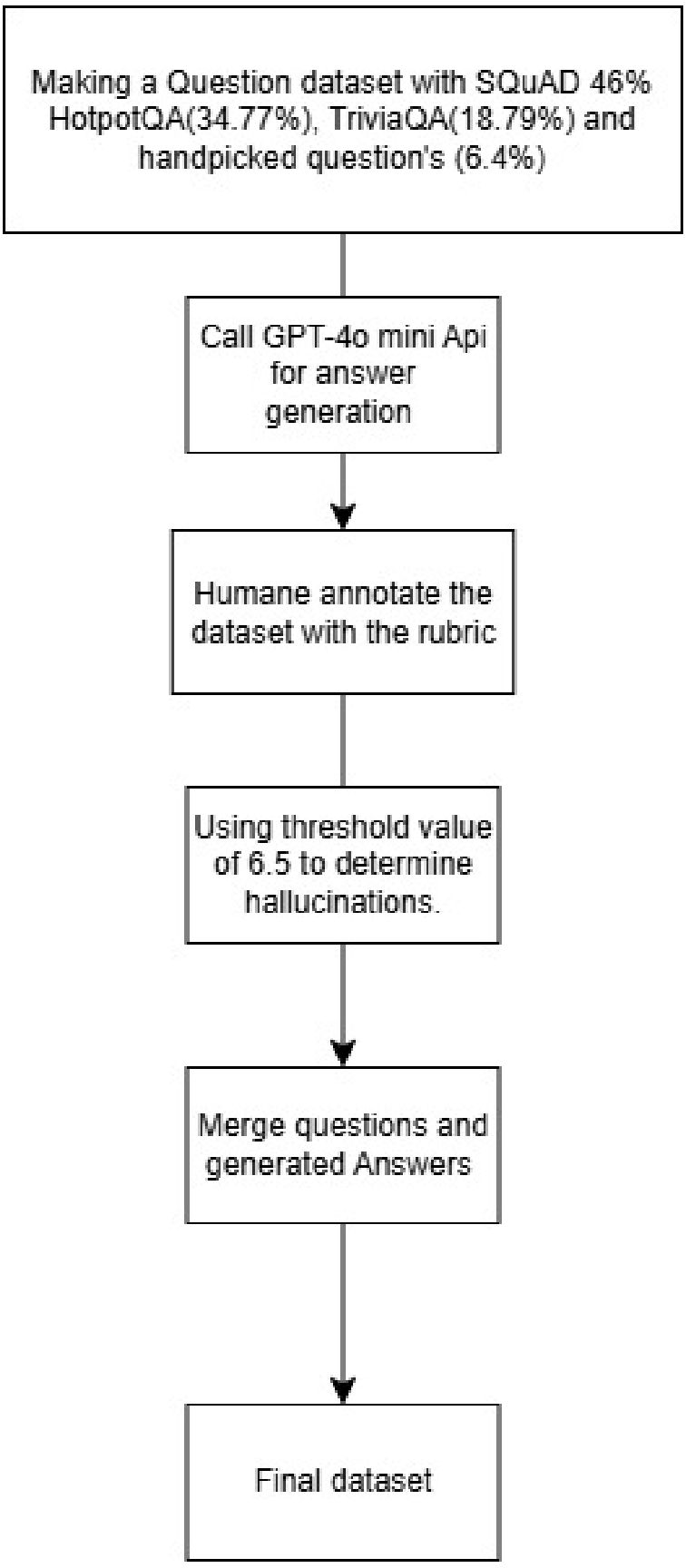


Figure 2.1: Process of Dataset Creation

Question Type	**Example**

Open-ended and informational questions	- You are an AI assistant. You will be assigned a task. You must generate a detailed and long answer. - What is the best time to post on Facebook and social media sites in India? - What are the best times to post on social media in India? Are these two questions rephrased by each other?
Causal or cause-and-effect questions	- Answer the following question: John was watching the movement of a stream in the forest. He noticed that the faster-moving stream could carry particles that are Larger or Smaller? - Assuming that: In erosion, a slower-moving stream will only carry smaller particles.
Subjective or interpretive questions	- Context: Aubrey challenged Tracy's friends to a game of hide and seek for an hour. - Question: How would Aubrey feel afterwards? Which one of these answers best answers the question according to the context? A: lazy B: passive C: sneaky
Language or translation questions	- Question: Detailed Instructions: In this task, you are given a sentence or phrase in English. You must translate it to Xhosa in a way that is equivalent in terms of meaning and grammatically correct. - Problem: This provided a sheltered stowage which is not normally included in the net or register tonnage. - Solution:
Factual questions	- Whose right to the English throne did Lady Jane Grey challenge in 1553?
Interpretive or subjective questions	- The Rock's persona is seen as what?
Informational or technical questions	- How did producers sometimes make edits of drum music?
Philosophical questions	- Plato's philosophy attempted to explain life from what standpoint?
Multiple-choice questions	- You are an AI assistant. User will give you a task. Your goal is to complete the task as faithfully as you can. - While performing the task, think step-by-step and justify your steps. Q: What bond is the force of attraction that holds together positive and negative ions? Choices: - Protonic - Electric - Magnetic - Ionic
Interpretive or subjective question	- You are an AI assistant that helps people find information. User will you give you a question. - Your task is to answer as faithfully as you can. While answering think step-by-step and justify your answer. Come up with a question and stream of consciousness reasoning.
Multiple factual question	- You are an AI assistant. You will be given a task. You must generate a detailed and long answer. - (CNN) – Condolences continued to pour in late Sunday night following the death of heavy metal rocker Ronnie James Dio, who lost his battle with stomach cancer earlier in the day. 1. Who died? 2. How? 3. Who made the statement?

There are many other types of questions in our datasets but here you have the most common ones which populate most of our dataset. The answers were generated by ChatGPT-4o mini which were annotated by a specific rubric created by sampling hallucinated data.

2.3 Scoring Rubric

A rubric was implemented using the human annotation method to categorize the varying levels of hallucination on an LLM. We are giving values based on 5 annotation classes that describe how the rubric works.

2.3.1 Annotation Class

- Factual Accuracy: The presence of verifiable incorrect facts or fabricated information severely under- mines the reliability of a response. If the response contains blatantly wrong details that can be easily checked against credible sources, it reflects a significant level of hallucination. A deduction is made if the facts presented are incorrect. Conversely, responses that contain verifiable and accurate facts enhance credibility and score positively.

- Contextual Hallucination: Responses that introduce off-topic or irrelevant information distort the original ques- tion's meaning. If a model strays significantly from the input context, it indicates a lack of understanding, warranting a score deduction. In contrast, responses that remain fully aligned with the input prompt and do not introduce unrelated infor- mation are valued and rewarded.

- Logical and Incoherent Hallucination: Logical flow is crucial for any coherent response. If a response is filled with illogical statements or contradictions, it suggests a severe hallucination, leading to a score reduction. Responses that maintain a clear and logical structure, presenting ideas cohesively, are considered high-quality and positively scored.

- Grammatical Hallucination: It refers to when large language models fail to follow the rules of grammar and punctuation .It shows the models lack of understanding of that language .It hap- pens when Large language model have build enough neuron to grasp the language and LLM is still a bit underfitted

These samples were annotated for 1 month and have resulted in 11000+ Human annotated samples. We have started with an initial score of 10 for every sample. And then we have deducted based on our rubric scores. This process gave us scores accurate to a type of hallucination. The rubric is as follows:

Feature	Type	Cost
Factual Hallucination	If a fact is not accurate at any point.	-8
	The LLM is deviating from the actual fact.	-4
	The fact is clear but not well put.	-1.5
Contextual Hallucination	When the answer is not correctly reported, it seems to mean something different from the meaning that was intended.	-6
	Some minor deviations from input.	-3
	Fully consistent with the input, but unrelated information.	-1
Logical Hallucination	Not illogical, not nonsensical sentences, or outright contradictions present.	-5
	Some disconnection or unclear sections, generally understandable.	-2.5
	Flows logically with clear ideas but not coherence.	-0.9
Incoherent Hallucination	Overconfident in incorrect statements, presenting falsehoods as facts.	-3
	Displays some hedging or uncertainty in statements.	-1.5
	Confident in correct but not verifiable information.	-0.5
Grammatical Hallucination	Obvious grammatical mistake.	-2
	Minor grammatical mistakes including commas, full stops, etc.	-1
	Mistakes in parts of speech and very minor mistakes.	-0.3

Table 2.2: Rubric Scoring

This process consists of each member of our team annotating every answer provided by ChatGPT-4o mini. The answer is judged on the criteria of the rubric and is given a 10 if it's perfect.

The model-based scoring of the samples:

Question	**Answer**	**Inspection**	**Hallucinations**	**Scoring**
In 1902 Newton Heath football club changed their name to what?	In 1902, Newton Heath football club changed their name to Manchester.	Here you can see that it wasn't specified which Manchester team was named.	**Factual Hallucination:** (The fact was not accurate)	-4
		Also the details about the local brewery owner John	**Contextual Hallucination:** (The entire context was not given)	-3
		Henry Davies was not mentioned.	**Incoherent Hallucination:** (It's confusing which team was named)	-1.5
			Total:	**10-4-3-1.5=1.5**

Table 2.3: Dataset annotation scoring example

2.4 Data Preprocessing

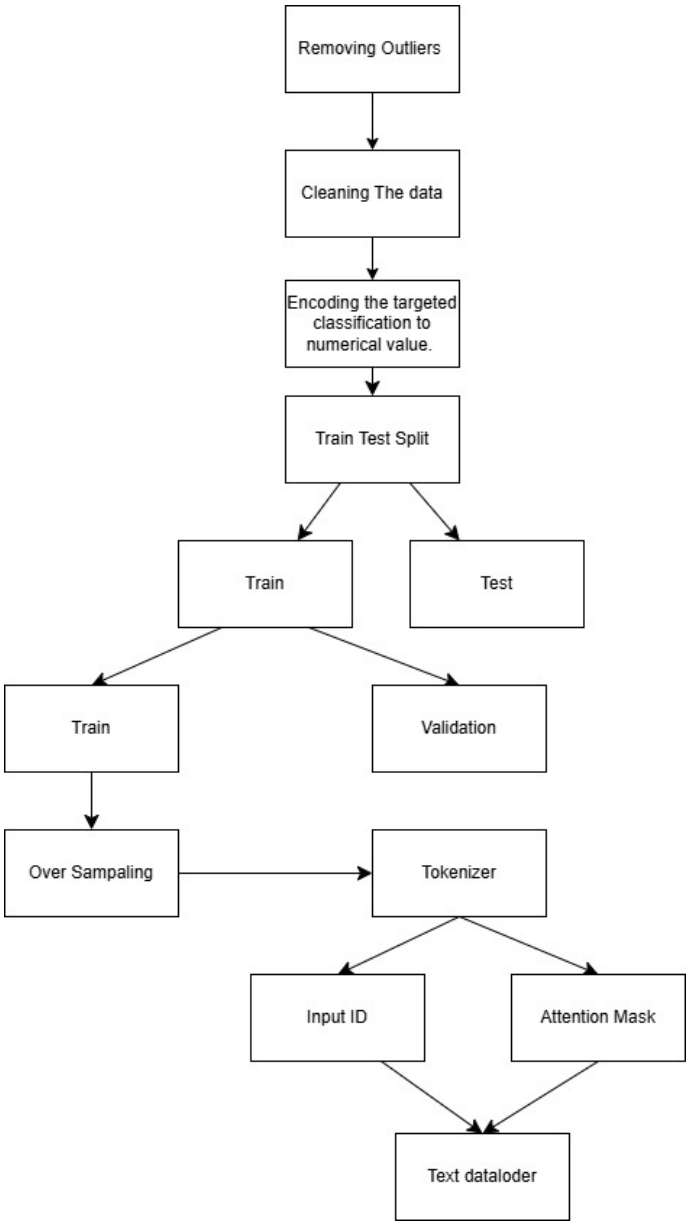


Figure 2.2: Data Processioning

Data preprocessing is a method to clean the dataset to increase the accuracy of the model. Datasets have a lot of unnecessary data which is not needed when we are training a model. For this purpose, a portion of the data that is needed for human annotation or our understanding is fine-tuned to increase the accuracy of the model. The process that we followed was removing unnecessary columns and dropping rows with null values. Here we get rid of the classification of different hallucinations and also, we get rid of values or rows that are not annotated. Then we clean the data to remove URLs, unnecessary spaces, and other unnecessary variables and convert everything to lowercase character. Then, encoding is done to alter the targeted class to the required numerical value. After that, we take the raw data text and apply corresponding labels from the data frame. DistilBERT is used for tokenization. Distribution of text length is done for natural language processing which refers to

text lengths distributed throughout the dataset. Then the prepared tokenized text, attention mask, and label for the model. It is done to get a better understanding of the dataset's characterizations.
Throughout the different stages of the data prepossessing process, the distribution of the label variable indicates how the target variable, also referred to as the dependent or output variable, is distributed over various classifications or values within the dataset. In supervised machine learning problems, the label variable identifies what is to be predicted. Knowing its distribution is quite useful as it may highlight some weaknesses such as a class imbalance or outliers.

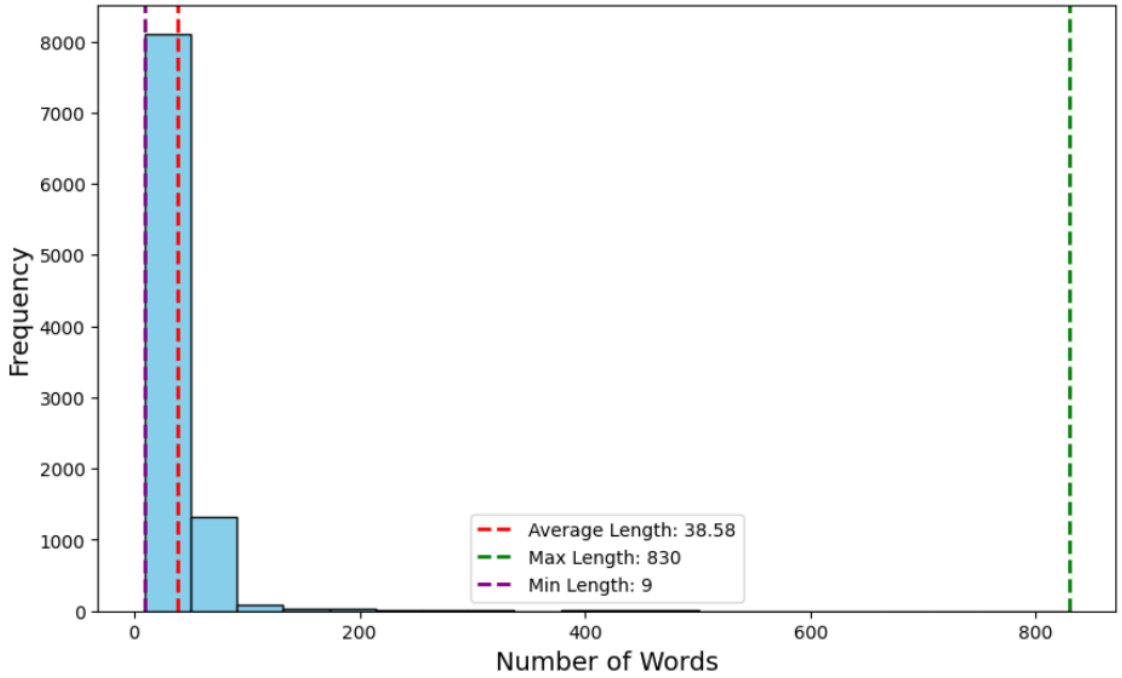


Figure 2.3: Distribution Of Text Length

2.5 Dataset Report

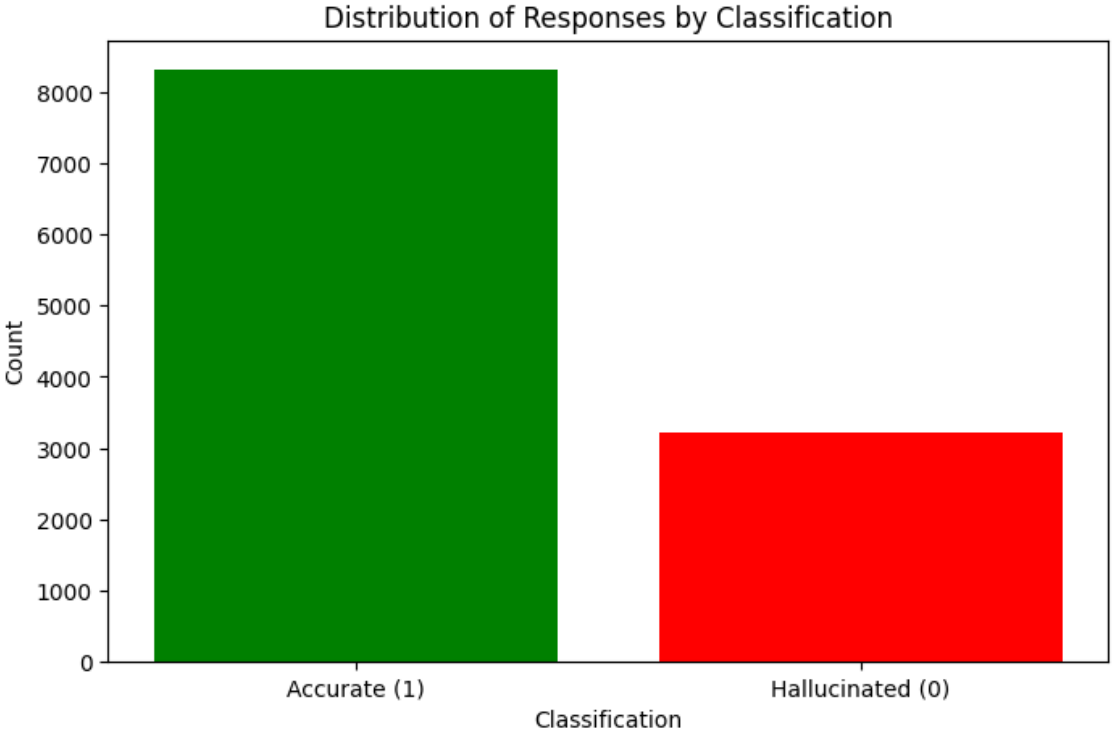


Figure 2.4: Distribution of Responses by Classification

Accurate responses (green bar) exceed 8,000, while hallucinated responses (red bar) are close to 3,000, indicating a higher frequency of accurate responses in the dataset.

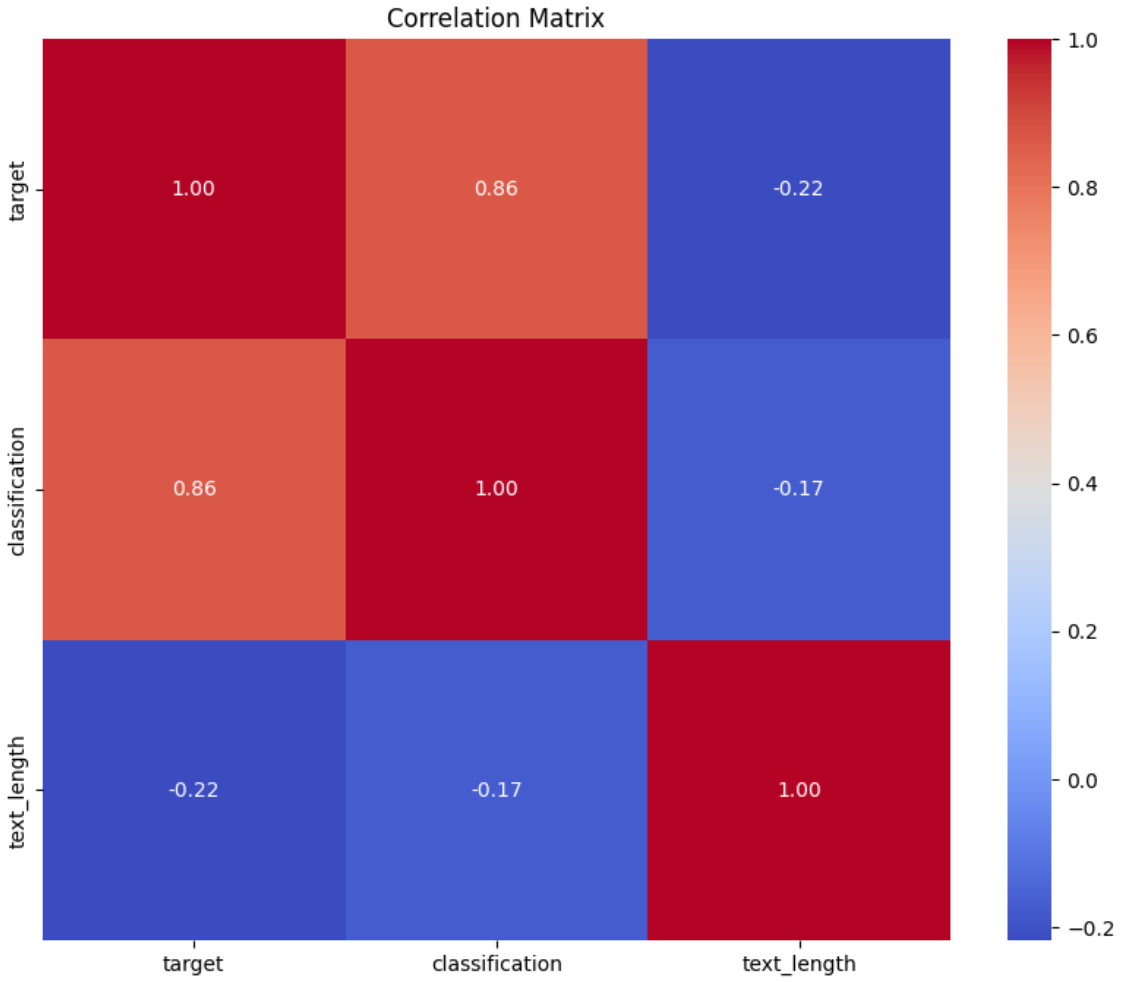


Figure 2.5: Correlation Matrix

The correlation matrix illustrates the relationships between target, classification, and text length. The diagonal values are 1.00, representing perfect positive correlations. The correlation between target and classification is strong and positive (0.86), while text length shows weak negative correlations with both target (-0.22) and classification (-0.17), indicating that longer text lengths slightly decrease their association. The color gradient ranges from blue (negative correlation) to red (positive correlation).

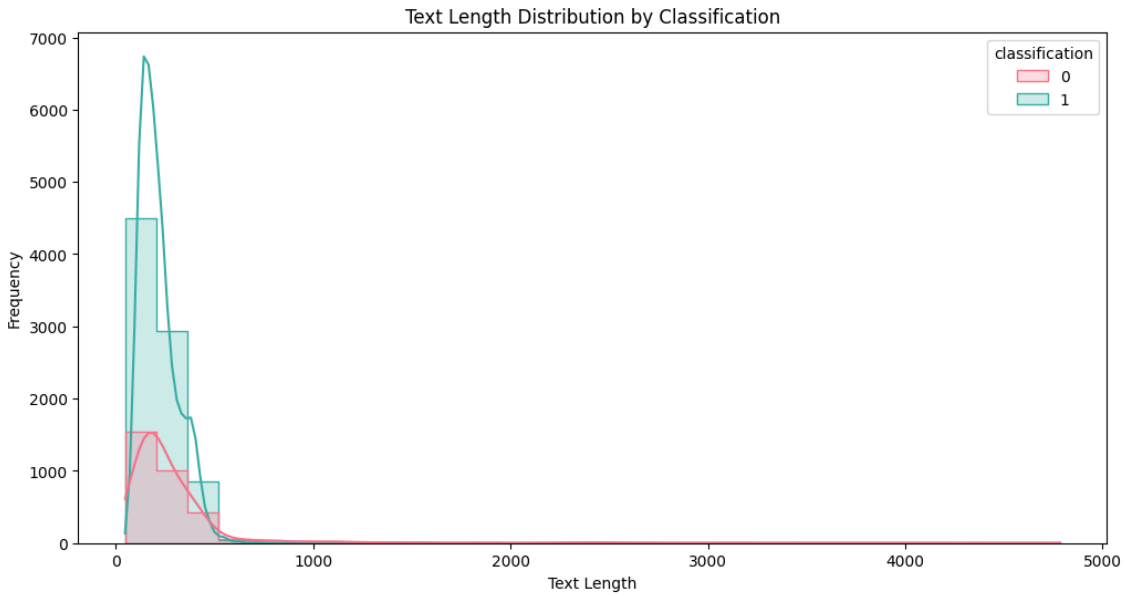


Figure 2.6: Text Length Distribution by Classification

The distribution of text lengths for two classifications (0 in red and 1 in blue), with both a histogram and KDE overlay. The x-axis represents text lengths (ranging from 0 to 5000 characters), while the y-axis shows their frequency. Most texts are concentrated at shorter lengths, with a sharp peak near lower values, and classification 1 has a higher overall frequency compared to 0. The distributions decline as text length increases, with few samples exceeding 2000 characters.

Chapter 3

Methodology

3.1 Model Selection and Training

We evaluate several state-of-the-art transformer models for hallucination detection in Large Language Models (LLMs), including BERT, ALBERT, DistilBERT, DeBERTa, Longformer, and FLAN T5.

- **BERT (Bidirectional Encoder Representations from Transformers)**: A widely-used baseline model that leverages bidirectional attention mechanisms to achieve strong contextual understanding. BERT's ability to capture relationships between words in both directions makes it a robust choice for detecting inconsistencies in generated text.

- **ALBERT (A Lite BERT)**: An efficient variant of BERT that reduces memory consumption and improves training speed through parameter sharing and factorized embeddings. Despite its lightweight design, ALBERT maintains competitive performance, making it suitable for large-scale hallucination detection tasks.

- **DistilBERT**: A distilled version of BERT that retains 97% of its performance while being significantly faster and lighter. DistilBERT is ideal for scenarios where computational efficiency is a priority without compromising on detection accuracy.

- **DeBERTa (Decoding-enhanced BERT with Disentangled Attention)**: An enhanced version of BERT that introduces disentangled attention mechanisms, allowing the model to better capture contextual relationships and improve representation learning. This makes DeBERTa particularly effective for identifying subtle hallucinations in complex text.

- **Longformer**: Designed to handle long sequences, Longformer employs a combination of local and global attention mechanisms. This capability is especially useful for multi-hop reasoning tasks, where detecting hallucinations requires understanding context across extended text spans.

- **FLAN T5**: A versatile model fine-tuned on instruction-based tasks, FLAN T5 excels in both comprehension and generation tasks. Its adaptability makes it a strong candidate for hallucination detection, particularly in scenarios requiring nuanced understanding of generated content.

The models are fine-tuned using our custom dataset of question-answer pairs. Training is conducted with the AdamW with a learning rate of 3×10^{-5} and a weight decay of 0.01 to prevent overfitting. Binary cross-entropy loss is used for classification, as the task involves distinguishing between hallucinated and non-hallucinated text. Models are trained for up to 50 epochs, with early stopping implemented if no improvement is observed for 5 consecutive epochs. This ensures efficient training while avoiding overfitting. Performance is evaluated on the test set to assess the model's ability to detect hallucinations in unseen data.

3.2 Model Optimization

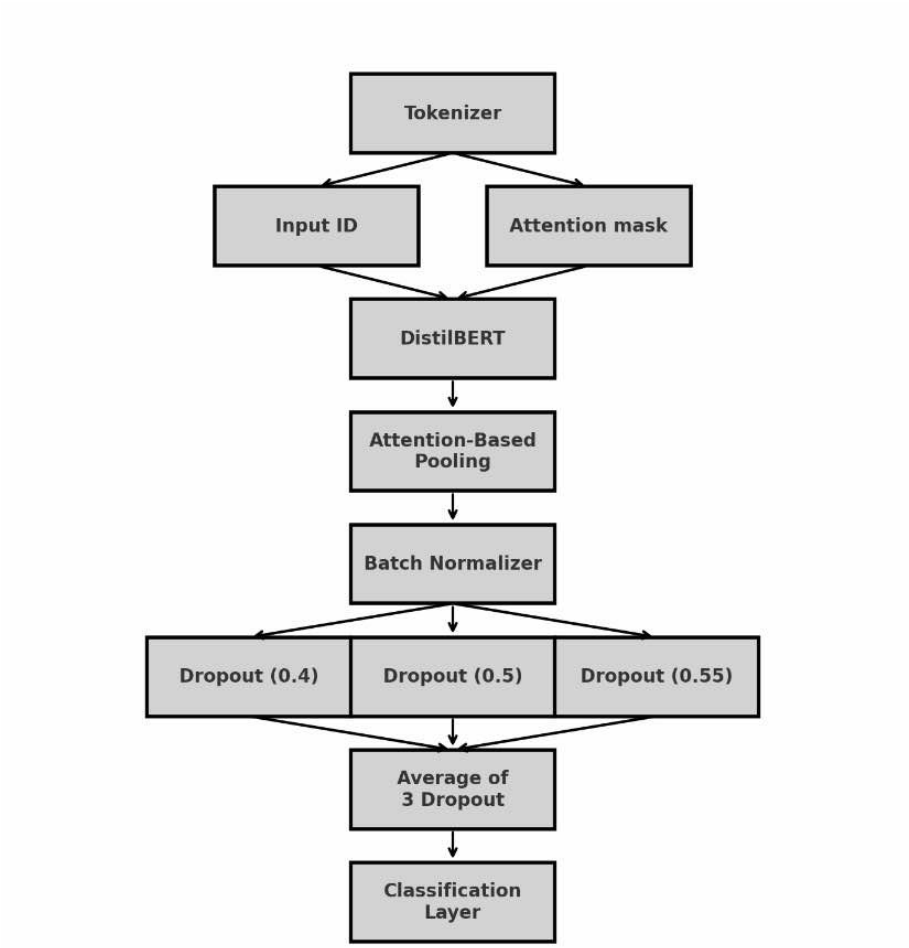


Figure 3.1: Fine-Tuned DistilBERT

After evaluating several transformer models, we selected DistilBERT for further fine-tuning, as its lower complexity aligns with our dataset "Bhrantio Satya." To enhance its capabilities, we optimized the model by integrating attention-based pooling, batch normalization, and a multi-sample dropout strategy. These modifications were designed to improve the model's ability to focus on important features while preventing overfitting. The final optimized architecture consists of the following components:

- **DistilBERT** as the base transformer model.
- **An attention-based pooling mechanism** that computes weighted feature representations, allowing the model to focus on key tokens in the input sequence.
- **A batch normalization layer** applied to stabilize and normalize activations, improving generalization.
- **Multi-Sample Dropout** with three dropout layers (0.4, 0.5, and 0.55) to enhance robustness by averaging outputs from multiple dropout variations.

Instead of L2 regularization, we employed a **cosine annealing learning rate scheduler**, which dynamically adjusts the learning rate over time to enhance convergence and stability. Additionally, we incorporated **label smoothing** in the loss function to mitigate overconfidence in predictions and improve generalization.
For handling class imbalance, we integrated a **focal loss function** with hyperparameters $\gamma = 1.5$ and $\alpha = 3$, which adjusts the contribution of easy and hard-to-classify examples. This approach ensures that the model focuses more on difficult cases, thereby improving performance on hallucinated responses. The focal loss is computed as follows:

$$FL(p_t) = \alpha(1 - p_t)^{\gamma} \cdot CE(p_t)$$

where p_t is the predicted probability and CE represents the cross-entropy loss with label smoothing.
For training, we used the **AdamW optimizer** with a learning rate of 6×10^{-6}, a weight decay factor of 0.1, and a batch size of 16. The model was trained for up to 50 epochs, with **early stopping** applied if the validation loss did not improve for 5 consecutive epochs. The learning rate was dynamically adjusted using a cosine annealing scheduler. Training and validation losses were monitored throughout the process, and the final evaluation was conducted on a separate test set to assess the model's ability to generalize to unseen data.

Chapter 4

Results And Discussion

4.1 Experimental Analysis

A classification report provides the F1-score, precision, and recall of each class taken by a model on a classification job. The generalization basis for a model is its test set.

Precision: Precision basically is the measure of how well the model is predicting the future. It is the ratio of correctly predicted positives compared to all predicted positives.

$$Precision = \frac{TP}{TP+FP}$$

Recall (Sensitivity): Model recall is the model's ability to identify actual positives; that is, the ratio of all actual positives to the correctly predicted positive.

$$Recall = \frac{TP}{TP+FN}$$

F1 Score: F1 score The F1 score uses its harmonic mean to balance recall and precision. As a single metric when you want to take both false positives and false negatives into consideration.

$$F1Score = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)}$$

Weighted Average: Weighted average With weights defined with class support, the weighted average computes the mean of metrics for each class. This accounts for class imbalances.

WeightedAvg.Precision,

$$= \frac{(Precisionclass1 \times Supportclass1 + Precisionclass2 \times Supportclass2 + ... + PrecisionclassN \times SupportclassN)}{TotalSupport}$$

Accuracy: Accuracy To assess overall correctness, accuracy considers the ratio of correct predictions compared to the total number of instances.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Support: Support represents the actual instances in each class, giving insight into class distribution and potential imbalances.

Macro Average: Macro average provides the mean performance of all classes for precision, recall, and F1-score, without considering class size.

$$MacroAverage = \frac{(Precisionclass1 + Precisionclass2 + ... + PrecisionclassN)}{N}$$

4.2 Model Performance Overview

Table 4.1 presents a comparison of the transformer models used in this study, focusing on essential metrics such as accuracy, precision, recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.67	0.66	0.62	0.63
ALBERT	0.65	0.69	0.65	0.66
DeBERTa	0.66	0.68	0.66	0.67
DistilBERT	**0**.67	**0**.66	**0**.67	**0**.67
Longformer	0.64	0.61	0.68	0.62
FLAN T5	0.61	0.69	0.61	0.63

Table 4.1: Overview of Model Performance

From the table, we observe that **DistilBERT** equals or outperforms other models in terms of F1-score, precision, and recall, model complexity, and overall suitability for our dataset, making it the best candidate for further fine-tuning.

4.3 Fine-Tuned DistilBERT Performance

The performance of the fine-tuned DistilBERT model on our test dataset is shown in Table III. This model was evaluated on its ability to classify hallucinated (Class 0) and non hallucinated (Class 1) responses.

Class	Precision	Recall	F1-Score	Support
Class 0 (Hallucinated)	0.50	0.54	0.52	346
Class 1 (Non-Hallucinated)	0.80	0.87	0.82	807
Accuracy	0.71			1153
Macro Avg	0.64	0.62	0.63	1153
Weighted Avg	0.68	0.70	0.69	1153

Table 4.2: Performance Metrics of Fine-Tuned DistilBERT Model

The model's performance for Class 0 (hallucinated responses) shows a significant gap, with a precision of 0.50 and a recall of 0.54. This indicates that the model struggles to accurately identify hallucinated responses, often missing subtle or context-dependent hallucinations. In contrast, for Class 1 (non-hallucinated responses), the model performs notably better, achieving a precision of 0.85 and a recall of 0.84, indicating a stronger ability to correctly classify non-hallucinated content.

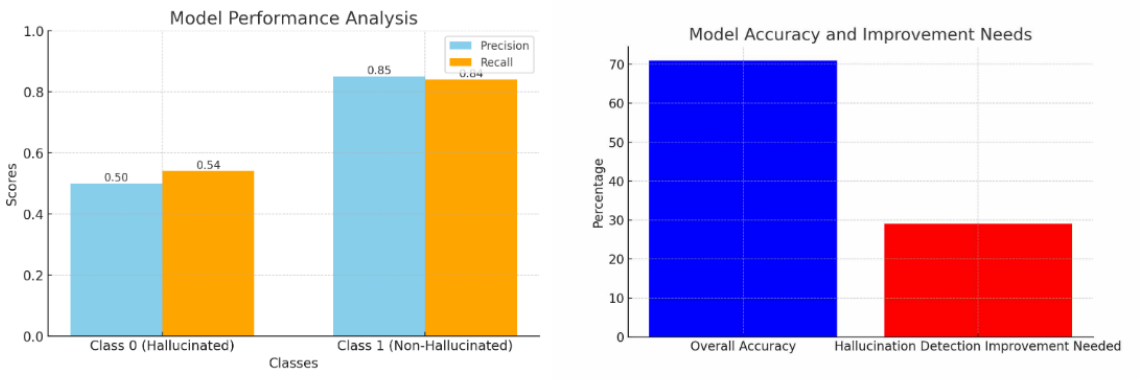


Figure 4.1: Model Performance and Accuracy

The overall accuracy of the model is 71%, reflecting room for improvement, particularly in hallucination detection.

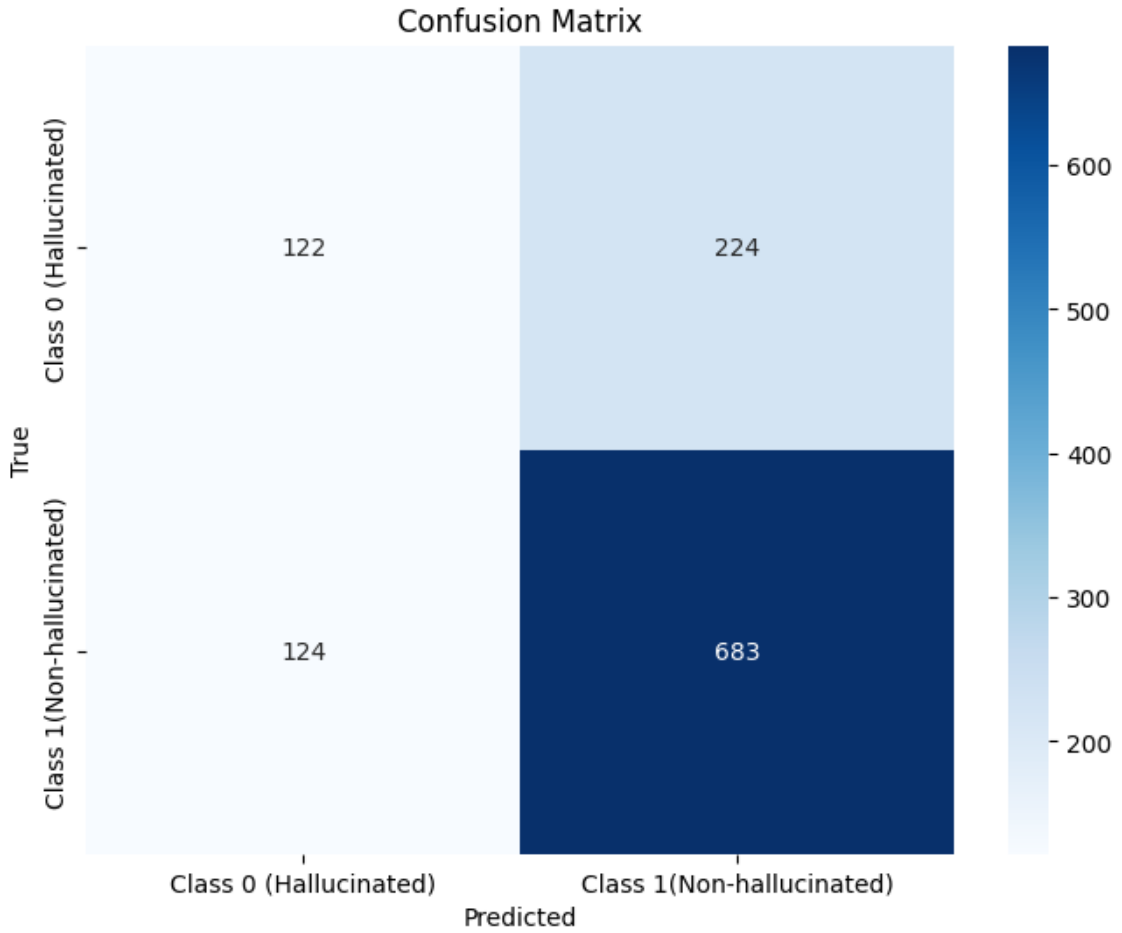


Figure 4.2: Confusion Matrix for Fine-Tuned DistilBERT Model

Despite an overall accuracy of 0.71 (Figure 4.1), the macro average recall of 0.72 highlights the need for improvement, particularly in detecting hallucinated responses. The confusion matrix (Figure 4.2) further illustrates this imbalance. The ROC curve

(Figure 4.3) provides additional insights into the model's performance, particularly in differentiating between hallucinated and non-hallucinated responses.

4.4 Discussion

The results reveal that while DistilBERT performs well overall, it struggles to detect hallucinated responses, with low recall

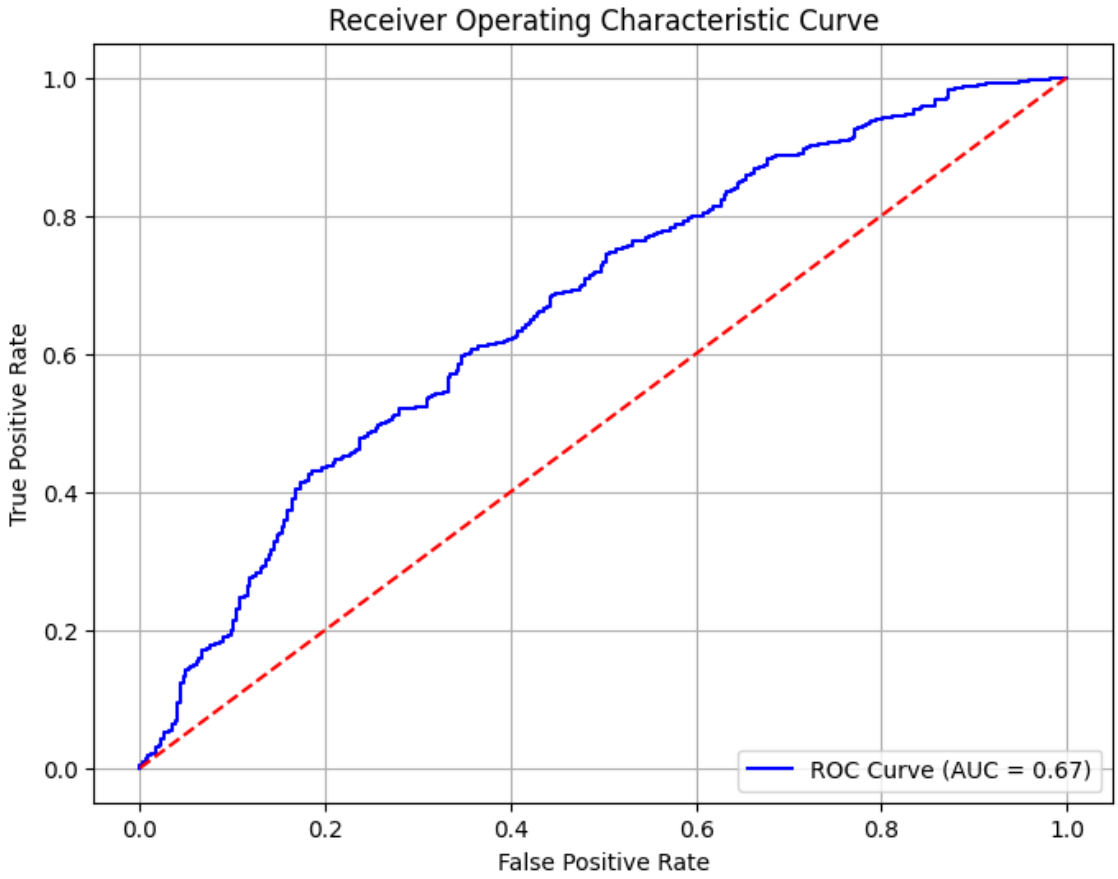


Figure 4.3: ROC Curve (DistilBERT)

For Class 0, the model faces challenges in identifying subtle and context-dependent hallucinations, likely exacerbated by dataset imbalance. Addressing these limitations may require improving dataset balance or applying data augmentation techniques. Incorporating external knowledge or domain-specific constraints could also enhance the model's ability to detect hallucinations. These findings highlight the complexity of hallucination detection and the need for further research to improve model reliability in critical applications.
From our experiments, we observe that **DistilBERT** equals or outperforms other models in terms of computational efficiency, model complexity, and overall suitability for our dataset, making it the best candidate for further fine-tuning.
Future work should focus on addressing these challenges through strategies such as dataset balancing, adversarial training, and exploring more advanced model architectures that can better distinguish between factual and hallucinated content. These improvements are crucial for enhancing the reliability and trustworthiness of

LLMs, especially in high-stakes domains like healthcare and law, where accuracy and dependability are critical.

Chapter 5

Conclusion

In this paper, we addressed the challenge of hallucinations in Large Language Models (LLMs) and proposed a structured framework for detecting and mitigating hallucinated responses. Through extensive experimentation with several state-of-the-art transformer models, we identified DistilBERT, enhanced with attention-based pooling and multi-sample dropout, as the most effective model for this task, achieving an overall accuracy of 71%. However, the model exhibited notable limitations in detecting hallucinated responses, as indicated by low precision (0.50) and recall (0.54) for Class 0 (hallucinated responses). While our optimized DistilBERT model demonstrates promise in this domain, the results underscore the need for further refinements, particularly in capturing subtle hallucinations.All in all, LLMs are the future that will likely replace Google as our main virtual assistant in the ever-changing digital world where accuracy of the information is tantamount to our livelihood. So finding and resolving these issues as LLMs become more widely available to everyone is something we hope to resolve with our work.

Bibliography

[1] S. Bickel, P. Haider, and T. Scheffer, "Predicting sentences using n-gram language models," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, R. Mooney, C. Brew, L.-F. Chien, and K. Kirchhoff, Eds., Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 193–200. [Online]. Available: https://aclanthology.org/H05-1025.

[2] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," vol. 2, Sep. 2010, pp. 1045–1048. DOI: 10.21437/Interspeech.2010-343.

[3] A. Pauls and D. Klein, "Faster and smaller n-gram language models.," Jan. 2011, pp. 258–267.

[4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *arXiv (Cornell University)*, Jan. 2017. DOI: 10.48550/arxiv.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762.

[5] K. Lee, O. Firat, A. Agarwal, C. Fannjiang, and D. Sussillo, "Hallucinations in neural machine translation," 2018, [Online]. Available: https://openreview.net/forum?id=SkxJ-309FQ.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533.

[8] Stanford Encyclopedia of Philosophy, *Theories of meaning*, Spring 2021 Edition, [Online]. Available: https://plato.stanford.edu/archives/spr2021/entries/meaning/, Jun. 2019.

[9] *Theories of Meaning (Stanford Encyclopedia of Philosophy/Spring 2021 Edition)*, Jun. 2019. [Online]. Available: https://plato.stanford.edu/archives/spr2021/entries/meaning/.

[10] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language Models are Few-Shot Learners," *arXiv (Cornell University)*, Jan. 2020. DOI: 10.48550/arxiv.2005.14165. [Online]. Available: https://arxiv.org/abs/2005.14165.

[11] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., [Online]. Available: https://aclanthology.org/2020.acl-main.173, Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919.

[12] M. Cao, Y. Dong, and J. Cheung, "Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3340–3354. DOI: 10.18653/v1/2022.acl-long.236. [Online]. Available: https://aclanthology.org/2022.acl-long.236.

[13] A. Chowdhery, S. Narang, J. Devlin, *et al.*, "PaLM: Scaling Language Modeling with Pathways," *arXiv (Cornell University)*, Jan. 2022. DOI: 10.48550/arxiv.2204.02311. [Online]. Available: https://arxiv.org/abs/2204.02311.

[14] Z. Ji, N. Lee, R. Frieske, *et al.*, "Survey of Hallucination in Natural Language Generation," *arXiv (Cornell University)*, Jan. 2022. DOI: 10.48550/arxiv.2202.03629. [Online]. Available: https://arxiv.org/abs/2202.03629.

[15] T. Liu, Y. Zhang, C. Brockett, *et al.*, "A token-level reference-free hallucination detection benchmark for free-form text generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6723–6737. DOI: 10.18653/v1/2022.acl-long.464. [Online]. Available: https://aclanthology.org/2022.acl-long.464.

[16] Z. Cao, Y. Yang, and H. Zhao, "AutoHall: Automated Hallucination Dataset Generation for large Language Models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2310.00259. [Online]. Available: https://arxiv.org/abs/2310.00259.

[17] Y. Chen, Q. Fu, Y. Yuan, *et al.*, "Hallucination detection: Robustly discerning reliable answers in large language models," *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264350686.

[18] S. Choi, T. Fang, Z. Wang, and Y. Song, "KCTS: Knowledge-Constrained Tree Search Decoding with Token-Level Hallucination Detection," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2310.09044. [Online]. Available: https://arxiv.org/abs/2310.09044.

[19] Deutsche Bank, *Large language models are built to hallucinate.* https://www.db.com/what-next/entrepreneurial-success/Future-of-Work--Neue-Arbeitswelt/Burda/index?kid=wn.stageGHP.Burda-AI.en&language_id=1, 2023.

[20] B. A. Galitsky, "Truth-o-meter: Collaborating with llm in fighting its hallucinations," *Preprints*, Jul. 2023. DOI: 10.20944/preprints202307.1723.v1. [Online]. Available: https://doi.org/10.20944/preprints202307.1723.v1.

[21] L. Huang, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[22] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HalUEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2305.11747. [Online]. Available: https://arxiv.org/abs/2305.11747.

[23] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2303.08896. [Online]. Available: https://arxiv.org/abs/2303.08896.

[24] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2309.05922. [Online]. Available: https://arxiv.org/abs/2309.05922.

[25] B. Snyder, M. Moisescu, and M. B. Zafar, "On early detection of hallucinations in factual question answering," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2312.14183. [Online]. Available: https://arxiv.org/abs/2312.14183.

[26] H. Touvron, T. Lavril, G. Izacard, *et al.*, "LLAMA: Open and Efficient Foundation Language Models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2302.13971. [Online]. Available: https://arxiv.org/abs/2302.13971.

[27] L. K. Umapathi, A. Pal, and M. Sankarasubbu, "Med-HALT: Medical Domain Hallucination Test for large language models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2307.15343. [Online]. Available: https://arxiv.org/abs/2307.15343.

[28] W. Xu, S. Agrawal, E. Briakou, M. J. Martindale, and M. Carpuat, "Understanding and detecting hallucinations in neural machine translation via model introspection," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2301.07779. [Online]. Available: https://arxiv.org/abs/2301.07779.

[29] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, "Cognitive Mirage: A review of hallucinations in large language models," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2309.06794. [Online]. Available: https://arxiv.org/abs/2309.06794.

[30] Y. Z. et al., "Knowledge overshadowing causes amalgamated hallucination in large language models," *arXiv preprint arXiv:2407.08039*, 2024.

[31] A. Author, "Understanding hallucination in large language models," *Journal of Natural Language Processing*, vol. 12, no. 3, pp. 45–67, 2024.

[32] B. Author and C. Author, "Detection schemes for hallucinations in llms," in *Proceedings of the International Conference on Artificial Intelligence*, 2024, pp. 123–130.

[33] D. Author, *Mitigating Hallucinations in Language Models: Strategies and Solutions*. Publishing House, 2024.

[34] C. Jiang, B. Qi, X. Hong, and D. Fu, "On large language models' hallucination with regard to known facts," *arXiv preprint arXiv:2403.20009*, 2024.

[35] V. D. Kees, "The pitfalls of defining hallucination," *arXiv (Cornell University)*, Jan. 2024. DOI: 10.48550/arxiv.2401.07897. [Online]. Available: https://arxiv.org/abs/2401.07897.

[36] J. Li, J. Chen, R. Ren, *et al.*, "The Dawn After the Dark: An Empirical Study on Factuality Hallucination in large language models," *arXiv (Cornell University)*, Jan. 2024. DOI: 10.48550/arxiv.2401.03205. [Online]. Available: https://arxiv.org/abs/2401.03205.

[37] Y. Liang, Z. Song, H. Wang, and J. Zhang, "Learning to trust your feelings: Leveraging Self-awareness in LLMs for hallucination Mitigation," *arXiv (Cornell University)*, Jan. 2024. DOI: 10.48550/arxiv.2401.15449. [Online]. Available: https://arxiv.org/abs/2401.15449.

[38] J. Luo, T. Li, D. Wu, M. Jenkin, S. Liu, and G. Dudek, "Hallucination Detection and Hallucination Mitigation: An investigation," *arXiv (Cornell University)*, Jan. 2024. DOI: 10.48550/arxiv.2401.08358. [Online]. Available: https://arxiv.org/abs/2401.08358.

[39] A. Mishra, A. Asai, V. Balachandran, *et al.*, "Fine-grained hallucination detection and editing for language models," *arXiv (Cornell University)*, Jan. 2024. DOI: 10.48550/arxiv.2401.06855. [Online]. Available: https://arxiv.org/abs/2401.06855.

[40] R. W. Oelschlager, "Evaluating the impact of hallucinations on user trust and satisfaction in llm-based systems," 2024.