Enhancing Hallucination Detection in NLG: Novel Architectures and Optimization

Mohammad Sufyan Azam
IIIT Delhi
sufyan20312@iiitd.ac.in

Aamleen Ahmed
IIIT Delhi
aamleen20002@iiitd.ac.in

Subhanshu Bansal
IIIT Delhi
subhanshu20135@iiitd.ac.in

Siddhant Singh
IIIT Delhi
siddhant20338@iiitd.ac.in

***Abstract*—Abstract: In the domain of Natural Language Generation (NLG), ensuring semantic accuracy alongside fluency is paramount. Our NLG project addresses the pervasive challenge of hallucination detection in neural model outputs, where systems generate grammatically sound but semantically incorrect text. Leveraging custom architectures, pretrained models, and MetaModelling techniques, our approach operates in a post hoc setting to identify and mitigate hallucinations. Across diverse NLG tasks, including machine translation, paraphrase generation, and definition modeling, our methodology achieves an accuracy score of 82% and a Spearman's Rho score of approximately 74%. By focusing on hallucination detection, our project contributes to enhancing the reliability and trustworthiness of NLG systems in practical applications.**

***Index Terms*—Hallucination Detection, Natural Language Generation, Semantic Inconsistencies, Machine Translation, Paraphrase Generation, Dialogue Systems, Cross-Attention Mechanisms, MetaModel Framework**

I. Introduction

Hallucination detection in Natural Language Generation (NLG) systems is a critical task, as it addresses the discrepancy between fluency and semantic accuracy in generated outputs.The project is motivated by the prevalent issue of hallucination in NLG outputs, where generated text may be grammatically correct but semantically inaccurate. By addressing this challenge, the project endeavors to improve the accuracy and reliability of NLG systems, ensuring that generated text aligns more closely with the intended meaning. This has significant implications across various applications, including machine translation, paraphrase generation, and dialogue systems, where the correctness of generated text is paramount for effective communication and comprehension.

Our NLG project aims to tackle this challenge by developing a robust hallucination detection system that operates in a post hoc setting, without direct access to the generating model. Our approach integrates a variety of techniques, including custom architectures, pretrained models, and MetaModelling methods, to effectively identify and mitigate hallucinations across diverse NLG tasks.

One of the key components of our methodology is the development of custom architectures tailored specifically for hallucination detection. While existing approaches often rely on off-the-shelf models or standard architectures, our project introduces novelty by designing architectures optimized for this particular task. These custom architectures leverage insights from the NLG domain and incorporate specialized components to enhance the detection of semantic inconsistencies in generated text.

In addition to custom architectures, our approach leverages pretrained models fine-tuned specifically for hallucination detection. By fine-tuning models on hallucination detection datasets, we aim to improve their ability to discern between grammatically sound but semantically incorrect outputs and genuinely accurate ones. This fine-tuning process is crucial for adapting pretrained models to the nuances of hallucination detection, thereby enhancing their effectiveness in real-world NLG applications. Through the integration of custom architectures, fine-tuned pretrained models, and MetaModelling techniques, our project offers a comprehensive solution to the challenge of hallucination detection in NLG systems, contributing to the advancement of reliability and trustworthiness in NLG applications.

II. Related Works

a) [4] ***Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback****:* It explores advanced strategies for managing hallucinations in Large Vision Language Models (LVLMs). It introduces a framework that includes fine-grained AI feedback for generating detailed annotations at the sentence level to facilitate both hallucination detection and mitigation. The method involves creating a detect-then-rewrite pipeline for constructing preference datasets and integrating a Hallucination Severity-Aware Direct Preference Optimization (HSA-DPO) to prioritize the mitigation of more severe hallucinations. Extensive experimental results show that this approach not only improves the accuracy of hallucination detection but also enhances the model's ability to generate more reliable and faithful responses.

*b) [5]****Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection****:* It investigates the phenomenon of hallucinations in neural machine translation (NMT). The study first identifies internal model symptoms of hallucinations through contrastive analysis of hallucinated vs. non-hallucinated outputs generated by source perturbations. It then uses these symptoms to create a lightweight hallucination detector that outperforms existing methods on English-Chinese and German-English translation

test beds. The paper presents two main findings: the identification of distinctive source contribution patterns that characterize hallucinations and the development of a robust hallucination classifier. The classifier, based on Layerwise Relevance Propagation (LRP), demonstrates superior performance in detecting hallucinations compared to model-free baselines and classifiers based on quality estimation or large pre-trained models.

*c) [2]**A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models**:* It presents a comprehensive survey of over 32 techniques developed to mitigate hallucination in Large Language Models. The authors introduce a detailed taxonomy categorizing these methods based on various parameters such as dataset utilization, common tasks, feedback mechanisms, and retriever types. Additionally, the paper analyzes the challenges and limitations inherent in these techniques, providing a solid foundation for future research in addressing hallucinations and related phenomena within the realm of LLMs. The paper also delves into the detection of hallucinations, with methods such as mFACT, contextual information-based frameworks, and the investigation of self-contradiction as a contributing factor. By synthesizing essential features characterizing these techniques, the paper provides a foundation for more structured future research within the domain of hallucination mitigation.

*d) [1]**Hallucination Detection and Hallucination Mitigation: An Investigation**:* The paper focuses on the challenges posed by hallucinations in LLMs like ChatGPT, Bard, and Llama. The report provides a comprehensive review of current literature on hallucination detection and mitigation. It also discusses various methods and benchmarks for hallucination detection at both token and sentence levels, as well as mitigation strategies. They highlight the importance of accurate and reliable LLM outputs, especially in critical applications, and the need for effective techniques to detect and mitigate hallucinations to ensure the practical deployment of LLMs. The paper includes detailed discussions on metrics for evaluating hallucination detection and mitigation, and presents reproduced results of existing works to illustrate the effectiveness of different approaches.

III. Dataset

The SHROOM dataset for the SemEval 2024 Task on Hallucination Detection consists of three components: train, test, and validation. Each component includes data samples for three tasks: PG (paraphrase generation), MT (machine translation), and DM (definition modeling). The training data is unlabelled, while the validation and testing data have labels for hallucination detection.

IV. Baseline Approaches for Hallucination Detection

In this section, we present our baseline approaches for hallucination detection in Natural Language Generation (NLG) outputs. Hallucination, characterized by the generation of grammatically sound but semantically incorrect text, poses a significant challenge in ensuring the reliability and trustworthiness of NLG systems. Our baseline strategies aim to assess the effectiveness of various models and architectures in detecting semantic inconsistencies.

A. Overall Baseline Architecture

Our overall baseline architecture comprises several key components designed to capture semantic inconsistencies in NLG outputs. Beginning with an input layer, the architecture encodes the input text using pre-trained word embeddings to capture semantic information effectively. Subsequently, either convolutional layers or recurrent neural networks (RNNs), such as gated recurrent units (GRUs), long short-term memory (LSTM) networks, or bidirectional LSTMs (BiLSTMs), are employed to capture sequential dependencies and semantic relationships within the text. To facilitate attention mechanisms, attention heads are incorporated to dynamically weigh the importance of different parts of the input text. Dropout layers are strategically inserted to prevent overfitting, followed by pooling layers to aggregate features across the sequence. Finally, a dense layer with a softmax activation function is employed to produce probability scores for hallucination detection.

B. Task-based Architecture

In task-based architectures, the architecture is tailored to the specific characteristics of each NLG task, such as machine translation (MT), paraphrase generation (PG), and definition modeling (DM). Specialized components are integrated to handle the unique challenges posed by each task, leveraging task-specific features and information. This task-specific customization enables the architecture to effectively capture semantic inconsistencies within task-specific contexts, potentially improving hallucination detection accuracy.

C. Combined Architecture

In combined architectures, multiple NLG tasks are integrated within a unified framework. However, the integration of both context and target sentences in combined architectures poses significant challenges, leading to a notable dip in performance. Providing both sentences together, with the context representing what the output should have been and the target representing the sentence generated by NLG tasks, results in the oversight of relative positions and the merging of important features from hallucinated sentences. Consequently, the accuracy of combined architectures is compromised, underscoring the complexities associated with joint processing of context and target sentences.

D. Overall Observations

Our experimentation shed light on the performance disparities between task-based and combined architectures for hallucination detection in NLG outputs. Notably, task-based architectures consistently outperformed their combined counterparts, achieving accuracy scores around 0.6 compared to

TABLE I
BASELINE MODELS PERFORMANCE

Model	Task	Accuracy
Bi-LSTM + Bahadanau Attention	MT	0.68
LSTMs + Concatenate	PG	0.64
LSTM + Attention	DM	0.67
Bi-LSTM + Positional	Combined 3	0.515
Transformer (MHA)	Combined 3	0.51
Transformer (MHA) + Residual Connections	Combined 3	0.59
Transformer (MQA)	Combined 3	0.46
Transformer (MQA) + Residual Connections	Combined 3	0.50
Transformer (GQA)	Combined 3	0.6
Transformer (GQA) + Residual Connections	Combined 3	0.66

approximately 0.5 for combined architectures. This observation underscores the effectiveness of task-specific customization in capturing semantic inconsistencies, highlighting the importance of tailored approaches in hallucination detection. However, despite achieving moderate success, both task-based and combined architectures fell short of achieving optimal accuracy, with scores plateauing around 0.6. This observation suggests the presence of inherent challenges in effectively capturing and mitigating hallucinations within NLG outputs, necessitating further exploration and refinement of detection methodologies.

The integration of both context and target sentences in combined architectures posed significant challenges, leading to a notable dip in performance. Specifically, we provided both sentences together, with the context representing what the output should have been and the target representing the sentence generated by NLG tasks. However, this approach resulted in the oversight of relative positions and the merging of important features from hallucinated sentences. Consequently, the accuracy of combined architectures was compromised, underscoring the complexities associated with joint processing of context and target sentences. These findings highlight the need for nuanced approaches to address the challenges of combined learning in hallucination detection. In the subsequent sections, we propose a refined methodology aimed at mitigating these issues and enhancing the accuracy of hallucination detection in NLG systems.

V. METHODOLOGY

A. Our Custom Architecture: *HIT-e: Hallucination Interactive Tri-encoder*

To improve the detection of mistakes in NLG outputs in baseline, we've devised a new method within our neural network architecture. We aim to address previous issues and make our approach more effective by separately working on the sentences. [3]

1) **Hypothesis Encoder:** This crucial component is where we begin dealing with the text generated by the LLM, such as translations, paraphrases, or dialogues. Here, we convert this text into numerical representations, taking into account the order of words and their meaning. To ensure our model understands the sequence's relative order, we incorporate positional embeddings into the representation. These embeddings help the model understand where each word fits in the sentence.

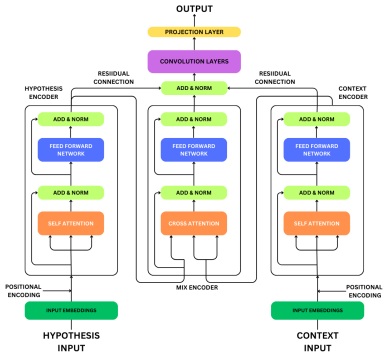

Fig. 1. HITe: Hallucination Interactive Tri-Encoder

2) **Context Encoder:** Working alongside the Hypothesis Encoder, this component handles the context or background information relevant to the task. Similar to the Hypothesis Encoder, it transforms this contextual data into numerical representations, paying close attention to word order and meaning. Utilizing attention mechanisms and other techniques, it refines the representation to focus on the essential aspects of the text.
3) **Mixed Encoder:** After processing through the Hypothesis and Context Encoders, the representations from both paths are combined in the Mix Encoder. This central component facilitates a detailed comparison between the LLM's output and the provided context, using multi-head cross-attention mechanisms to identify discrepancies indicative of hallucinations. The output of the Mix Encoder is then added and normalized with the outputs from the Hypothesis and Context Encoders in a residual layer. Subsequently, the residual layer's output is passed through convolutional layers and finally through projection layers.
4) **Residual Layer and Convolution:** To mitigate challenges such as gradient overflow during training, residual connections and convolutional layers are integrated into the architecture. These enhancements help maintain information flow and gradients throughout the network, contributing to improved learning and convergence.

In both the Hypothesis and Context Encoders, the process follows a similar flow: the input text undergoes embedding, where each word is represented as a numerical vector, and positional embeddings are added to capture word order. Next, the encoded representations are refined using attention mechanisms and feed-forward networks to extract meaningful semantic features from the text.

The incorporation of these elements has resulted in a notable enhancement in accuracy scores, from approximately 0.5 to 0.7, highlighting the effectiveness of our refined approach in bolstering hallucination detection methodologies in NLG systems.

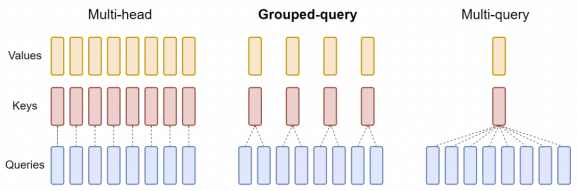

Fig. 2. Multi Head, Multi Query & Grouped Query Attention

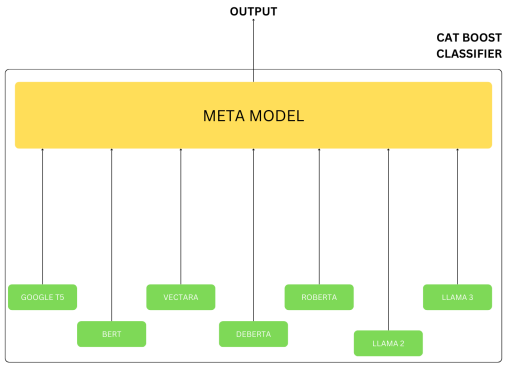

Fig. 3. Meta Model WorkFlow

B. Base Models

To detect hallucinations, we leveraged a total of 13 pre-trained models as the base models for our ensemble model. These models are deberta self check nli, deberta self check nli inverse, facebook roberta large, google t5, deberta base, microsoft deberta base, mistral, llama, microsoft deberta v2 xxlarge, microsoft deberta v2 xxlarge mnli, openchat, sentence transformer, and the vectara hallucination model.

We evaluated the performance of each of these models based on accuracy and Spearman coefficient. The results showed that the probabilities produced by these models were in the range of 0.4 to 0.5, indicating a need for further refinement to improve performance.

To address this, we implemented a meta-model through catboost that utilizes the predicted scores from the base models and trains on them. This meta-model aims to combine the strengths of the individual models and enhance the overall performance in detecting hallucinations.

C. Meta Model

In our preliminary investigations, we evaluated several base models, including renowned architectures such as RoBERTa, Google T5, Mistral 7b, LLAMA2, and LLAMA3. Despite their sophistication, these models exhibited suboptimal accuracy scores, primarily due to probabilities residing near the critical switch zone of 50%. This ambiguity posed a significant challenge in accurately discerning hallucination occurrences.

To address this limitation, we devised a series of optimization strategies. Firstly, we introduced two probabilities for each sentence sample, evaluating both entailment and contradiction, aiming to strengthen the predictive capabilities of our models. However, the breakthrough came with the integration of these probabilities into a MetaModel framework. This MetaModel, empowered by CatBoost and XGBoost Regressor, leveraged the collective insights from the base models to refine predictions and mitigate uncertainty. Through this integrated approach, we observed a substantial improvement in accuracy scores, achieving approximately 0.83. Moreover, our methodology yielded promising qualitative results, as evidenced by a notable increase in RHO (Spearman) scores, highlighting the enhanced reliability and robustness of our model predictions. The architecture of our model is given in *Fig 3*.

In the training process, we first train multiple base models, totaling 13 in our case, on the given dataset. These base models generate entailment and contradiction probabilities for each sentence sample. Subsequently, we employ MetaModeling techniques, specifically CatBoost and XGBoost Regressor, to learn from the outputs of these base models. The MetaModel integrates the probability scores generated by the base models and refines predictions based on collective insights. This refined approach enhances the accuracy and reliability of hallucination detection by leveraging the diverse perspectives captured by the ensemble of base models. Finally, the MetaModel filters across all the base models and generates the final labels, indicating whether the generated text contains hallucinations or not.

VI. Results

Our evaluation metrics primarily comprised accuracy and RHO (Spearman), providing comprehensive insights into the efficacy of our methodologies in hallucination detection. Accuracy reflects the model's ability to correctly classify hallucinated and non-hallucinated text samples, while RHO (Spearman) assesses the correlation between predicted probabilities and ground truth labels.

A. Performance of Our Own Architecture

In evaluating the efficacy of our novel architecture for hallucination detection, we observed notable improvements in accuracy compared to baseline approaches. Our architecture, featuring a mixed cross-encoder design, comprises three main components: the Hypothesis Encoder, Context Encoder, and Mix Encoder. Through rigorous experimentation, we demonstrated the architecture's ability to enhance accuracy scores, with performance escalating from an initial baseline of 0.5 to a commendable 0.7. RHO Score was around 0.59

The mixed cross-encoder architecture adopts a unique approach, leveraging separate pathways for processing context and target sentences. By extracting essential features from each sentence independently and subsequently integrating them through cross-attention mechanisms, our architecture facilitates effective comparison and contrast between the generated output and the provided context. This design empowers the model to identify semantic inconsistencies and detect instances of hallucination more accurately across various NLG tasks,

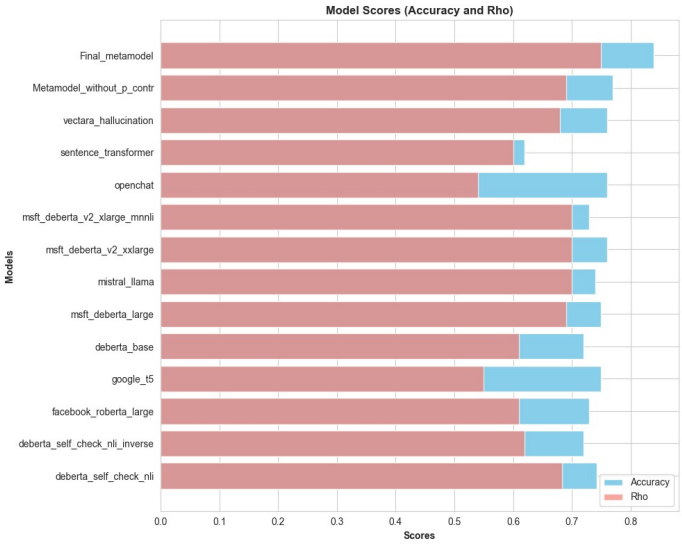

Fig. 4. Comparision of Spearman and Accuracy scores of Base Models

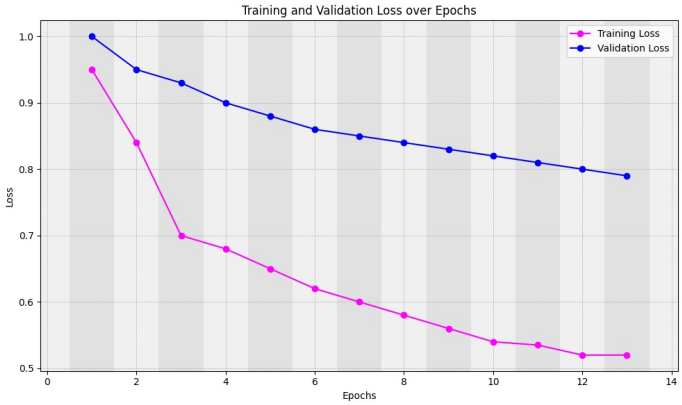

Fig. 5. Loss Curve For Meta Model

demonstrating promising potential for practical deployment in real-world scenarios.

B. Performance of MetaModel Framework

In parallel with our architectural endeavors, we explored the capabilities of the MetaModel framework in enhancing hallucination detection accuracy. Leveraging insights gleaned from a comprehensive set of base models, including RoBERTa, Google T5, Mistral 7b, LLAMA2, and LLAMA3, we developed a MetaModel framework enriched with CatBoost and XGBoost Regressor. Initial assessments revealed suboptimal accuracy scores, primarily attributed to probabilities hovering near the critical switch zone of 50%. To address this challenge, we introduced two probabilities for each sentence sample, assessing both entailment and contradiction, thereby bolstering the predictive capabilities of our models.

Integration of these probabilities into the MetaModel framework yielded a substantial improvement in accuracy, achieving approximately 0.83. Moreover, qualitative assessments using RHO (Spearman) scores demonstrated enhanced reliability and robustness in our model predictions. Visualization techniques, including loss curves *Fig 5* and a bar graph depicting model accuracies and spearman score for each base model *Fig 4*, provided valuable insights into the framework's effectiveness in hallucination detection across diverse NLG tasks.

These results *Fig 4* collectively underscore the promising potential of both our architectural innovation and the MetaModel framework in advancing the state of the art in hallucination detection within NLG systems.

VII. Discussion

The exploration of hallucination detection in Natural Language Generation (NLG) outputs revealed intriguing insights into the efficacy of various methodologies and techniques. Our study encompassed a systematic investigation of baseline approaches, specialized neural network architectures, and MetaModeling techniques, shedding light on the complexities and challenges inherent in this domain.

Baseline approaches served as the foundation for our exploration, providing valuable insights into the performance disparities between task-based and combined architectures. Task-based architectures, tailored to specific NLG tasks such as machine translation (MT), paraphrase generation (PG), and dialogue tasks (DT), demonstrated superior accuracy compared to their combined counterparts. This observation underscores the importance of task-specific customization in capturing semantic inconsistencies effectively.

The introduction of a specialized neural network architecture marked a significant advancement in hallucination detection. Comprising Hypothesis, Context, and Mix Encoders, this architecture enabled separate processing paths for context and target sentences, facilitating more nuanced analysis of semantic coherence. Leveraging cross-attention mechanisms, the model effectively compared the generated text with the input context, enhancing the detection of semantic discrepancies. The integration of residual layers and convolution further improved model robustness, addressing issues such as gradient overflow and enhancing overall performance.

MetaModeling techniques introduced a novel approach to refining predictions and mitigating uncertainty in hallucination detection. By aggregating entailment and contradiction probabilities from multiple base models, the MetaModel enhanced decision-making and improved detection accuracy. The collaborative learning across diverse base models enriched the detection process, enabling more comprehensive identification of semantic inconsistencies.

Entailment and contradiction probabilities serve as key indicators in assessing the semantic coherence and consistency of generated text.

Entailment probability measures the likelihood that the generated text logically follows from the input context or source text. In other words, it evaluates whether the generated text is a valid inference or consequence of the provided information. A high entailment probability suggests that the generated text aligns well with the context and is logically coherent.

On the other hand, contradiction probability evaluates the likelihood of inconsistency or contradiction between the generated text and the input context. It assesses whether the generated text contradicts or conflicts with the information provided in the context. A high contradiction probability indicates that the generated text contains semantic inconsistencies or falsehoods compared to the context.

Despite these advancements, several technical challenges persist in hallucination detection. The inherent ambiguity and subjectivity in assessing semantic coherence pose significant obstacles in designing effective detection methodologies. Additionally, the generalization and scalability of proposed approaches require further investigation and validation across diverse datasets and NLG tasks.

In conclusion, our study represents a significant step forward in advancing the technical capabilities of hallucination detection in NLG outputs. The integration of task-specific architectures, specialized neural network models, and MetaModeling techniques offers promising avenues for enhancing detection accuracy, reliability, and robustness. Continued research and development efforts are crucial to overcoming remaining technical challenges and realizing the full potential of hallucination detection in NLG applications.

VIII. Social Impact and Ethical Considerations

A. Social Impact

Hallucination detection technologies in natural language generation (NLG) systems have the potential to significantly impact various aspects of society. By improving the accuracy and coherence of generated text, these technologies can enhance communication across diverse applications, including machine translation, paraphrase generation, and dialogue systems. This advancement holds promise for facilitating clearer and more effective communication in domains such as cross-cultural interactions, education, and customer service. Moreover, by mitigating the risk of semantic inconsistencies and false information in generated text, hallucination detection technologies contribute to fostering trust and credibility in online communication platforms, thereby bolstering the integrity of digital interactions and promoting informed decision-making.

B. Ethical Considerations

While hallucination detection technologies offer benefits in improving communication accuracy, ethical considerations must be carefully addressed to mitigate potential risks and safeguard societal interests. One key concern pertains to the responsible use of NLG systems to prevent the spread of misinformation and deceptive content. Additionally, ensuring privacy protection and data security in the collection and processing of user data is essential to uphold individuals' rights and maintain trust in NLG technologies. Moreover, efforts to address algorithmic bias and promote fairness in NLG outputs are crucial for preventing discrimination and promoting inclusivity in online communication channels. By adhering to ethical principles of transparency, accountability, and user empowerment, stakeholders can ensure that hallucination detection technologies serve the broader societal good while upholding ethical standards and promoting responsible innovation.

IX. Conclusion

In conclusion, our research has addressed the critical challenge of hallucination detection in natural language generation (NLG) systems. Through a comprehensive exploration of baseline approaches, including task-based and combined architectures, as well as the development of novel methodologies such as the mixed cross-encoder architecture and MetaModel optimization, we have demonstrated promising advancements in identifying semantic inconsistencies in generated text. Our findings underscore the importance of tailored approaches and multi-model integration in enhancing the accuracy and reliability of hallucination detection systems.

Moving forward, our work opens avenues for further research and development in several key areas.

X. Future Work

1) **Enhanced Model Architectures:** Further refinement and optimization of model architectures, such as incorporating additional layers or exploring alternative attention mechanisms, can improve the robustness and effectiveness of hallucination detection systems.
2) **Multi-Modal Integration:** Investigating the integration of multi-modal information, such as incorporating visual or contextual cues alongside textual inputs, could enhance the contextual understanding and accuracy of hallucination detection models.
3) **Ethical Considerations:** Continued exploration of ethical considerations, including privacy protection, bias mitigation, and fairness in NLG outputs, is essential to ensure responsible deployment and use of hallucination detection technologies.
4) **Real-World Applications:** Extending our methodologies to real-world applications and domains, such as social media moderation, content verification, and medical diagnosis, can provide valuable insights into the practical utility and impact of hallucination detection systems.
5) **User-Centric Evaluation:** Conducting user-centric evaluations and usability studies to assess the effectiveness, user satisfaction, and potential biases of hallucination detection systems in real-world settings.

By addressing these areas of future work, we aim to advance the field of hallucination detection in NLG systems and contribute to the development of more reliable, trustworthy, and ethically sound AI technologies.

References

[1] Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation, 2024.

[2] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
[4] Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback, 2024.
[5] Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564, 2023.