SHROOM: Hallucination Detection in NLG Tasks

Aamleen Ahmed(2020002) || Mohammad Sufyan Azam (2020312) || Subhanshu Bansal (2020315) || Siddhant Singh (2020338)

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Abstract

In the domain of Natural Language Generation (NLG), ensuring semantic accuracy alongside fluency is paramount. However, NLG models often suffer from hallucination, where they produce grammatically sound yet semantically incorrect outputs, posing reliability challenges. Our project tackles this issue by developing a robust hallucination detection system.

We employ a blend of innovative methodologies, including custom architectures, fine-tuned pretrained models, and MetaModelling techniques, to effectively identify hallucinatory outputs. Spanning various NLG tasks such as machine translation, paraphrase generation, and definition modeling, our approach offers a comprehensive solution to this pervasive problem, demonstrating its potential to enhance the reliability and trustworthiness of NLG systems.

Baseline + Pretrained Models

Through our baseline experiments, we trained various architectures tailored to individual NLG tasks and their combinations, employing simple models like BiLSTM layers, RNN, and GRU. Surprisingly, machine translation exhibited superior accuracy compared to paraphrase generation and definition modeling. Notably, definition modeling showed higher hallucination rates, likely due to the challenge of defining sentences with single words, leading to contextual loss over time.

Moreover, we observed a potential hindrance in learning when providing both context and target together, as it might lead to loss of crucial features while focusing on lengthy sentences simultaneously. To address this challenge, we introduced our novelty approach.

Model	Task	Metric Score
Custom Attention + biLSTM	MT	0.68
LSTMs + Concatenate	PG	0.64
Attention + LSTM	DM	0.67
biLSTM_4 + positional	Combined 3	0.515
Novelty Architecture	Combined 3	0.6
Novelty Arch + Residual Connections	Combined 3	0.66

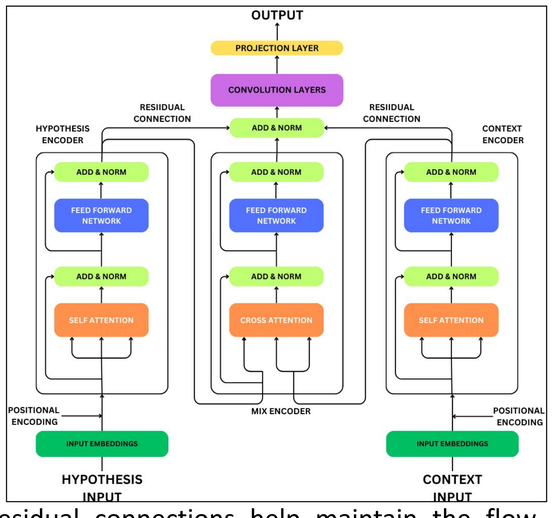
Separate Cross Attentions

The architecture consists of 3 main components:

1. Hypothesis Encoder: This processes the output from the LLM for one of the tasks—MT, PG, or DT. It converts the text into numerical representations, adds positional information, and then refines the representation using attention and feed-forward networks.

2. Context Encoder: This encoder processes the relevant context or source information for the given task. For MT, it would be the source text in the original language; for PG, the text to be paraphrased; and for DT, the dialogue history or the prompt.

3. Mix Encoder: This central encoder integrates the information from both the Hypothesis and Context Encoders. Through multi-head cross-attention mechanisms, it allows the model to compare and contrast the LLM's output with the provided context, looking for discrepancies that might indicate hallucinations—instances where the output is not supported by the input context.

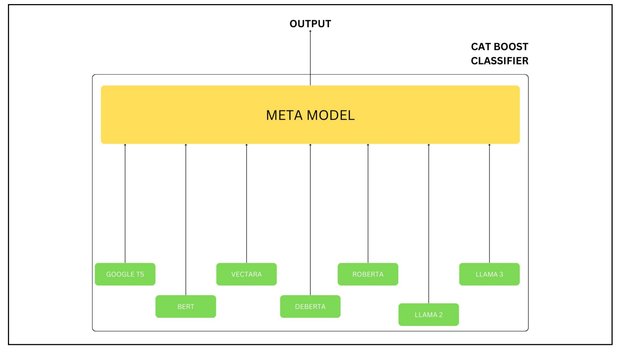


Residual connections help maintain the flow of information and gradients throughout the network, which is crucial for learning. The convolution layers toward the end of the architecture could help in detecting patterns indicative of hallucinations, and the output layer would provide the final assessment of whether the LLM's output is hallucinated or not.

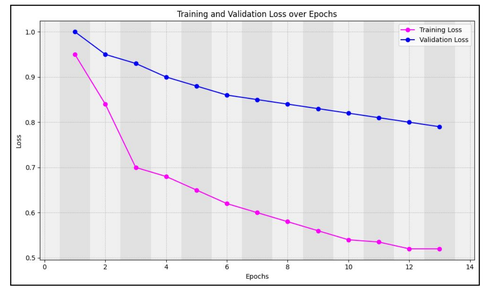
BaseModels + Meta Model

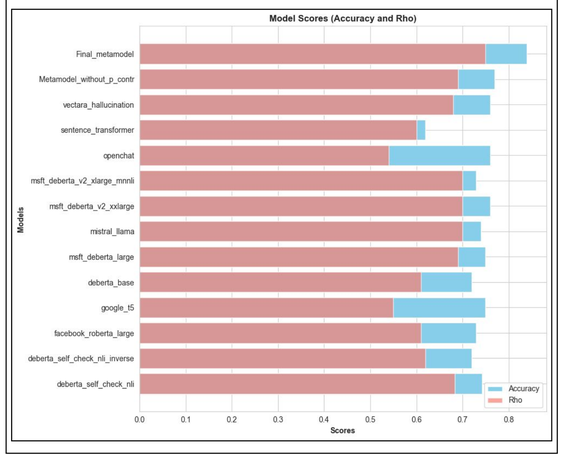
In our preliminary investigations, we evaluated several base models, encompassing prominent architectures like RoBERTa, Google T5, Mistral 7b, LLAMA2, and LLAMA3. Despite their sophistication, these models yielded suboptimal accuracy scores, primarily due to probabilities residing near the critical switch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

In response to this limitation, we devised a series of optimization strategies. Firstly, we introduced two probabilities for each sentence sample, assessing both entailment and contradiction, in a bid to fortify the predictive capabilities of our models. However, the true breakthrough came with the integration of these probabilities into a MetaModel framework. This MetaModel, empowered by CatBoost and XGBoost Regressor, harnessed 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, underscoring the enhanced reliability and robustness of our model predictions.



Results





Conclusion

In conclusion, our project on hallucination detection in NLG outputs has showcased the efficacy of integrated optimization techniques in addressing the challenge of semantic inconsistencies. Through rigorous experimentation and refinement, we improved accuracy scores to approximately 0.83, with notable enhancements in RHO (Spearman) scores, indicating the increased reliability and robustness of our model predictions.

By leveraging diverse base models, including RoBERTa, Google T5, Mistral 7b, LLAMA2, and LLAMA3, we have made significant progress in enhancing the reliability of NLG systems. This work holds significant implications for various practical applications, offering insights into the development of more trustworthy NLG systems.