Faithful Benchmark for Information-Seeking Dialogue (Fact Hallucinations Detection and Prevention)

Balu Harshavardan Koduru
University at Buffalo
Buffalo, New York
baluhars@buffalo.edu

Jaswanth Reddy Angeri
University at Buffalo
Buffalo, New York
jangeri@buffalo.edu

Gayeethri Polamreddy
University at Buffalo
Buffalo, New York
gayeethr@buffalo.edu

Abstract

Chatbots have become an integral part of present modern day life. Inspite of recent breakthroughs in chatbots such as GPT-3, GPT-4, BERT etc., there still exists a major problem which is haunting them - hallucination. Uttering unfaithful, irrelevant, and nonsensical responses deteriorates the performance of the chatbots. In this work, we explored what hallucination is, it's sources and cause, and ways to mitigate hallucinations in a chatbot. We implemented a model which classifies whether an utterance by a chatbot is hallucinated or not and our model is even capable of producing responses which are 82% halluciation free.

1 Introduction

AI chatbots have gained significant popularity in recent years as a means of engaging with customers, providing customer service, and even assisting with mental health care. However, one major challenge that these chatbots face is the issue of hallucination. Sometimes AI chatbots generate some factually incorrect or illogical statements which are not faithful to the information source. This phenomenon is termed as hallucination (Ji et al., 2022). Hallucination can either arise from data or from the training inference. There are several recent citings of hallucinations in large language models such as the GPT3 engine (chatGPT, BingAI).

Hallucination can arise from two sources: data and model inference as shown in figure - 1. Data hallucination occurs when the training data provided to the chatbot is noisy, biased, or flawed, leading the chatbot to generate factually incorrect responses. Model inference hallucination, on the other hand, occurs when the chatbot acquires the trait of hallucination while training due to some fault in the training methodology or the model architecture, which results in hallucinated responses.

The impact of hallucination in AI chatbots can be significant. It can lead to a loss of user trust and satisfaction, as well as potential legal and ethical issues if the chatbot provides incorrect information in fields such as medicine or law. Therefore, understanding and addressing the issue of hallucination is crucial for the successful deployment of AI chatbots.

Recent research has explored various approaches for mitigating the effects of hallucination in AI chatbots. For instance, researchers have proposed techniques for detecting and correcting factual errors in chatbot responses (Rajani et al., 2019).

In this work, our goal is to understand the phenomenon of hallucination and explore several ways to mitigate and avoid hallucinations in the responses of the NLG models.

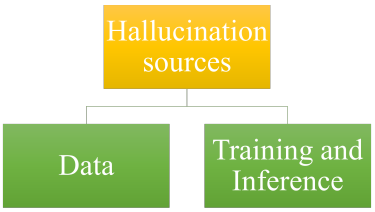

Figure 1: Hallucination Sources

2 Related Work

Several methods were proposed to mitigate hallucinations in the Natural Language Generation (NLG) models. As shown in fig 1, hallucinations can arise from two sources. Based on the origin/source of hallucination in the generated text, the mitigation methods can be classified as - Data related methods and model-related methods (Ji et al., 2022). In this work, we explore a data-driven method that concentrates on preparing a hallucination-free dataset.

2.1 Data Related Methods

Oftentimes, datasets are noisy and this leads to hallucinations in the learned models that were trained on these datasets. One of the ways to reduce hallucinations is to generate a faithful dataset. Different

authors proposed several approaches to building a faithful dataset. In (Zhou et al., 2018), the authors propose a document-grounded dataset that contains conversations that are grounded with respect to the knowledge base which is Wikipedia articles about popular movies in this case. In (Liu et al., 2021), the authors propose two methods to generate faithful texts which will reduce the hallucinations.

2.2 Model Related Methods

Sometimes hallucinations arise irrespective of whether the dataset is faithful or not. This is because of the wrong inference of data by the NLG models. In (Nie et al., 2019), authors proposed an NLG model which reduces hallucinations by establishing meaningful correspondence between the meaning representations and corresponding spoken or written sentences. In (Shuster et al., 2021), authors explored neural retrieval in the loop architectures to mitigate hallucinations. In (De Bruyn et al., 2020), authors explore ways to fine-tune a BART model to obtain faithful and grounded next utterances from the model.

3 Problem Statement

In recent years, dialog systems have become increasingly prevalent in various industries and applications, ranging from customer service chatbots to virtual assistants. While the use of such systems has increased efficiency and convenience, it has also led to an upsurge in the occurrence of hallucinations in the responses generated by these systems.

To address this issue, our project proposes an approach to identify instances where a dialog system provides a response that is not supported by prior knowledge. We leverage a pre-trained transformer-based neural network architecture, specifically the RoBERTa model, to encode the knowledge and response inputs and classify them into hallucination and non-hallucination categories. The model is trained and evaluated on the Faith-Dial dataset (Dziri et al., 2022) consisting of conversational exchanges between a user and a conversational agent. Each exchange is annotated with labels indicating whether a response contains hallucination or not.

In addition to identifying hallucinations, the proposed approach also focuses on generating faithful responses based on prior knowledge that is free of hallucination. By providing a reliable and effective method for detecting hallucinations in conversational agents, our approach can be useful in real-world applications where the trustworthiness and reliability of conversational agents are critical. The performance of the approach is evaluated using various metrics such as accuracy, F1-score, and confusion matrix.

Figure 2: Dataset Pre-processing Pipeline

These are our present tasks - Task 1: Hallucination Critic: Given the conversation history and knowledge, you will have to determine whether the response is hallucinated. Task 2: BEGIN and VRM Multi-Class Multi-label Classification: Given the conversation history and knowledge, you will have to identify the speech acts (VRM taxonomy such as disclosure, edification, question, acknowledgment, etc.) and the response attribution classes (BEGIN taxonomy) such as hallucination, entailment, etc. Task 3: Dialogue Generation: Given the conversation history and knowledge, you will have to generate a response that is faithful to the conversation history and knowledge.

4 Dataset

FAITHDIAL is a benchmark dataset for knowledge-grounded dialogue, consisting of 50,761 turns that are spread across 5,649 conversations. Through extensive human validation, it has been revealed that 94.4% of the utterances in FAITHDIAL are faithful. This dataset provides supervision for hallucination critics, which are used to determine whether an utterance is faithful or not.

The FAITHDIAL dataset includes useful metadata annotations, such as information about the topic of conversation, the gender of the speakers, and the context in which the conversation took place. Additionally, it is annotated with dialogue acts, which label the communicative function of each turn in the conversation. These annotations provide a comprehensive and nuanced understanding of the conversations in the dataset.

Overall, FAITHDIAL is a valuable resource for research in knowledge-grounded dialogue, offering

a robust and validated dataset with metadata and dialogue act annotations.

Statistics of FAITHDIAL dataset: 6200 total talks, 18 distinct domains, 348 unique question templates, Average conversation duration of 8 turns Average of 2.5 questions per conversation, 2.8 sentences per turn on average. Features of the dataset are summarized in the table 1.

As shown in figure - 2, the responses were first all converted into lower case. Then all the stop words and special characters were removed from the dataset and the responses were then tokenized using BERT-based uncased.

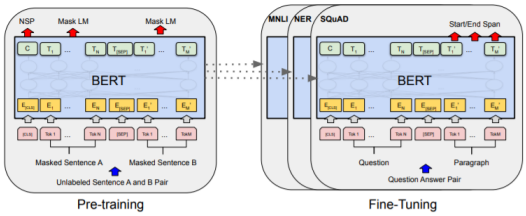

Figure 3: BERT - Network Architecture (Sharma, 2022)

5 Model Architecture

5.1 Milestone - 2 Model

In this project, for baseline-2, we have built our model using RoBERTa large version (Liu et al., 2019), a pre-trained Masked Language (MLM) Model. RoBERTa, which stands for Robustly Optimized BERT-Pretraining Approach is a modified version of BERT with better performance. RoBERTa is an improved implementation version of BERT in terms of the size of the training corpus and a more efficient form of training. RoBERTa alters BERT's pre-training by removing the Next Sentence Prediction (NSP) task and adding dynamic masking, which causes the masked token to change throughout the training epochs.

Figure 4: Intermediate Model end-to-end Pipeline

The model architecture of RoBERTa is similar to that of the BERT - multi-layer bidirectional transformer encoder. The BERT model architecture is depicted in figure - 3. The main differences between BERT and RoBERTa are summarized in the table - 2.

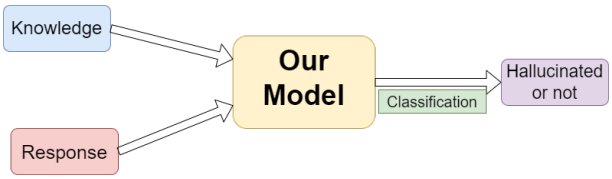

Figure 5: Task-1 model outline

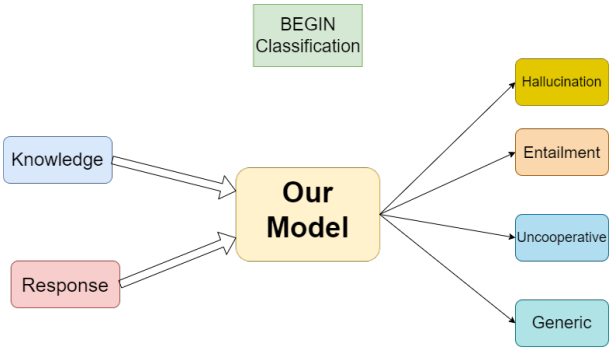

Figure 6: Task-2 BEGIN Classification model outline

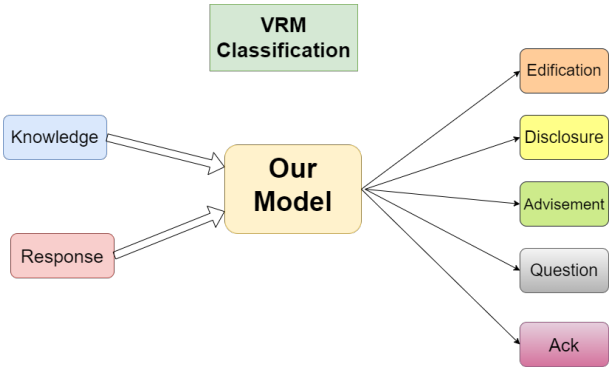

Figure 7: Task-2 VRM Classification model outline

5.2 Intermediate Model

For this model, extra layers were added to the existing RoBERTa model as shown in figure - 4. Extra Layers added in the model are -

- Mulit-Head Attention layer
- Layer Normalization
- Dense layer
- Desne Layer (Output Layer)

The above layers were added repeatedly, 4 times, after the RoBERTa model layers.

Features	Description
History - List(string)	The dialogue history
Knowledge (string)	The source knowledge on which the bot wizard should ground its response
Speaker (string)	The current speaker
Original response (string)	The WoW original response before editing it
Response (string)	The new Wizard response
BEGIN - List (string)	The BEGIN labels for the Wizard response
VRM - List (string)	The VRM labels for the wizard response

Table 1: Features present in FaithDIAL Dataset (Dziri et al., 2022)

	BERT	RoBERTa
Size	Base:110 Million; Large:340 Million	Base:110 Million; Large:340 Million
Training Time	64 TPU Chips x 4 days (280 x V100 x 1 day)	1024 x V100 x 1 day (5 times longer than BERT)
Performance	Out-performed the state-of-the-art (Oct, 2018)	Upto 20% improvement over BERT
Training Data	16 GB Data Corpus	16 GB BERT Data + 144 Additional GB
Architecture	BERT (Bi-directional Transformer with MLM and Next Sentence Prediction (NSP)	BERT without NSP

Table 2: Differences between BERT and RoBERTa (Suleiman Khan, 2021)

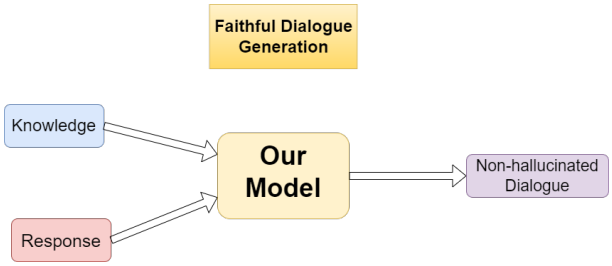

Figure 8: Task-3 model outline

For task-3, we used GPT-neo (125 million) to generate the responses (hallucinated and non-hallucinated).

5.3 Final Model Architecture

For the final model, for tasks 1 and 2, we modified the internal layers of the RoBERTa-large model. We added a linear scheduler and epsilon for the Adam optimizer. We modified the weight decay of model layers to 0.1 except for bias, position embedding, embedding, and normalization layers. For task - 3, we used the same model as the intermediate model.

6 Results

6.1 Task - 1

Accuracy and F1-Score for this task are shown in the table - 3. The confusion matrix for the Final model is shown in figure - 9.

Figure 9: Task-1 Final model BEGIN Classification Confusion Matrix

6.2 Task-2

Accuracy and F1-Score for all the models is shown in figure - 4 and 5
The confusion matrix for BEGIN Classification of the final model is shown in the table - 10

The confusion matrix for VRM Classification of the final model is shown in figure - 11

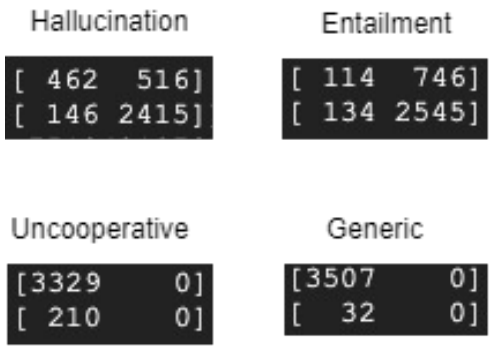

Figure 10: Task-2 Final model BEGIN Classification Confusion Matrix

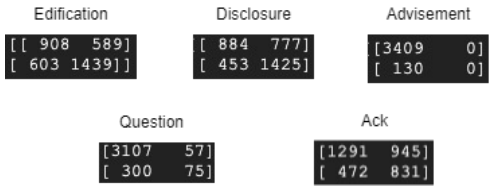

Figure 11: Task-2 Final model VRM Classification Confusion Matrix

6.3 Task -3

BLEU, ROGUE, and BERT score are shown in the table 6 for all the models. The percentage of utterances identified as unfaithful is 18%.

7 Contribution

My contribution to this project is summarized in the table 7.

8 Discussion and Error Analysis

8.1 Milestone - 2: Baseline Model

Based on the current results of the model, it is evident that the model is not efficient in predicting responses accurately, with most of the responses being classified as 'Hallucination' by the critic model. The dialog generation model is also generating responses that are not faithful to the knowledge, and in Task 2, the model is predicting 'Hallucination' for the 'BEGIN' class and 'Disclosure' for the 'VRM' class. However, we can make incremental changes to different parameters of the model and conduct experiments to achieve better results. As we observed with each epoch there were no significant changes in accuracy or f1 score. The current model is insufficient in evolving with given learning rate. We observed that the predicted class in most cases is defined by the most prevalent class in the training data. We need to ensure this do not happen.

To begin with, we plan to change the pre-trained model used for tokenization. Currently, we are using Roberta-base, but we could experiment with models such as Bert, GPT, Albert, or Roberta-Large, as tokenization is a critical aspect of NLP models. By changing the model early in the process, we may be able to obtain better results.

Another way to improve the model's performance is to use a larger model with more layers, which can improve the training. We can also add extra layers to the training model to make it more robust, including neural network layers such as Batch Normalization, Pooling, and Dropout, which can enhance the model's efficiency. Furthermore, we can replace the existing model used for response creation with a better one to produce more accurate responses. We should also adjust the hyperparameters when implementing these changes, such as the learning rate, batch size, number of epochs, activation function, and dropout rate.

In addition, we can experiment with regularization techniques and try using other optimizers such as SGD or Adamax. However, we believe that changing the hyperparameters of the existing optimizer would yield better results than replacing it with another optimizer.

Lastly, we can implement early stopping to ensure that the model reaches its peak accuracy without overfitting. By increasing the number of epochs and stopping the training process before the model begins to overfit, we can obtain a more accurate model.

8.2 Milestone - 3: Final Model

8.2.1 Task-1

In Task-1, we have implemented various techniques such as scheduler, weight-decay, and warm-up to enhance the performance of the roBERTa model. These methods have helped in improving the parameters of each layer, resulting in better model performance. From the confusion matrix, we can infer that the model is producing satisfactory results. However, the prediction of the entailment label can be further improved. On the other hand, the model's ability to predict hallucination labels is quite impressive. One possible reason for this could be that the training data contains more hallucination labels than entailment labels, indicating a possible bias towards hallucination.

	Accuracy	F1-Score
Baseline Model	72.82%	0.42
Intermediate Model	78.14%	0.57
Final Model	83.73%	0.65

Table 3: Task - 1 Evaluation Results

	Accuracy	F1-Score
Baseline Model	72.82%	0.42
Intermediate Model	74.35%	0.38
Final Model	87.39%	0.43

Table 4: Task - 2 BEGIN Classification Evaluation Results

	Accuracy	F1-Score
Baseline Model	72.82%	0.42
Intermediate Model	41.7%	0.24
Final Model	44.82%	0.3

Table 5: Task - 2 VRM Classification Evaluation Results

	BLEU Score	ROGUE	BERT Score
Baseline Model	0.08314	0.19	0.53
Intermediate Model	0.09	0.22	0.56
Final Model	0.09	0.22	0.56

Table 6: Task - 3 Evaluation Results

Milestone-1	Milestone-2	Milestone-3
Familiarizing and exploring various evaluation metrics and Linguistic features and tools to improve the performance of the model	Data set loading, pre-processing, input id, and label creation, evaluation 163 metrics	Loading Train and test datasets, Preprocessing the data to generate input ids, attention masks, and labels, Implemented loss function, and worked on critic model modification

Table 7: My Contribution

To improve the model further, we can experiment with techniques such as POS tagging and varying the padding length of input ids. Additionally, we can also try using different optimizers and varying hyperparameters such as learning rate and epsilon to enhance the model's performance. Although we used the Adam optimizer in our approach, there could be other optimizers that could yield better results.

In summary, the roBERTa model's performance has improved significantly with the implementation of various techniques. However, there is still room for improvement, and by experimenting with different techniques and hyperparameters, we can enhance the model's performance and reduce possible biases.

8.2.2 Task-2

The classification model for BEGIN had a bias towards hallucination and entailment labels due to the limited number of instances for the uncooperative and generic class labels. This disparity in the number of examples for each class resulted in the model performing better on hallucination and entailment labels compared to the remaining two classes. To improve the accuracy of this task, it is essential to increase the number of training examples and ensure an equal number of instances for each class in the training data to avoid such bias. The same issue applies to the Advisement and Question labels of VRM classification, where the limited number of training examples for these labels resulted in reduced classification accuracy.

For task-1, we used categorical cross-entropy as there was only one class for classification. However, in task-2, we used multi-class multi-label binary cross-entropy with logit loss. It is possible that experimenting with different loss functions may further improve the model performance.

In summary, to address the bias in classification models, we need to ensure a sufficient number of instances for each class in the training data. Additionally, experimenting with different loss functions may further improve the model performance in classification tasks.

8.2.3 Task-3

In our text generation task, we observed that 82% of the generated text was non-hallucinated. This is an encouraging result considering that we used a critic generated from task-1, which had a slight bias towards predicting responses as hallucinations. The high percentage of non-hallucinated responses indicates that the generator is performing well.

During the experiment, we faced memory exhaustion issues, and as a workaround, we used the GPT-neo version with 125 million parameters. However, we believe that using the same model with more parameters could potentially improve the text generation performance even further.

In summary, our text generation experiment yielded promising results, with the majority of the generated text being non-hallucinated. Additionally, using a more powerful model with more parameters could further enhance the text generation performance.

9 Conclusion

Our work focuses on identifying and reducing hallucination in the utterances of NLG models. In essence, our project is an essential contribution to the ongoing endeavors of improving the performance and dependability of conversational agents, especially with respect to hallucination detection and response generation. This can have a profound impact on various industries and applications. The use of trustworthy and reliable conversational agents can lead to increased efficiency, productivity, and customer satisfaction, among other benefits. We seek to address a significant and currently unresolved problem in dialog systems by utilizing cutting-edge natural language processing methods and a pre-trained neural network architecture.

10 Contribution

My contribution to this project is summarized in the table - 7.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[1] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

References

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. Bart for knowledge grounded conversations.

[1] https://www.aclweb.org/portal/content/acl-code-ethics

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13415–13423.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning.

Drishti Sharma. 2022. A gentle introduction to roberta.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

Ph.D. Suleiman Khan. 2021. Bert, roberta, distilbert, xlnet-which one to use?

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.