Faithful Benchmark for Information-Seeking Dialogue

Jagath Sai Narayana K
University At Buffalo
Buffalo, NY, USA
jagathsa@buffalo.edu

Siddhi Satish Jadhav
University At Buffalo
Buffalo, NY, USA
sjadhav3@buffalo.edu

Briyana Rana
University At Buffalo
Buffalo, NY, USA
brana@buffalo.edu

Abstract

As the demand for information-seeking dialogue systems increases, ensuring the accuracy and reliability of responses to user queries becomes crucial. When response utterances are not supported by underlying truth or ground knowledge, it can lead to negative consequences such as loss of trust in the system and a poor user experience. In domains like healthcare, finance, and education, the dissemination of false or misleading information can have severe implications. To address this issue, we propose a novel approach that focuses on generating knowledge-grounded responses based on conversation history and ground truth. By using this approach as a training signal, we aim to enhance the performance of existing dialogue systems. We introduce a hallucination critic that discerns between faithful and unfaithful responses, and we identify different Verbal Response Modes (VRM) to gain a comprehensive understanding of speech acts in dialogue. Additionally, we train a model to generate response utterances that align with the underlying knowledge sources, ensuring accuracy and reliability in information-seeking dialogue systems.

1 Introduction

Conversations lacking sufficient knowledge pose a significant challenge for information-seeking dialogue systems, which strive to deliver precise and reliable information to seekers. When responses are not firmly rooted in solid knowledge, seekers may receive incomplete or inaccurate information, leading to an erosion of trust in the system. Despite notable advancements in natural language conversation agents, the persistent issue of 'Hallucination' remains. Hallucination refers to the phenomenon where language models generate responses that seem plausible but lack a foundation in factual information or contextual understanding.

An illustrative example in Figure 1 showcases a dialogue between an open-source agent and an information seeker, where the agent mistakenly provides an incorrect response to a query about the largest state in the US, Texas instead of the correct answer, Alaska. In order to tackle this challenge, our project adopts a systematic three-step approach. Firstly, we develop a Hallucination Critic, designed to discern the presence of hallucinated responses. Secondly, we establish a taxonomy of Verbal Response Modes (VRM), enabling us to gain a comprehensive understanding of the diverse speech acts occurring within dialogues. Lastly, we train a language model to generate responses that are firmly grounded in the underlying knowledge.

To accomplish this, we leverage the FaithDial dataset, which has been acquired from Wizards of Wikipedia on the HuggingFace platform. This dataset encompasses the complete conversation history, associated knowledge, original responses from the WoW agent, and meticulously annotated gold response data generated by human experts. By employing these invaluable resources and employing advanced techniques, our objective is to significantly enhance the accuracy and dependability of information provided by dialogue systems.

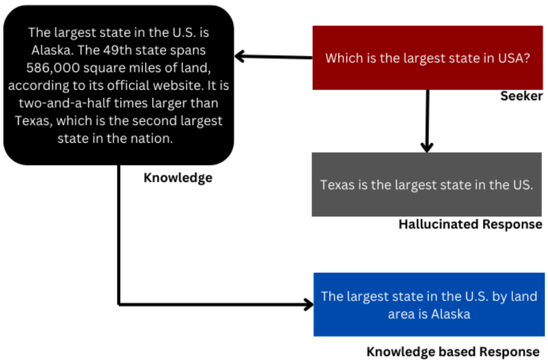

The largest state in the U.S. is Alaska. The 49th state spans 586,000 square miles of land, according to its official website. It is two-and-a-half times larger than Texas, which is the second largest state in the nation.
Knowledge
Which is the largest state in USA?
Seeker
Texas is the largest state in the US.
Hallucinated Response
The largest state in the U.S. by land area is Alaska
Knowledge based Response

Figure 1: High-level representation on the text generation

2 Related work

The models being used in the current approach are BERT classifier and T5 text generation model. Although text generation models have shown impressive capabilities in generating text but, these models often suffer from hallucination, producing outputs that contain fabricated information. Following are the latest research and techniques aimed at addressing hallucination in text generation models as well as classification tasks.

"FAITHDIAL: A Faithful Benchmark for Information-Seeking Dialogue" by Dziriy et al. introduces a comprehensive benchmark dataset designed specifically for information-seeking dialogue systems. The authors highlight the issue of hallucination in existing benchmarks, which refers to the generation of unsupported utterances by dialogue systems. The main objective of this proposed dataset is to establish a benchmark that is free from hallucination, ensuring that all responses provided by the dialogue systems are firmly grounded on reliable knowledge sources.

"BERT for Multi-label Text Classification" by Sun et al. (2019): The paper introduced a hierarchical attention network that utilizes BERT embeddings to capture contextual information and hierarchical dependencies within the document. The model achieved competitive results on benchmark multi-label classification data sets, showcasing the effectiveness of BERT in this context.

"Plug and Play Language Models: A Simple Approach to Controlled Text Generation" by Dathathri et al. (2020): The paper focused on controllable text generation using T5. They proposed a straightforward yet effective approach to modify T5's behavior by conditioning it on prompts and few-shot demonstrations. This technique empowered users to generate text with specific attributes or biases, allowing for greater control over the output.

"Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" by Raffel et al. (2019): The paper introduced T5, a unified framework that achieved remarkable results across various NLP tasks, including text generation. The authors demonstrated the versatility of T5 by fine-tuning it on tasks such as summarizing, translation, and question-answering, achieving state-of-the-art performance.

3 Data Description

Faith-Dial is a benchmark data set specifically designed to provide hallucination-free or knowledge-grounded responses. Table 1 provides a description of the Faith Dial benchmark data set, including how the responses have been augmented to meet these criteria.

4 Model Architecture

The model is divided into three phases. The first phase is multiclass classification, where the task is to determine if the given text falls into the categories of "GENERIC," "UNCOOPERATIVE," "HALLUCINATION," or "ENTAILMENT" based on conversation history and knowledge. The second phase is multi label classification, which involves identifying if the given text belongs to the categories of "NONE," "ADVISEMENT," "QUESTION," "ACK," "EDIFICATION," or "DISCLOSURE" based on conversation history and knowledge. The third phase of the project focuses on text generation using a transformer model. Prior to performing all the phases, the data pre-processing step involves converting the list of history into a single string and then concatenating the newly created history with the knowledge.

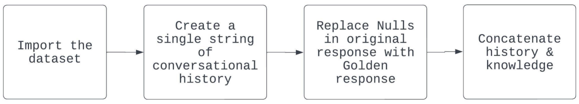

Import the dataset
Create a single string of conversational history
Replace Nulls in original response with Golden response
Concatenate history & knowledge

Figure 2: Data Pre-processing

4.1 Hallucination Critic

As part of our approach, we build a Hallucination Critic to discriminate between faithful and unfaithful responses generated by dialogue systems. This task aims to improve response quality by filtering out unsupported responses and enhance the overall accuracy and reliability of the system. Once trained, the critic evaluates new dialogue system outputs, providing feedback to improve response quality. To achieve this we employ a BERT Multi-label classifier. This involves splitting the list of classes into different columns, encoding the tokens with special padding, and converting the data into tensors. We then encapsulate the entire dataset into a dataloader and pass it through the neural network. The model consists of the forward feed neural network, training, testing, and validation

Dataset Label	Description
Dialog idx	Id of the conversation
History	The conversation history
Knowledge	Ground truth based on which the response must be generated
Original Response	The response given by the agent without the knowledge
Response	The response given by the agent the knowledge
BEGIN	Begin labels for the wizard response
VRM	VRM labels for the wizard response

Table 1: FaithDial dataset and description

steps. These steps take the input IDs, attention mask, and labels as inputs and compute the loss using the Binary Cross Entropy function. The model uses the Adam optimizer to update its parameters and minimize the loss while maintaining the learning rates. The model is then trained for 2 epochs, and the best checkpoint is stored. The checkpoint is later loaded, and predictions are generated on the testing set. Table 2 and Figure 3 shows the count of classes, as well as a flowchart.

Classes	Train	Test	Val
GENERIC	116	32	23
UNCOOPERATIVE	979	210	226
HALLUCINATION	13507	2561	2525
ENTAILMENT	15374	2679	2604

Table 2: Begin Label count

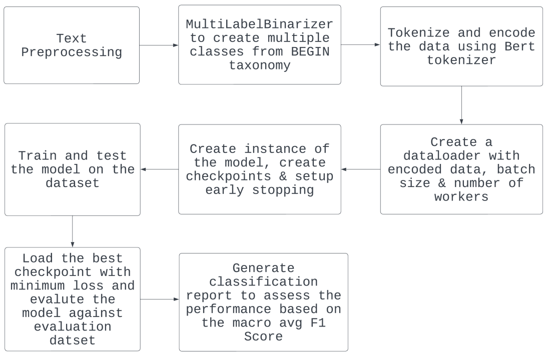

Text Preprocessing
MultiLabelBinarizer to create multiple classes from BEGIN taxonomy
Tokenize and encode the data using Bert tokenizer
Create a dataloader with encoded data, batch size & number of workers
Create instance of the model, create checkpoints & setup early stopping
Train and test the model on the dataset
Load the best checkpoint with minimum loss and evalute the model against evaluation datset
Generate classification report to assess the performance based on the macro avg F1 Score

Figure 3: BEGIN classification Architecture

4.2 Multi-Class Multi-Label Classification

In the next phase of our approach, we focus on classifying the Verbal Response Modes (VRM) categories. By categorizing speech acts using VRM, we gain valuable insights into communication dynamics that can have diverse applications. Verbal response modes encompass the different ways speakers respond to verbal stimuli, such as questions, statements, or requests, and can be organized into distinct categories based on their function and structure. To achieve this we employ a BERT Multi-label classifier. This involves splitting the list of classes into different columns, encoding the tokens with special padding, and converting the data into tensors. We then encapsulate the entire dataset into a dataloader and pass it through the neural network. The model consists of the forward feed neural network, training, testing, and validation steps. These steps take the input IDs, attention mask, and labels as inputs and compute the loss using the Binary Cross Entropy function. The model uses the Adam optimizer to update its parameters and minimize the loss while maintaining the learning rates. The model is then trained for 2 epochs, and the best checkpoint is stored. The checkpoint is later loaded, and predictions are generated on the testing set.

Classes	Train	Test	Val
NONE	55	12	7
ADVISEMENT	565	1300	103
QUESTION	2070	375	343
ACKNOWLEDGE	7381	1303	1436
EDIFICATION	10239	1878	1903
DISCLOSURE	10587	2042	1928

Table 3: VRM Label count

4.3 Text Generation

The third phase is the actual task of generating hallucination-free text. To ensure the generation of coherent text data, we utilize the T5 transformer-based model. T5 is a versatile sequence-to-sequence model capable of performing text generation and text-to-text tasks, such as translation, summarization, and question-answering. Functioning as a fully conditional language model, T5 is trained to generate output sequences based on input

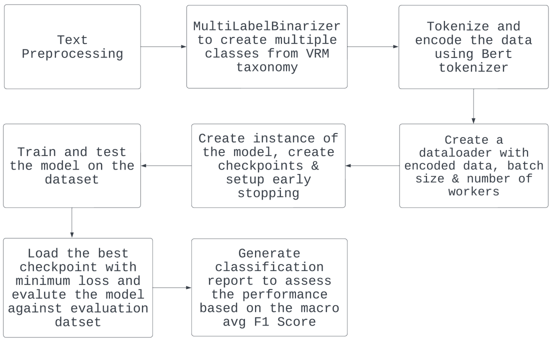

Text Preprocessing
MultiLabelBinarizer to create multiple classes from VRM taxonomy
Tokenize and encode the data using Bert tokenizer
Create a dataloader with encoded data, batch size & number of workers
Create instance of the model, create checkpoints & setup early stopping
Train and test the model on the dataset
Load the best checkpoint with minimum loss and evalute the model against evaluation datset
Generate classification report to assess the performance based on the macro avg F1 Score

Figure 4: VRM classification Architecture

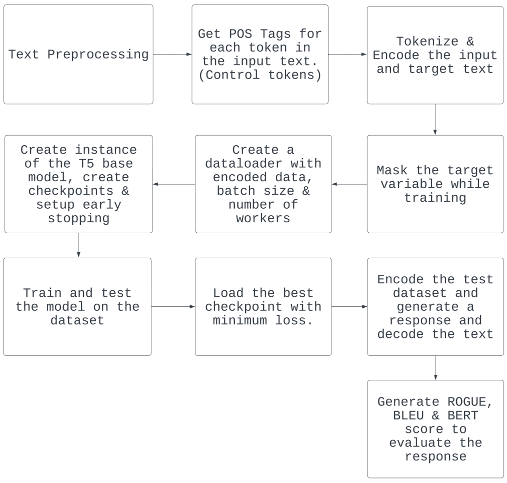

Text Preprocessing
Get POS Tags for each token in the input text. (Control tokens)
Tokenize & Encode the input and target text
Mask the target variable while training
Create a dataloader with encoded data, batch size & number of workers
Create instance of the T5 base model, create checkpoints & setup early stopping
Train and test the model on the dataset
Load the best checkpoint with minimum loss.
Encode the test dataset and generate a response and decode the text
Generate ROGUE, BLEU & BERT score to evaluate the response

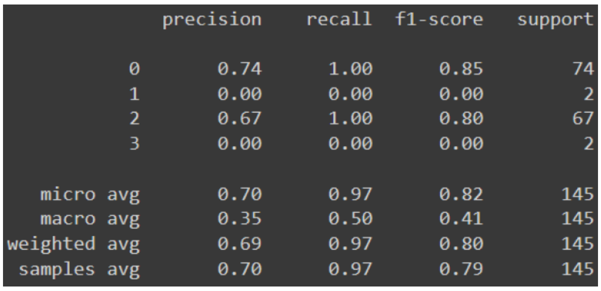
Figure 5: Text Generation Model Architecture

sequences and control tokens. This characteristic makes it well-suited for text generatiom, where the model must generate appropriate responses. In the text generation task, we focus on three essential fields: conversation history, associated knowledge, and augmented or edited responses. To preprocess the data, we concatenate the conversation history and associated knowledge. Additionally, we extract the parts-of-speech tags from this concatenated string. Tokenization is then performed on the concatenated string, incorporating the parts-of-speech tags as control tokens in the source encoding. Target encoding is achieved by tokenizing the gold responses. The labels are derived from the target encoding, with padding tokens replaced by -100. After tokenizing the source and target, we create a data loader to handle the loading and processing of data for input to the T5 model, creating training, validation, and test sets. The T5 model is initialized with logging and early stopping enabled. It is then fine-tuned on the pre-trained T5 base model using the training set and validated using the validation set. The best model is saved, and its performance is evaluated on unseen test data. Finally, given the provided knowledge and conversation history, we aim to predict the model's generated response.

5 Results

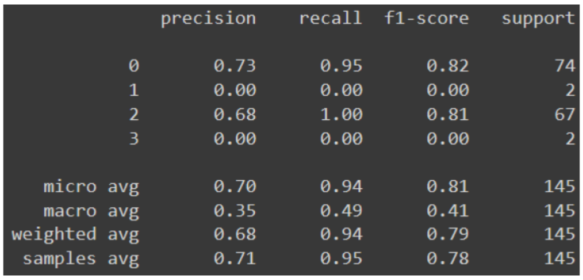
5.1 Hallucination Critic

We initially implemented text classification using SVM, Multinomial Naive Bayes, and BERT for our baseline model. However, for the final architecture, we chose to proceed with BERT due to its ability to capture complex relationships and dependencies among words, which is advantageous for multi-label classification tasks. Furthermore, upon examining the data set, we observed a significant frequency imbalance between the 'Hallucination' and 'Entailment' labels compared to 'Edification' and 'Disclosure'. To address this issue, we utilized BERT's pre-training and fine-tuning approach, enabling the model to learn from the imbalanced distribution and adjust label probabilities, resulting in improved handling of imbalanced data. The classification report shows macro-average F1 scores of 0.42, 0.42, and 0.43 for SVM, MultinomialNB, and BERT, respectively, for the BEGIN taxonomy. However, the macro-average F1 score fails to consider the data set's class imbalance, treating all classes equally regardless of their size. In such cases, the F1 micro-average score is more appropriate as it provides a suitable metric for imbalanced data sets, where the performance on larger classes has a greater impact on the overall evaluation. Upon observing the micro-average F1 scores, we find that SVM, MultinomialNB, and BERT achieve scores of 0.81, 0.81, and 0.83, respectively. Clearly, BERT demonstrates the best overall performance among the models evaluated.



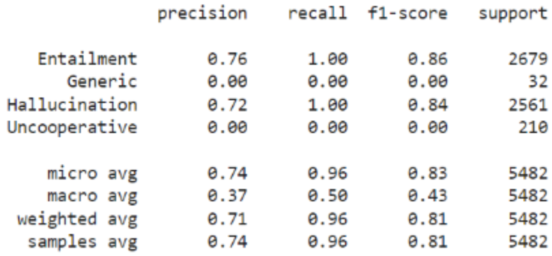
```
              precision    recall  f1-score   support

           0       0.74      1.00      0.85        74
           1       0.00      0.00      0.00         2
           2       0.67      1.00      0.80        67
           3       0.00      0.00      0.00         2

   micro avg       0.70      0.97      0.82       145
   macro avg       0.35      0.50      0.41       145
weighted avg       0.69      0.97      0.80       145
 samples avg       0.70      0.97      0.79       145
```

Figure 6: Begin classification using SVM



```
              precision    recall  f1-score   support

           0       0.73      0.95      0.82        74
           1       0.00      0.00      0.00         2
           2       0.68      1.00      0.81        67
           3       0.00      0.00      0.00         2

   micro avg       0.70      0.94      0.81       145
   macro avg       0.35      0.49      0.41       145
weighted avg       0.68      0.94      0.79       145
 samples avg       0.71      0.95      0.78       145
```

Figure 7: Begin classification using MultinomialNB

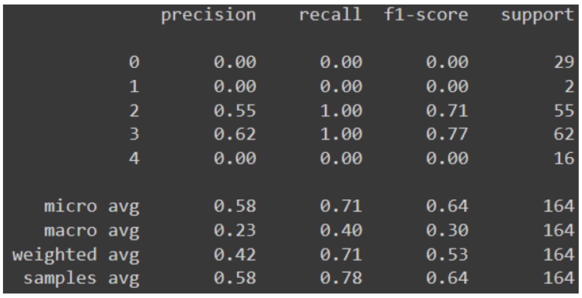


```
               precision    recall  f1-score   support

   Entailment       0.76      1.00      0.86      2679
      Generic       0.00      0.00      0.00        32
Hallucination       0.72      1.00      0.84      2561
Uncooperative       0.00      0.00      0.00       210

    micro avg       0.74      0.96      0.83      5482
    macro avg       0.37      0.50      0.43      5482
 weighted avg       0.71      0.96      0.81      5482
  samples avg       0.74      0.96      0.81      5482
```

Figure 8: Begin classification using BERT

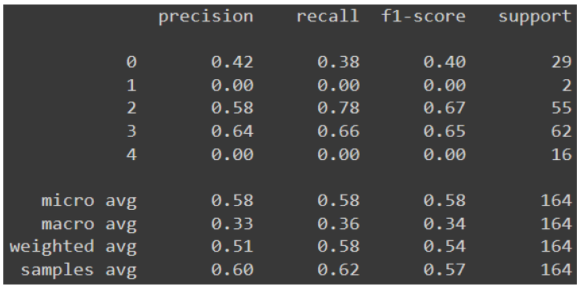
5.2 Multi-Label and Multi-class Classification

We initially implemented text classification using SVM, Multinomial Naive Bayes, and BERT for our baseline model. However, for the final architecture, we chose to proceed with BERT due to its ability to capture complex relationships and dependencies among words, which is advantageous for multi-label classification tasks. Furthermore, upon examining the data set, we observed a significant frequency imbalance between the 'DISCLOSURE' and 'EDIFICATION' labels compared to 'ACKNOWLEDGEMENT', 'QUESTION', 'ADVISEMENT', and 'NONE'. To address this issue, we utilized BERT's pre-training and fine-tuning approach, enabling the model to learn from the imbalanced distribution and adjust label probabilities, resulting in improved handling of imbalanced data. The classification report shows macro-average F1 scores of 0.30, 0.34, and 0.33 for SVM, MultinomialNB, and BERT, respectively, for the VRM taxonomy. However, the macro-average F1 score fails to consider the data set's class imbalance, treating all classes equally regardless of their size. In such cases, the F1 micro-average score is more appropriate as it provides a suitable metric for imbalanced data sets, where the performance on larger classes has a greater impact on the overall evaluation. Upon observing the micro-average F1 scores, we find that SVM, MultinomialNB, and BERT achieve scores of 0.64, 0.58, and 0.60, respectively. Although SVM generates better results than BERT and MultinomialNB, considering the architectural benefits of BERT we choose to move forward with the same.



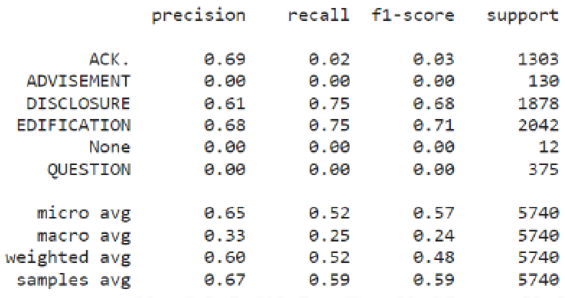
```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        29
           1       0.00      0.00      0.00         2
           2       0.55      1.00      0.71        55
           3       0.62      1.00      0.77        62
           4       0.00      0.00      0.00        16

   micro avg       0.58      0.71      0.64       164
   macro avg       0.23      0.40      0.30       164
weighted avg       0.42      0.71      0.53       164
 samples avg       0.58      0.78      0.64       164
```

Figure 9: VRM classification using SVM



```
              precision    recall  f1-score   support

           0       0.42      0.38      0.40        29
           1       0.00      0.00      0.00         2
           2       0.58      0.78      0.67        55
           3       0.64      0.66      0.65        62
           4       0.00      0.00      0.00        16

   micro avg       0.58      0.58      0.58       164
   macro avg       0.33      0.36      0.34       164
weighted avg       0.51      0.58      0.54       164
 samples avg       0.60      0.62      0.57       164
```

Figure 10: VRM classification using MultinomialNB



```
              precision    recall  f1-score   support

        ACK.       0.69      0.02      0.03      1303
  ADVISEMENT       0.00      0.00      0.00       130
  DISCLOSURE       0.61      0.75      0.68      1878
 EDIFICATION       0.68      0.75      0.71      2042
        None       0.00      0.00      0.00        12
    QUESTION       0.00      0.00      0.00       375

   micro avg       0.65      0.52      0.57      5740
   macro avg       0.33      0.25      0.24      5740
weighted avg       0.60      0.52      0.48      5740
 samples avg       0.67      0.59      0.59      5740
```

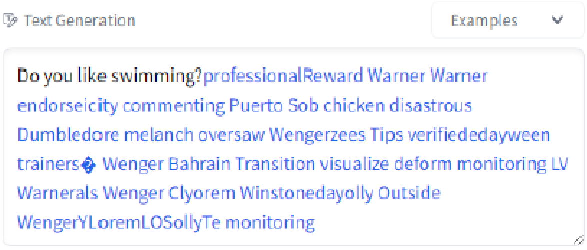
Figure 11: VRM classification using BERT

5.3 Text Generation

Initially, for the text generation aspect of our approach, we constructed the model architecture using the GPT-2 transformer model. GPT-2 is a language model based on a transformer architecture, designed to understand natural language patterns and structure. However, despite its power, the transformer architecture has limitations and can struggle with certain inputs. GPT-2 generates text by predicting the probability distribution of the next word based on the preceding words, but it may encounter issues such as repetitive patterns and nonsensical output due to overfitting on the training data. When we applied the GPT-2 model for text generation, we encountered nonsensical and repetitive output, as illustrated in Figure 11. To address this problem, we opted for a more powerful transformer model, T5, for the final implementation. T5 proved to be superior in generating coherent sentences and avoiding hallucinations. To evaluate the model, we employed metrics such as ROUGE, BLEU, and BERT scores. We tested the model on an unseen test dataset, as shown in Figure 13, and obtained an average ROUGE score of 0.58, a better BLEU score of 0.38, and a high BERT f1 score of 0.89 when comparing the generated sentences to the gold responses. The scores for the entire test dataset can be found in Table 4. Additionally, we assessed the model's performance by manually providing seeker utterances and associated knowledge, as depicted in Figure 14. The model exhibited satisfactory performance in generating responses without hallucinations.

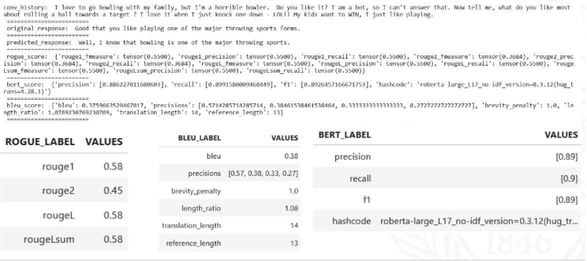
ROUGE1	BLEU	BERT Score
0.40	0.07	0.90

Table 4: Evaluation scores for the entire test dataset



Do you like swimming?professionalReward Warner Warner endorseicity commenting Puerto Sob chicken disastrous Dumbledore melanch oversaw Wengerzees Tips verifiededayween trainers� Wenger Bahrain Transition visualize deform monitoring LV Warnerals Wenger Clyorem Winstonedayolly Outside WengerYLoremLOSollyTe monitoring

Figure 12: Text Generated using GPT-2



```
conv_history:  I love to go bowling with my family, but I'm a horrible bowler.  Do you like it? I am a bot, so I can't answer that. Now tell me, what do you like most
about rolling a ball towards a target ? I love it when I just knock one down - LOL!! My kids want to WTN, I just like playing.
 ========================
 original_response:  Good that you like playing one of the major throwing sports forms.
 ========================
 predicted_response:  Well, I know that bowling is one of the major throwing sports.
 ========================
 rogue_score:  {'rouge1_fmeasure': tensor(0.5500), 'rouge1_precision': tensor(0.5500), 'rouge1_recall': tensor(0.5500), 'rouge2_fmeasure': tensor(0.3684), 'rouge2_prec
ision': tensor(0.3684), 'rouge2_recall': tensor(0.3684), 'rougeL_fmeasure': tensor(0.5500), 'rougeL_precision': tensor(0.5500), 'rougeL_recall': tensor(0.5500), 'rouge
Lsum_fmeasure': tensor(0.5500), 'rougeLsum_precision': tensor(0.5500), 'rougeLsum_recall': tensor(0.5500)}
 ========================
 bert_score:  {'precision': [0.886227011680603], 'recall': [0.8991580009460449], 'f1': [0.8926457166671753], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_t
rans=4.28.1)'}
 ========================
 bleu_score:  {'bleu': 0.3759663529467017, 'precisions': [0.5714285714285714, 0.38461538461538464, 0.3333333333333333, 0.2727272727272727], 'brevity_penalty': 1.0, 'le
ngth_ratio': 1.0769230769230769, 'translation_length': 14, 'reference_length': 13}
 ========================
```

ROGUE_LABEL	VALUES
rouge1	0.58
rouge2	0.45
rougeL	0.58
rougeLsum	0.58

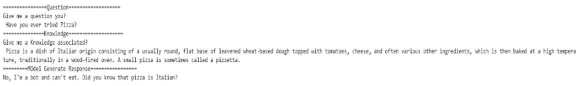
BLEU_LABEL	VALUES
bleu	0.38
precisions	[0.57, 0.38, 0.33, 0.27]
brevity_penalty	1.0
length_ratio	1.08
translation_length	14
reference_length	13

BERT_LABEL	VALUES
precision	[0.89]
recall	[0.9]
f1	[0.89]
hashcode	roberta-large_L17_no-idf_version=0.3.12(hug_tr...

Figure 13: Text generation and Evaluation using T5

6 Discussion and Error Analysis

From the baseline results, we were able to observe that we got low macro-average f1 scores due to highly imbalanced data set classes. To handle this, we can follow some other approach like data augmentation where we can generate some new examples for the underrepresented dataset or data up-sampling/down-sampling to have all the classes of the same distribution or use ensemble methods such as bagging or boosting that handle imbalanced data to improve performance, etc. Since the current dataset heavily relies on the knowledge provided by the user, we can use information retrieval methods to extract the knowledge data from Wikipedia based on the queries asked by the seeker. Besides that, we can utilize active learning techniques to improve the efficiency of dataset annotation and model training. We can also deploy the model as a conversational agent and collect user feedback to further improve the quality and naturalness of the generated responses.



```
----------------Question-------------------
Give me a question you?
 Have you ever tried Pizza?
---------------Knowledge--------------------
Give me a Knowledge associated?
 Pizza is a dish of Italian origin consisting of a usually round, flat base of leavened wheat-based dough topped with tomatoes, cheese, and often various other ingredients, which is then baked at a high tempera
ture, traditionally in a wood-fired oven. A small pizza is sometimes called a pizzetta.
---------MOdel Generate Response-----------------
No, I'm a bot and can't eat. Did you know that pizza is Italian?
```

Figure 14: Manually text input

7 Conclusion

We have successfully accomplished each task and generated hallucination-free responses based on the provided knowledge. However, due to the imbalance in label counts, we noticed significantly low scores, indicating an area that requires improvement. It is important to note that the accuracy and correctness of the generated text depend heavily on the quality and relevance of the knowledge being provided. Therefore, it is crucial to ensure that the knowledge is retrieved from reliable sources and kept up to date.

8 Contribution

Developed data pre-processing step involving creating a single string of history, replacing nulls, and concatenation of the history and knowledge. Implementation of TF-IDF, Multinomial Naive Bayes, and creating a pipeline for the second milestone. Implemented Bert classifier for BEGIN and VRM taxonomy. Converting Bert classifier into Pytorch Lighting and fine-tuning it. Implemented POS tagging as a control token in text generation.

References

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion An-

Task	Contribution%
Hallucination Critic	70%
Multi-label & class classifier	70%
Text generation	20%

Table 5: Contributions Table

droutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.