Creating a Realistic, Airline Chatbot

Abstract

Current customer service chatbots have three main issues: a lack of personalization, repetitive responses, and a lack of empathy/understanding for the customer.[1] Firstly, there's a big lack of personalization in their responses as chatbots tend to offer very generic, unhelpful answers, failing to recognize and adapt to the unique circumstances and needs of a customer's concerns. Secondly, a significant challenge with chatbots is due to their common dependence on keyword recognition instead of more intelligent techniques, leading them to miss the nuances of human language and interpret different requests/queries as the same. Finally, tying all these together is the issue that chatbots have a large lack of empathy. When customers face personal issues, many chatbots respond in tones inappropriate for the corresponding situation, which ignites severe anger, especially when they are stressed. All of these as a combination leave customers feeling undervalued and more frustrated from when they started, which is especially a huge issue in the airline industry, prone to having stressed customers with time-sensitive concerns. **The goal of our project is to create a chatbot catered for customer service by fine-tuning DialoGPT on various human-to-human customer service conversation datasets to better mimic humans and add empathy in responses specifically for professional use** i.e the airline industry, and behaves as close to how a human representative would, providing appropriate responses and calming down the customer according to the situation at hand. We evaluate our results across multiple models we create from different types of customer service datasets from 3 perspectives: similarity to sample response (measured by BLEU and METEOR), length of response, and human judgment through qualitative review. **Our results indicate significant improvement in producing responses that are more human-like** (i.e. from 0 to 30 BLEU score improvement) and show empathy in our fine-tuned models, compared to the original DialoGPT model. These conclusions were carefully made through extensive hyperparameter tuning and cross-validating our many variations of fine-tuned models on different customer service datasets. We also explored some non zero-shot prompt-engineering techniques, but due to DialoGPT already being fine-tuned from GPT-2, we found our efforts to produce less significant results compared to fine-tuning on new datasets.

Introduction

According to Zendesk's 2022 CX Trends report, around 60 percent of customers see disappointment when dealing with chatbots, [2] and close to 75 Billion Dollars in potential business revenue is lost for reasons tied to poor customer service.[3] These statistics confirmed our motivation to pursue this high-potential project to create a conversational chatbot that talks as close to humans as possible, reducing potential anger and rage brought on by existing customer service chatbots that mainly just use keyword-matching. For this project, we are using the dialoGPT[4] model from HuggingFace, which is trained

on 147M conversation-like text from sources such as Reddit, and is optimized from the existing GPT-2 model to cater more towards conversational texts and generating human responses. GPT-2 is already known for its proficiency in generating human-like responses, and we will fine-tune this using more specialized datasets tailored to customer support in the aviation industry to further enhance the conversational abilities of the model and provide a more efficient interaction experience for angry passengers seeking assistance. **Creating this "more human and empathetic" customer service chatbot will be of huge significance due to its direct impact on customer satisfaction rates which are a driving force behind many airlines' revenue.** This goal has driven the rationale behind our methodologies and evaluation techniques, **putting a heavy emphasis on ensuring the output texts are as human-like as possible** (measured quantitatively through BLEU and METEOR) and through qualitative inspection. Our objective is to experiment with various fine-tuned versions of the DialoGPT model and prompt engineering on customer service datasets, and coming up with the most optimal model that successfully achieves our evaluation metrics and criterias outlined above.

Literature Review

We test on 6K multi-ref dataset from Reddit. The results are summarized in below

Experiment	NIST2	NIST4	BLEU2	BLEU4	METEOR	ENT-4	DIST-1	DIST-2	Avg. Len
Human response	3.41	4.25	17.90%	7.48%	10.64%	11	14.50%	63.00%	13.1
DialoGPT 117M	2.39	2.41	10.54%	1.55%	7.53%	10.78	8.60%	39.90%	12.8
DialoGPT 345M	3	3.06	16.96%	4.56%	9.81%	9.13	6.80%	26.30%	12.2
DialoGPT 762M	2.84	2.9	18.66%	5.25%	9.66%	9.72	7.76%	29.93%	11.2

Figure 1: DialoGPT Performance on 6K Reddit Dataset[5]

Relevance: A and B, which one is more relevant to the source prompt.

System A	A Wins (%)	Ties (%)	B Wins (%)	System B
DialoGPT 345M	2671 (45%)	513 (9%)	2816 (47%)	Human responses
DialoGPT 345M	3281 (72%)	394 (9%)	882 (19%)	PersonalityChat
DialoGPT 345M w/ MMI	**2871** (48%)	522 (9%)	2607 (43%)	Human responses

Informativeness: A and B, which one is more contentful and informative.

System A	A Wins (%)	Ties (%)	B Wins (%)	System B
DialoGPT 345M	2722 (45%)	234 (4%)	3044 (51%)	Human responses
DialoGPT 345M	3490 (77%)	206 (5%)	861 (19%)	PersonalityChat
DialoGPT 345M w/ MMI	**3011** (50%)	234 (4%)	2755 (46%)	Human responses

Figure 2: DialoGPT Performance Evaluated by Humans[6]

In this section, we will first provide an in-depth review about DialoGPT, a widely used and peer-reviewed model that we use as our baseline in this paper. Then, we will examine common industry practices and recent breakthroughs in both fine-tuning and prompt engineering techniques to tailor industry-standard LLMs for specific applications. DialoGPT has been trained on 147 Million response-pairs in Reddit discussions and as

documented on the official DialoGPT implementation[7], its results have been validated through both similarity metrics such as BLEU and METEOR as can be seen in Figure 1, and by testing on humans to compare human and DialoGPT generated responses to prompts as can be seen in Figure 2. In an in-depth competition organized by Microsoft, the DialoGPT responses have been proven to be selected by humans as being more relevant and informative than human responses, on average. Next, we will look into some papers detailing hyperparameter tuning methods. The first paper *Rethinking Learning Rate Tuning in Large Language Models*[8] discusses how in LLMs, there are so many parameters involved in training that often fewer epochs and different evaluation techniques can actually help model performance. We tried to adopt the thought processes in this analysis as was evident with our best performance in non-overfitting with only 5 epochs in training. Additionally, the utilization of BLEU and METEOR metrics in evaluating a response definitely helped us as is described in a lot more detail under the hyperparameter tuning section below.

Methodology

In our investigation, we experimented with various customer service datasets to fine-tune DialoGPT and evaluated their respective responses using similarity scores (BLEU, METEOR) to human response test data, put them in context by analyzing their length, and conducting a sanity check through qualitative inspection of sample responses. Below is a high level summary of the datasets, training methods, hyperparameter tuning, and evaluation techniques used to achieve our results described later in this paper.

Datasets Used for Fine-Tuning (Reference name in Quotations and description after colon):

- "24K": 24k rows organized in pairs of customer demands/questions and successful human operator responses[9]
- "Non-Augmented": Small Dataset of Action-seeking Customer Questions/Requests and Corresponding Responses[10]
- "Augmented": Manually Revised Version of the Non-Augmented dataset with additional customer queries and more empathetic human responses[11]

Training and Hyperparameter Tuning:

- For fine-tuning models on each dataset, we used an Adam Optimizer with Cross-Entropy Loss on 5 epochs with a learning rate of 5e-5 for "Augmented Data" and "Non-Augmented Data" (these values were decided after hyperparameter tuning on validation set and further detail provided in "Hyperparameter Tuning and Experiments" section) and 3 epochs for "24K" dataset.
- For our big "24K" dataset, we used a subset of the data (10%) to minimize GPU Cost.
- Hyperparameter Tuning on variables like learning rate, epochs, and dropout to find the optimal hyperparameters for our model to evaluate on test set and comparison against other models (Further Detail in "Hyperparameter Tuning and Experiments" section).

- We also experimented using prompt engineering techniques on native DialoGPT and on our 3 fine-tuned models. X-shot methods failed to work largely due to the fact that DialoGPT itself is fine-tuned already, and meta-prompting led to confusing or systematic error responses such as "I don't understand what you mean." The BLEU score was also consistently lower (highest achieved was 10 on 24K dataset) than without using prompt engineering, and therefore we did not include models including prompt engineering in our final evaluation.

Models Used for Final Evaluation:

- Native/Original DialoGPT
- "24K" fine-tuned model with 3 epochs and learning rate = 5e-5
- "Augmented" dataset fine-tuned model with 5 epochs and learning rate = 5e-5
- "Non-Augmented" dataset fine-tuned model with 5 epochs and learning rate = 5e-5

Evaluation Methodology:

- We analyzed the 4 models' performance on the test sets of the "24K" (referred to as "long_customer_dataset.csv") and the "Augmented" Dataset (referred to as "test.csv") as they were the most accurate and comprehensive customer support datasets on the internet, and by cross-testing we can have a comprehensive understanding of the strengths and weaknesses of each model.

Evaluation Metrics:

- *Similarity Score*: BLEU and METEOR values comparing the model-generated response to a query against the response data on the dataset. BLEU and METEOR are both industry-standard evaluation techniques that show n-gram precision for BLEU and both recall and precision in the case of METEOR. These scores have been proven to benchmark the level of "humanness" and similarity of two texts.
- *Length of Output*: Some questions warrant longer answers and some for shorter responses, and taking the average output length is also important to put the BLEU and METEOR scores in context as they are influenced by them.
- *Qualitative Inspection:* Randomly sampling answers by each model gives us great insight into the type of English language used and how accurate/actionable the responses are to customer queries.

Hyperparameter Tuning and Results

For our data, we decided to perform hyperparameter tuning to find the best performance of a fine-tuned model in a multitude of ways: via the number of epochs used to train, the learning rates, and the models we used. To evaluate the efficacy of the models, we used a combination of three main API calls. The BLEU, bilingual evaluation understudy, evaluates two sentences and measures the match between them. A few limitations of this scoring are the fact that it looks for more matches, and does not necessarily look at grammar or sentence flow. METEOR, like BLEU, is also used for evaluating machine translation. It addresses some of BLEU's shortcomings by considering both precision and

recall and by incorporating additional linguistic features. Which is ultimately why we chose to use both these scores in evaluation not only for sentence match quality from what is expected but also for general human-like responses. Lastly, we used average response length because nobody wants to use a chatbot that either outputs too long or too short of a response.

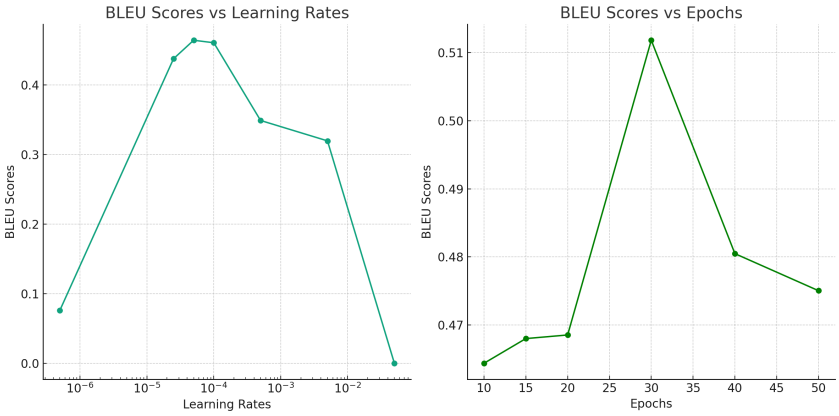

Figure 3: Initial Hyperparameter Exploration using Training Set

In the figure above, we see the results of our experiments from our initial model as a function of epochs and learning rates as a function of the average BLEU scores for the training data. To clarify, we used training data as this is not the hyperparameter tuning, but it was an experimentation to see the BLEU scores over different learning rates and epochs to see what may or may not have potential to yield the best scoring answers. We see above that the best scores are with 30 epochs, but this may be due to overfitting in the data since the test.csv is composed of similar-type data to that of the Customer-Conversation-Augmented dataset. To calculate this, we used the model that performed the best on the test.csv validation data, which was the augmented customer support data, and tweaked the hyperparameters. As visible, we found a natural parabolic-esque curve that peaked when the learning rate of the Adam optimizer was 5e-5. Lastly, we tweaked the models we used: the base DialoGPT model, DialoGPT trained on a Customer-Support dataset, DialoGPT trained on an augmented version of the Customer-Support dataset, and the DialoGPT model trained on the long_customer_support dataset with the largest amount of data. This was a pure experimentation to see which model type trained on what data had the best performance scores.

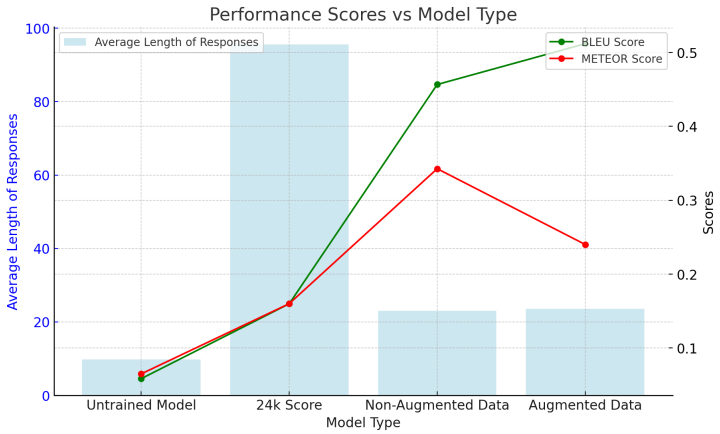

Figure 4: Performance Scores for Evaluation on the Augmented Test Set

Above, in Figure 4, we have a graph of performance scores vs the model type, using the BLEU and METEOR score as metrics as well as an average length of responses. This was all tested on the test.csv dataset. Below, we followed the same process except with a version of the long-customer-support dataset as the validation set to see how scores would be affected.

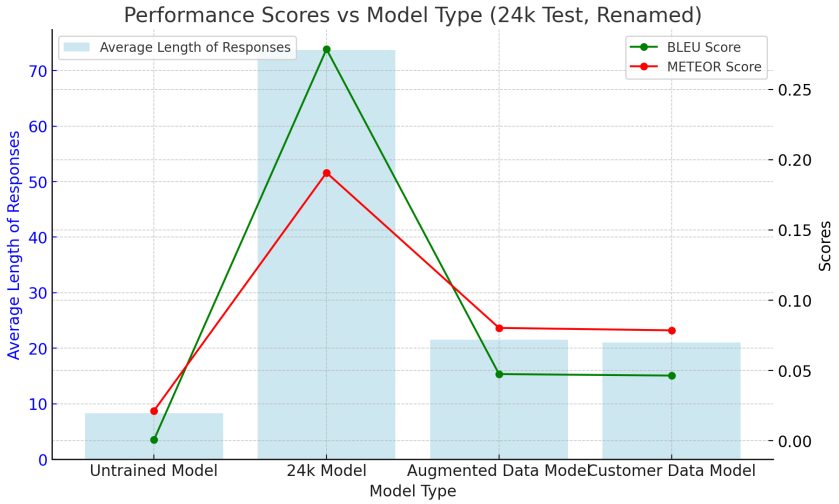

Figure 5: Performance Scores for Evaluation on the 24K Test Set

As visible above in Figure 5, the magnitude of scores significantly drops below the counterpart in the test.csv dataset. This is most-likely due to the fact that the validation data in test.csv is much more human-like, where the validation data in the long-customer-support is more robotic in their answers, following a very structured approach to answering queries. The 24k model, which is the long-customer-support data trained model, performs best in this case since it is full of more data following the structured approach to answering queries. However, this was a very interesting experiment to take in order to see the journey of our models and how our early model and later models compared. The earliest model would be the Customer-Data un-augmented and the latest model is the augmented customer data model along with the 24k model.
For the next portion of analysis, let's look into losses for different hyperparameters. As seen in Figures 6 and 7, we have validation and training losses with respect to different learning rates. As emphasized above, the best performance is with a learning rate of 5e-5 with 5 epochs to avoid overfitting. This is further exemplified with the regular training and validation loss curves (for a singular learning rate) and validation loss over different lengths of epochs. We see how the training loss as seen in Figure 7 strictly improves as more epochs are completed, but validation loss in Figure 6 is minimized at the 5 epoch number and with a learning rate, LR, of 5e-5. This overall tells us that 5 epochs of training on the Customer-Conversation-Augmented dataset yield the best results with a learning rate of 5e-5.

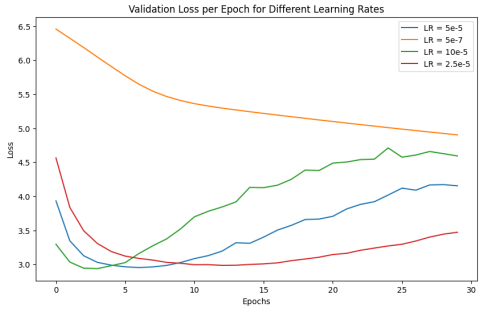

Figure 6: Validation Loss Per Epoch for Different Learning Rates

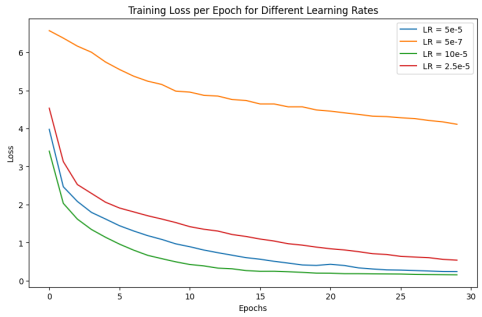

Figure 7: Training Loss Per Epoch for Different Learning Rates

Summary of Results

Tested on the 24K test set:

Model	BLEU	METEOR	Avg. Word Length
DialoGPT	0.0007	0.0214	10.4
”24K” Model	0.278	0.1907	82.8
”Augmented”	0.047	0.0803	26.4
”Non-Augmented”	0.045	0.0785	25.2

Tested on the ”test.csv”, the Augmented test set:

Model	BLEU	METEOR	Avg. Word Length
DialoGPT	0.0586	0.065	17.6
”24K” Model	0.1599	0.1601	155.5
”Augmented”	0.5118	0.2401	38.2
”Non-Augmented”	0.4566	0.3427	39.6

Qualitative Review

DialoGPT: Many of the results for the ”24K” test set were the same error messages, as seen in Figure 9 such as ”I’m not sure what you mean by this,” due to the very specific nature of the queries asking for their account numbers, flight details, etc. This shows a lack of adaptability to specific prompts and how it has been trained to give a generic answer when it is not sure what the prompt is asking for.

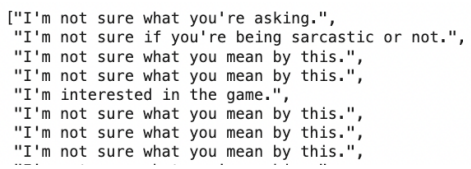

Figure 8: Sample Output from Native DialoGPT Model tested on 24k Dataset

"24K": The nature of the "24K" dataset consisted of very lengthy responses that go in-depth and provide instructions on what the user should do. This explains its average word length being much larger than other models, but as it can be seen from the below example, it provides very specific and accurate information to prompts. (This example asked "I want to unsubscribe from your newsletter. What do I do?") This was also given as a response from a query in the "Augmented" dataset, which is notably impressive given that it is a completely new dataset that it has never been trained on before.

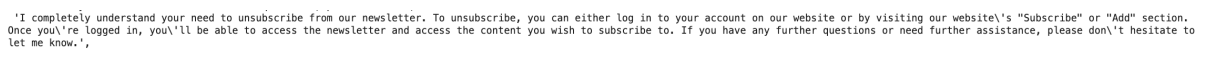

Figure 9: Sample Output from 24k Model Tested on Augmented Dataset

"Augmented" and "Non-Augmented": These models provided very similar outputs due to the nature of their training sets being very similar, and had many short instructions that were vague but addressing the issue at hand. One response, for example, was "I do not have the information at this moment but will let my boss know. What is your phone number?"

Discussion of Results and Next Steps

Overall, we can see significant improvement across all metrics from Dialo GPT to all other fine-tuned models, even when cross-testing with other datasets. We can attribute this to the success in fine-tuning our new models to cater to customer complaints and to address them in a way that shows empathy and understanding. Furthermore, we can claim that the "24K" model was the best performing, as despite its length (usually a negative contributor to BLEU score), it still performed at an acceptable level on the "Augmented" dataset having a BLEU score of close to 0.2 or 20%. This compares to the score DialoGPT got when used against traditional chats as shown in the "Literature Review" section and was enough for it to be regarded as an industry-standard model. We can further explore ways to improve our "24K" model by getting more computing units and further conducting hyperparameter tuning.

Appendix

Notes

[1]https://blog.bcwebwise.com/excessive-use-of-chatbots-can-lead-to-customer-dissatisfaction/
[2]https://www.cxtoday.com/speech-analytics/customers-frustrated-with-chatbots/
[3]https://hyken.com/customer-experience/75-billion-dollars-is-lost-due-poor-customer-service/
[4]https://huggingface.co/docs/transformers/model_doc/dialogpt
[5]https://huggingface.co/docs/transformers/model_doc/dialogpt
[6]https://huggingface.co/docs/transformers/model_doc/dialogpt

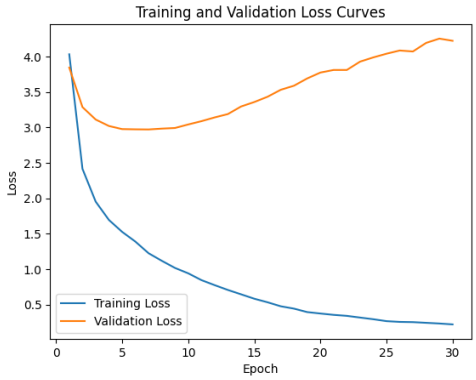

Figure 10: Training Loss Per Epoch for Different Learning Rates

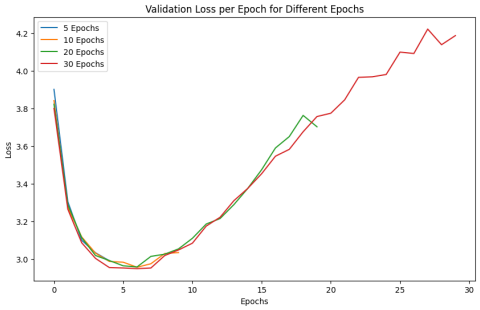

Figure 11: Validation Loss Per Epoch for Model Trained at Different Epochs

[7]https://github.com/microsoft/DialoGPT
[8]https://ar5iv.labs.arxiv.org/html/2309.08859
[9]https://huggingface.co/datasets/bitext/Bitext-customer-support-llm-chatbot-training-dataset
[10]https://huggingface.co/datasets/Kaludi/Customer-Support-Responses
[11]https://drive.google.com/file/d/1N8flgoy6TYdtgxgvBVtnXeDRMkWz-iYu/view?usp=sharing

GitHub Link: https://github.com/ShaamerKumar/cs182FinalProj