Overall, most of our peer reviews focused on the fact that we did not have a working model with results to be able to give substantial feedback on improvements but liked our direction and methodology overall. In this new iteration, we believe we addressed all the issues brought up as we have a very comprehensive methodology and substantial results to discuss. These are all included in our updated research paper that also includes many plots that support our claims made throughout the paper.

In black is an excerpt or summary of the 4 reviewers comments and in blue is our response.

1. What is the main goal of the project?
Develop an empathetic, personalized airline chatbot that closely mimics human customer service representatives. This chatbot aims to address common issues with current chatbots: llack of personalization, misunderstandings, and lack of empathy.

The main goal of this project is to build a chatbot that adds empathy in responses and “acts like a human would” for professional use. They cite the airline industry as an example in the goal statement, but it does seem like the main focus of the paper.

Response:

No response necessary, the goal was highlighted perfectly

What are the main claims?

Claims that by fine-tuning the DialoGPT model with a specialized dataset for airline customer service, the chatbot will significantly improve in handling customer queries with empathy and specificity, thus reducing customer frustration.

They claim that dependence on keyword recognition is the main issue. They claim that fine-tuning a chatbot on specialized datasets tailored to customer support in the aviation industry using a custom loss function that penalizes responses by their similarity to examples of “bot-like” responses that angered customers will solve the issue of empathy

Response:

No response, claims were touched on well.

What are the experiments?

Fine Tune the DialoGPT model with datasets specific to airline customer service, using custom loss functions to penalize bot-like responses and promote human-like interactions.

No experiments have been run or clearly defined.

Response:

Yes, we fine tune the DialoGPT model with different datasets, over different learning rates, and epochs to find the best results. We then compared the differently trained models with different NLP evaluation scores like BLEU and METEOR to find the best one

What is the evaluation protocol?

Cross-entropy loss and similarity metrics like BLEU and METEOR to measure the quality of the chatbot's responses, ensuring they align closely with empathetic, human responses.

There is no clear evaluation protocol listed; they cite a custom loss function using common evaluation metrics in NLP. The closest thing to an evaluation metric seems to be “customer anger”

Response:

Yes, BLEU and METEOR are two commonly used LLM metrics to align LLM responses to be more human-like

What is the data?

The data includes the Airline Customer Service Transaction Log Dataset and a Generic Customer Service Question and Answer Dataset, alongside human-labeled datasets for validating the chatbot's responses.

There is a query-response HuggingFace database titled “Customer Support Responses”, and another titled “3K Conversations Dataset for ChatBot” that seems to consist of conversation query-response pairs. The description lists casual or formal discussions, interviews, customer service interactions, or social media conversations. However, it only seems to contain casual conversations.

Response:

Yes! We have 3 main datasets. These are general customer-conversations, one large dataset with customer conversations that is very structured in its responses, and an augmented dataset with more airline-specific human conversation examples

What is the task?

The task is to fine-tune the DialoGPT model to generate responses that are formal, respectful, and tailored to handling airline customer complaints and queries.

The task is to generate respectful and formal text for airline-type issues.

Response:

Yes, thank you

How do the experiments support the goal/claims of the paper?

The experiments directly train the model with customer complaints and queries. It tries to improve the model's ability to handle complex and specific customer issues in the airline industry.

No experiments have been conducted, so it is hard to tell.

Response:

Yes thank you!

Are any of the limitations discussed in the paper?

The paper discusses limitations in data reformatting and training with large datasets, and the current lack of relevant results for revising methodology.

Response:

Yes, we go into detail discussing the difficulties with data formatting and training with large datasets as it was difficult to first find the data and then have it fit within our collab GPUs since LLMs are typically trained with lots of space, time, and data.

What are the strengths of the paper?

The strength of the paper lies in its detailed approach to customizing the chatbot for a specific industry need and its focus on mitigating customer frustration with empathetic responses.

They seem to understand the qualitative drawbacks of the current state of chatbots, and have found a conversational model that can serve as a quality base for fine-tuning. Given lack of detail in methodology and lack of results, it's hard to identify any other strengths.

Response:

Thank you! We try to tackle a challenging yet specific industry issue for our project to make an impact

What are the weaknesses of the paper?

The weakness of the paper is the general lack of experimentation or results. It needs significant work before being ready to submit.

It's difficult to pinpoint exactly what their custom loss is. The majority of the body focuses on custom loss functions, which all seem to rely on non-differentiable evaluation metrics, as most of these use discrete n-gram matching. In the TODOs, where they use cross-entropy as a base for their custom loss. If they go with the former, the model will not learn. The goal also seems very difficult to evaluate; customer service interactions, especially in high-stake industries like airlines, are inherently high-stress and prone to angry customers. In addition, there may not be much distinction between professional and chatbot responses in empathy, as anything beyond a variant of "I'm so sorry" might be seen as "unprofessional". Finally, there needs to be more specificity in implementation details like the fine-tuning methods.

Response:

Thank you for the feedback. This has been heavily tackled and adjusted. We have added lots of charts and data of our experimentations and results. Including hyperparameter tuning and the scoring of BLEU and METEOR results on the validation data for each of the different models

Provide a suggestion for improving the paper.

Start on your implementation and provide details. If you don't have implementation right now, it would be fine to just show the control. Benchmark the current model using the loss.

Response:

Thank you! We definitely have a working implementation now with all plots and charts clearly laid out.

What is the relevant related work?

The base model of Diablo from Microsoft. No other related work is mentioned.

Response:

Yes ! We have the base model from diablo, however, we also cited another paper *Rethinking Learning Rate Tuning in Large Language Models* which describes our results and how we chose to think about our tuning and training issues.

Is the paper reproducible?

Installing some libraries is reproducible, but there are no results or experiments in the code so there is nothing to reproduce of that sort.

Response:

Thank you! The paper is now reproducible by following the code uploaded to the github after downloading the corresponding datasets and making the right library imports.

Can you rerun the experiments?

The code containing the plot is missing, so I'm guessing that is from somewhere else?

Response:

All code is now clearly labeled and placed on the github, allowing (with a little uncommenting) to reproduce all charts and plots as well as being able to rerun experiments. For one of the exploration charts with BLEU scores vs learning rates and epochs, the charts won't be available on github along with the final charts since we retrieved the data by running the code with different parameters and used a third party software to plot the data. However, the charts are

still reproducible if you similarly extract the relevant data and plot it yourself. The main reason we didnt put these charts on github were due to the fact that we didn't know how to plot it and we wanted to have the consolidated information nicely laid out on one graph i.e the BLEU/METEOR/avg length scores with respect to each differently trained model version of DialoGPT.

Can you reproduce the results in the paper?

Given the current github, no.

Response:

As of now the results should be reproducible with the updated code by following the github and rerunning cells under different hyperparameters

Are all the plots in the paper clearly interpretable with well-defined and explained axes, with methodology clearly explained in the paper text?

The plot is well-defined and explained, but its methodology on how it was obtained is unclear.

Response:

All plots in the paper are clearly labeled with their axis and methodology are clearly laid out.

Is the English in the paper correct and clear?

Not really. In the first line, Are there three main issues or four?

English is clear and concise!

Response:

This issue has been fixed amongst others, thank you

Do you have any feedback on any TODOs that the authors have left at this stage? *

Prioritize resolving the training issues with the initial model and using your custom loss fucntion. Limit your scope a lot and maybe just finetuning a linear layer on top, for the sake of time.

more feedback given not stated here

Response:

Thank you for all your feedback, we have followed your suggestions on this