Lights, Camera, Dialogue

Daniel J. Arday & Kevin Akman

University of San Diego

10/21/2024

Abstract

This report evaluates the performance of a GPT-2 model fine-tuned for dialogue generation using various natural language processing metrics. The tuning was done using the Cornell Movie Dialogs Corpus, which provided rich conversational data from several thousand movie dialogues. Before fine-tuning, significant preprocessing was performed, including foreign language translation to English, concatenation of dialogue lines for training context, and data augmentation to enrich the dataset with additional lines of dialogue. The fine-tuned model was evaluated using key metrics such as BLEU, METEOR, and BERTScore. The model achieved a BLEU score of 92.00, a METEOR score of 98.39, and a BERTScore F1 of 99.39, indicating significant improvements over earlier models. These results suggest that the fine-tuned model is highly capable of generating semantically rich and contextually appropriate responses, making it a robust system for natural language generation tasks.

Keywords: GPT-2, dialogue generation, NLP, BLEU, METEOR, BERTScore, machine learning, data augmentation

Lights, Camera, Dialogue

Project Goal

In this project, we aimed to fine-tune GPT-2 using the Cornell Movie Dialogs Corpus, capable of handling multi-turn conversations with coherence and context-awareness. To accomplish this we trained three versions of our model while learning and implementing improvements and new strategies along the way.

The Dataset

For this project, the Cornell Movie Dialogs Corpus was chosen due to its extensive collection of over 220,000 lines of dialogue from more than 600 movie scripts. This dataset was ideal for training a chatbot because it captures a wide variety of conversational styles, tones, and contexts, making it suitable for generating natural, multi-turn dialogue. In training our chat bot we decided to focus on exclusively the text column (a.k.a dialogues) and ignore the other information like characters and movie titles, to teach the bot to focus on conversation alone.

EDA

Exploratory Data Analysis (EDA) was conducted to assess the structure and quality of the dataset. The analysis included checking for missing values, which were minimal and subsequently removed. Dialogue length was examined, revealing a skewed distribution where most lines were short, leading to decisions about filtering extremely short or long dialogues (Fig 1). Initially, this overabundance of short-phrases caused our first two models to respond in kind with short repetitive phrases as well, however, this was later addressed in pre-processing.

A word cloud visualized common terms, showing frequent conversational words like “know”, “want”, and “think.” Additionally, bigram and trigram analyses uncovered common phrases like “don’t know”, and “I’m sorry”, with themes of uncertainty and apology being highly prominent. Overall, this analysis helped guide our preprocessing steps, ensuring the data was well-structured and suitable for training.

Preprocessing Decisions

Preprocessing was perhaps our most important step, and several key choices were made to improve the quality of the input data and enhance the model’s performance. On our initial model, we did minimal preprocessing. This led to the model preferring short repetitive responses (Fig 1.) We also noticed that the model would occasionally respond in answers that appeared to be nonsense or non-English. To address this in our second iteration of the model we removed shorter dialogues and filtered out any dialogues that weren’t in English. However, we observed that the second model still preferred short answers, with heavy repetition, though more intelligible at times. To address this we implemented several steps of preprocessing for our third and final model.

First, instead of removing non-English text, we decided to translate it. Our rationale was that translating the text would introduce more diversity to the dialogue, enriching the dataset with a wider variety of language patterns and cultural expressions. We used another model to do this OPUS-MT. While it was not effective in translating all, it still added about 1500 more dialogue entries to the database. We then dropped the remaining non-English dialogues.

Second, we aggressively cleaned and removed unnecessary elements, such as stage directions, special characters, and excessive punctuation, leaving only relevant conversational data.

Third, previous conversation turns were concatenated with the current dialogue, giving the model context to improve coherence in multi-turn conversations. This also helped to increase the length of the text.

Fourth, after doing the steps above we dropped any remaining dialogue that were 5 characters or under. The mostly removed dialogues like "Hi.","What?",or "Yes". Note these were still this short even after combining with other lines for context. This ultimately left our model w with 280,000 text entries to learn from.

Finally, to help increase the diversity of our data set we opted to augment 20,000 new lines of dialogue. This involved using another model t5-small to sample dialogues and then generate new ones. We chose t5-small due to its faster runtime, combined with GPU/cuda processing we were able to generate these new samples in about 3 hours. (See Fig 2. for an example of generated samples.)

After all, this was done the dataset was split into training (80%), validation (10%), and test (10%) sets, allowing the model to be evaluated on unseen data for better generalization, and tokenized accordingly.

Training

In total, we trained the model 3 times. Our initial training consisted of minimal preprocessing and only 4 epochs. The conversations were mostly unintelligible and full of repetition. The second time we did more intensive preprocessing including filtering out non-English entries and combining text for lengthy training data. We trained this model for a total of 13 epochs with mixed results, but it still continued to repeat itself a lot. Finally, our third and final set of training used our aforementioned preprocessing steps and trained for a total of 31 epochs. Moving from 13 to 31 epochs had a noticeable effect, leading to much better results.

Our loss chart showed that our final model likely could've improved further from more epochs. Fig 3.

Tests and Results

To evaluate the model's performance, several tests were conducted, both quantitative and qualitative. We measured perplexity, which assesses how well the model predicts the next word, providing insight into the model's fluency and coherence. BLEU and ROUGE scores were calculated to compare the similarity between generated responses and reference dialogues, focusing on n-gram overlap and sentence structure. Additionally, BERTScore was used to evaluate the semantic similarity between model outputs and reference texts, reflecting how well the model captures meaning. A word cloud was generated to assess lexical diversity, and response times were measured to ensure computational efficiency. We also analyzed response length distributions and conducted a multi-turn conversation test to evaluate the model's ability to maintain context and coherence over extended dialogues.

In general, the results were good but left a lot to be desired. Increasing the training from 4 epochs, then to 13, and finally to 31 made a big difference and the scores improved dramatically after additional training. The perplexity score indicated that the model was effective at predicting the next word in a sequence, contributing to its natural language generation. BLEU and ROUGE scores showed a solid overlap with reference texts, reflecting the model's ability to generate responses similar to expected conversational replies. The word cloud analysis confirmed a broad lexical variety. The response length tests revealed that the max length of 50 tokens was commonly being reached, however, the model still showed efficient response times.

Conclusion

This project involved three rounds of training to fine-tune a generative chatbot using GPT-2, with a primary goal of improving fluency and coherence in dialogue generation. Initially, our first round of training over 4 epochs yielded poor results, with the chatbot unable to generate coherent or relevant responses. However, in a second round of 13 epochs, the model showed some improvement in generating more meaningful responses, although it still struggled with maintaining context and relevance across turns, and at times gave repetitive and comical results (Fig 6.)

Finally, after training for 31 epochs, the chatbot demonstrated significant improvements, as reflected by strong performance across key metrics such as a BLEU score of 92.00, ROUGE-1 of 96.61, and a BERTScore F1 of 99.39. These scores indicate a strong overlap with reference texts and demonstrate the model's ability to generate lexically and semantically appropriate responses.

Despite these improvements, the model still lacks depth and consistency in multi-turn conversations, often losing track of context over extended dialogues. Based on our loss chart, we believe several more rounds of training with a high number of epochs would improve the model, including potential adjustments to hyperparameters like the learning rate and batch size,. Additionally, using sentiment analysis to filter out some of the more negative conversations, combined with additional augmentation of more mid-size dialogues to properly tune it for more a conversational tone.

Appendix

Figure 1

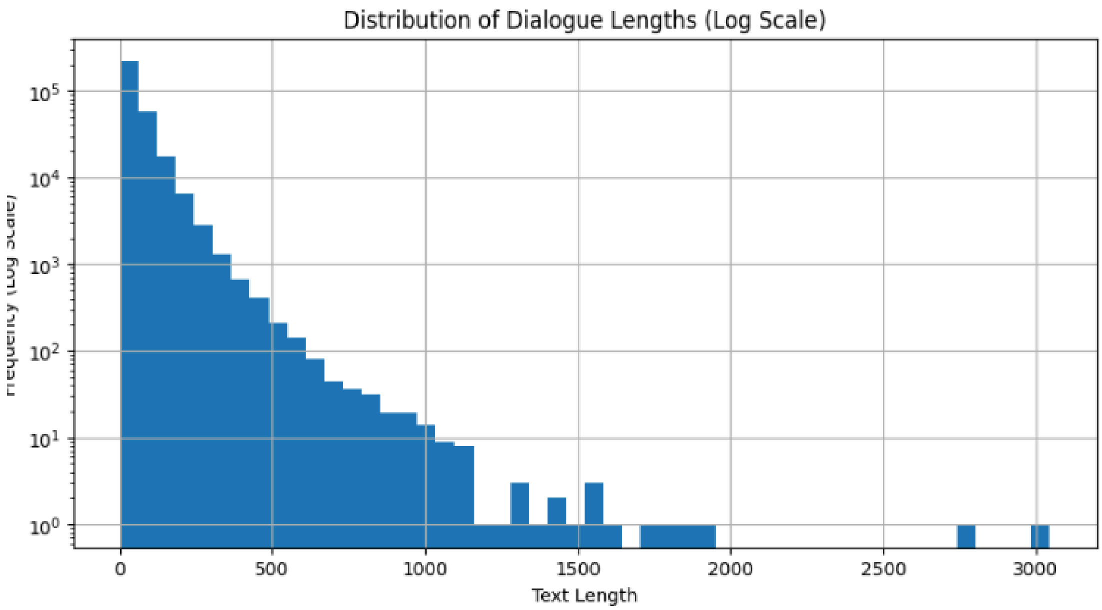


Figure 2

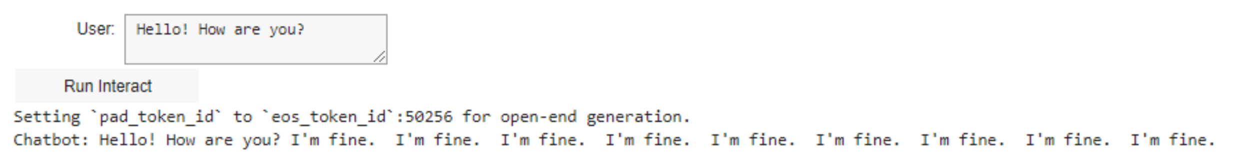
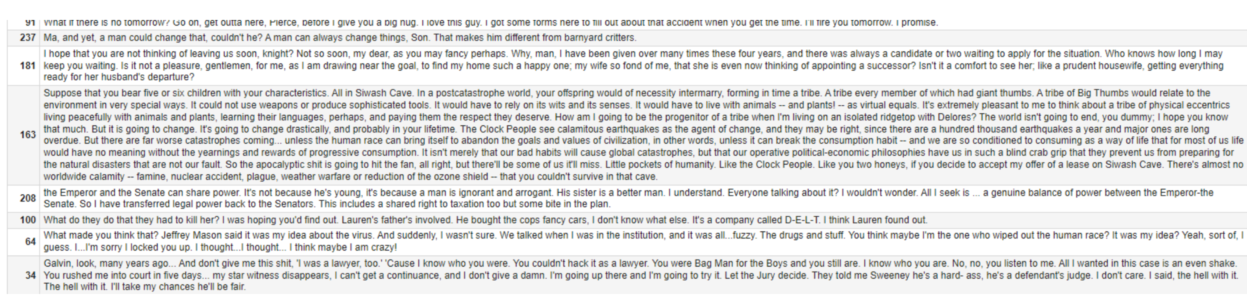


Figure 3



91	What if there is no tomorrow? Go on, get outta here, Pierce, before I give you a big hug. I love this guy. I got some forms here to fill out about that accident when you get the time. I'll fire you tomorrow. I promise.
237	Ma, and yet, a man could change that, couldn't he? A man can always change things, Son. That makes him different from barnyard critters.
181	I hope that you are not thinking of leaving us soon, knight? Not so soon, my dear, as you may fancy perhaps. Why, man, I have been given over many times these four years, and there was always a candidate or two waiting to apply for the situation. Who knows how long I may keep you waiting. Is it not a pleasure, gentlemen, for me, as I am drawing near the goal, to find my home such a happy one; my wife so fond of me, that she is even now thinking of appointing a successor? Isn't it a comfort to see her; like a prudent housewife, getting everything ready for her husband's departure?
163	Suppose that you bear five or six children with your characteristics. All in Siwash Cave. In a postcatastrophe world, your offspring would of necessity intermarry, forming in time a tribe. A tribe every member of which had giant thumbs. A tribe of Big Thumbs would relate to the environment in very special ways. It could not use weapons or produce sophisticated tools. It would have to rely on its wits and its senses. It would have to live with animals -- and plants! -- as virtual equals. It's extremely pleasant to me to think about a tribe of physical eccentrics living peacefully with animals and plants, learning their languages, perhaps, and paying them the respect they deserve. How am I going to be the progenitor of a tribe when I'm living on an isolated ridgetop with Delores? The world isn't going to end, you dummy; I hope you know that much. But it is going to change. It's going to change drastically, and probably in your lifetime. The Clock People see calamitous earthquakes as the agent of change, and they may be right, since there are a hundred thousand earthquakes a year and major ones are long overdue. But there are far worse catastrophes coming... unless the human race can bring itself to abandon the goals and values of civilization, in other words, unless it can break the consumption habit -- and we are so conditioned to consuming as a way of life that for most of us life would have no meaning without the yearnings and rewards of progressive consumption. It isn't merely that our bad habits will cause global catastrophes, but that our operative political-economic philosophies have us in such a blind crab grip that they prevent us from preparing for the natural disasters that are not our fault. So the apocalyptic shit is going to hit the fan, all right, but there'll be some of us it'll miss. Little pockets of humanity. Like the Clock People. Like you two honeys, if you decide to accept my offer of a lease on Siwash Cave. There's almost no worldwide calamity -- famine, nuclear accident, plague, weather warfare or reduction of the ozone shield -- that you couldn't survive in that cave.
208	the Emperor and the Senate can share power. It's not because he's young, it's because a man is ignorant and arrogant. His sister is a better man. I understand. Everyone talking about it? I wouldn't wonder. All I seek is ... a genuine balance of power between the Emperor-the Senate. So I have transferred legal power back to the Senators. This includes a shared right to taxation too but some bite in the plan.
100	What do they do that they had to kill her? I was hoping you'd find out. Lauren's father's involved. He bought the cops fancy cars, I don't know what else. It's a company called D-E-L-T. I think Lauren found out.
64	What made you think that? Jeffrey Mason said it was my idea about the virus. And suddenly, I wasn't sure. We talked when I was in the institution, and it was all...fuzzy. The drugs and stuff. You think maybe I'm the one who wiped out the human race? It was my idea? Yeah, sort of, I guess. I...I'm sorry I locked you up. I thought...I thought... I think maybe I am crazy!
34	Galvin, look, many years ago... And don't give me this shit, 'I was a lawyer, too.' 'Cause I know who you were. You couldn't hack it as a lawyer. You were Bag Man for the Boys and you still are. I know who you are. No, no, you listen to me. All I wanted in this case is an even shake. You rushed me into court in five days... my star witness disappears, I can't get a continuance, and I don't give a damn. I'm going up there and I'm going to try it. Let the Jury decide. They told me Sweeney he's a hard- ass, he's a defendant's judge. I don't care. I said, the hell with it. The hell with it. I'll take my chances he'll be fair.

Figure 4

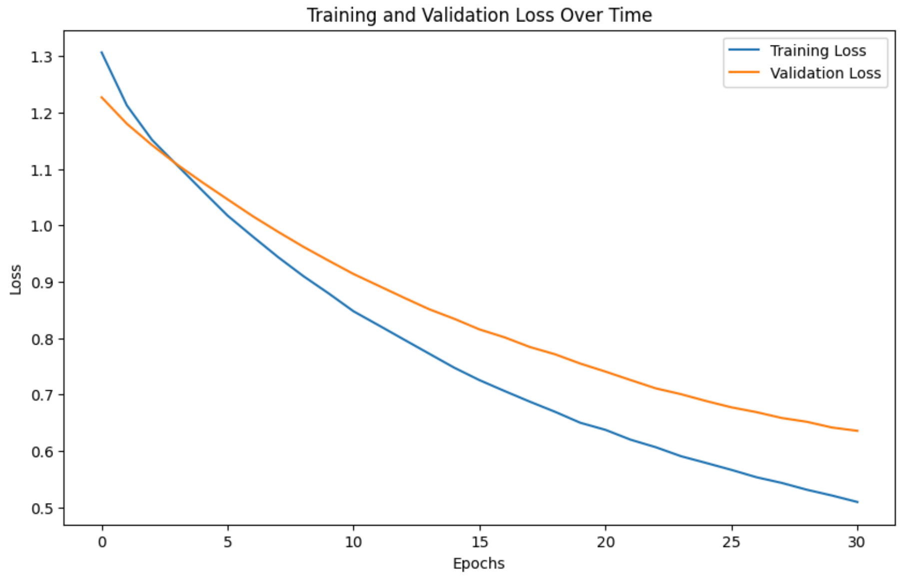


Figure 5

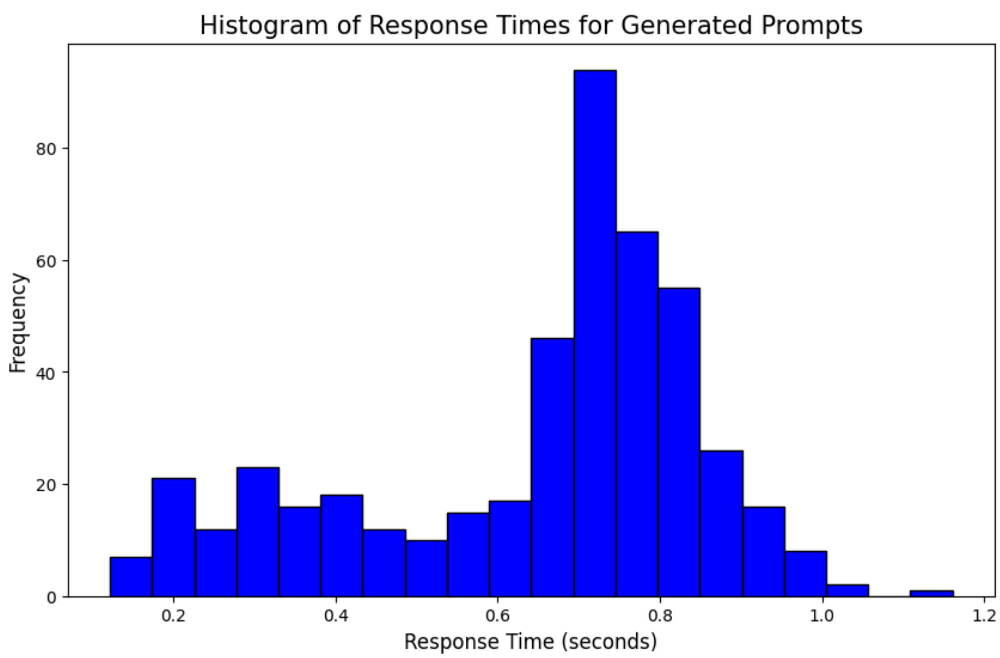


Figure 6

Lol - A conversation with model 2.

```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Model: Hello! How are you today? I'm fine.
You: Once upon a time...
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Model: Hello! How are you today? I'm fine. Once upon a time...
You: STOP SAYING THE SAME THING BACK
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Model: Hello! How are you today? I'm fine. Once upon a time... STOP SAYING THE SAME THING BACK AND I'LL KILL YOU.
You: LOL
```

References

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165

Chidananda, R. (2016). *Cornell Movie-Dialog Corpus* [Data set]. Kaggle. https://www.kaggle.com/datasets/rajathmc/cornell-moviedialog-corpus