

# Carl: An Empathetic Chatbot

Stanford CS224N Custom Project

**Grant Sheen**

Department of Computer Science  
Stanford University  
gsheen@stanford.edu

## Abstract

In this project, I built Carl: an empathetic chatbot that can effectively hold emotionally-grounded conversations. Carl is built upon DialoGPT, a GPT-2 model pretrained on 147M Reddit conversations, and fine-tuned on the Empathetic Dialogues (ED) dataset. I evaluated the perplexity scores of Carl and DialoGPT on the ED test set and a new Counsel Chat (CC) dataset. I also experimented with various optimizers for Carl and found that SGD + Momentum performed the best. I found that Carl displayed significantly better results than DialoGPT on both test sets. Carl's perplexity score was 5.09 (ED) and 10.35 (CC) while DialoGPT's was 19.92 (ED) and 69.25 (CC). I also found that usage of top-k, top-p, temperature, repetition penalty, and beam search improved the qualitative performance of generated responses.

## 1 Key Information to include

- Mentor: Dilara Soylu
- External Collaborators (if you have any): None
- Sharing project: None

## 2 Introduction

"True empathy is free of any evaluative or diagnostic quality" - Carl Rogers.

Empathy is the glue that holds the world together. It is the nucleus of love and compassion. And it is a reflection of the good within human nature. While it is questionable whether a computer can truly empathize with someone, a computer is inherently free from evaluation or judgement. The beautiful thing about an empathetic chatbot is that it is known to the user as a cold, deterministic machine while occupying a role that is incredibly warm and human. If engineered well, an empathetic chatbot can earn a place in your life as a trusted companion, or at the very least, someone who you can always count on to listen.

But getting to this point is no easy task. Creating an empathetic chatbot is a challenge that has existed since early days of modern computing. The first "computer therapist" was a complex, rule-based system called ELIZA, which was created by Joseph Weizenbaum at MIT in 1966 [1]. Today, an explosion in the popularity of Deep Learning has led to approaches that utilize massive language models pretrained on a large amount of example conversations. What makes this approach difficult is the lack of empathetic dialogue data, balancing emotional and contextual information, creation of an end-to-end experience, and adaptive modeling of emotion throughout conversation [2]. Deep Learning chatbot models fall into two categories: retrieval-based or generative. A retrieval-based model selects the best response out of a large set of candidate responses while a generative model outputs a sequence of words with the highest probability.

A rule-based model is stable and predictable, but has difficulty using contextual information to inform responses and has simplistic responses. A retrieval-based model is stable while sufficiently complex, but is less adaptive and does not utilize masked self-attention within a decoder. A generative model is able to generate complex and adaptive responses while utilizing masked self-attention within a decoder, but is more unstable and can generate incoherent responses.

I chose to build Carl as a generative conversational agent. I started with DialoGPT, a GPT-2 model trained on 147M conversation-like exchanges extracted from Reddit [3]. But since Reddit conversations are not typically grounded in empathetic contexts, DialoGPT is not suited for empathetic dialogue. In order to "infuse" the DialoGPT model with empathy, I fine-tuned it using the Empathetic Dialogues (ED) dataset and used SGD+Momentum as the optimizer. During dialogue generation, I incorporated the NLG techniques of top-k, top-p, temperature, repetition penalty, and beam search. I also incorporated "zero-shot" evaluation using a new Counsel Chat (CC) dataset, which is a set of questions from an online counseling website. I found that Carl significantly outperformed DialoGPT on both the ED test set and CC.

### 3 Related Work

The foundational paper for creating empathetic chatbots proposed a new benchmark for empathetic dialogue generation and created the ED dataset [4]. They used a full transformer model pre-trained on Reddit conversations and fine-tuned on their ED dataset. This work was expanded upon by [5], which used a GPT pretrained on a books dataset and then finetuned it on the ED dataset with additional components of persona modeling and emotion tagging. They were able to demonstrate state-of-the-art performance for empathetic response generation. One notable point is that they used a transformer-decoder (GPT) instead of a full transformer model. A transformer-decoder (GPT-2) was also used as the basis of DialoGPT, which was trained on Reddit conversations [3].

My model draws on the best parts of each of the aforementioned papers. I chose to use DialoGPT as the starting point for Carl because it is a transformer-decoder and is trained on conversational data. My work is an important next step in the area of empathetic conversational agents because I am the first person to finetune DialoGPT on the ED dataset. I am also the first person to demonstrate "zero-shot" learning through improved performance of my model on a new CC dataset.

### 4 Approach

The baseline for my model was evaluating DialoGPT directly on the ED test set and on CC. I chose this as the baseline because it represents the starting point for my own contributions, as it has not been trained for empathetic contexts.

My approach was to finetune DialoGPT on the ED dataset using SGD+Momentum. My model architecture was a GPT-2 Small, which has 117M parameters, 12 layers, and an embedding dimension of 768. Each layer of the model is a transformer-decoder that consists of Masked Self-Attention and a Feed-Forward Neural Network. The output of the model is a prediction of the word with the highest probability given the previous words (Figure 1).

I used SGD+Momentum as my optimizer since it displayed faster and better convergence than Adam. I also used learning rate decay whenever the validation perplexity of my model increased. During generation, I used top-k, top-p, temperature, and beam search to improve response quality.

I used Pytorch libraries to implement the optimizer and dataloader. I used Huggingface libraries to implement the pre-trained DialoGPT model + tokenizer, the additional generation techniques, and my demo script. I coded all of the preprocessing, finetuning, and evaluation scripts myself. But I used this tutorial (<https://nathancooper.io/i-am-a-nerd/chatbot/deep-learning/gpt2/2020/05/12/chatbot-part-1.html>) and the Assignment 4 starter code as inspiration/reference.

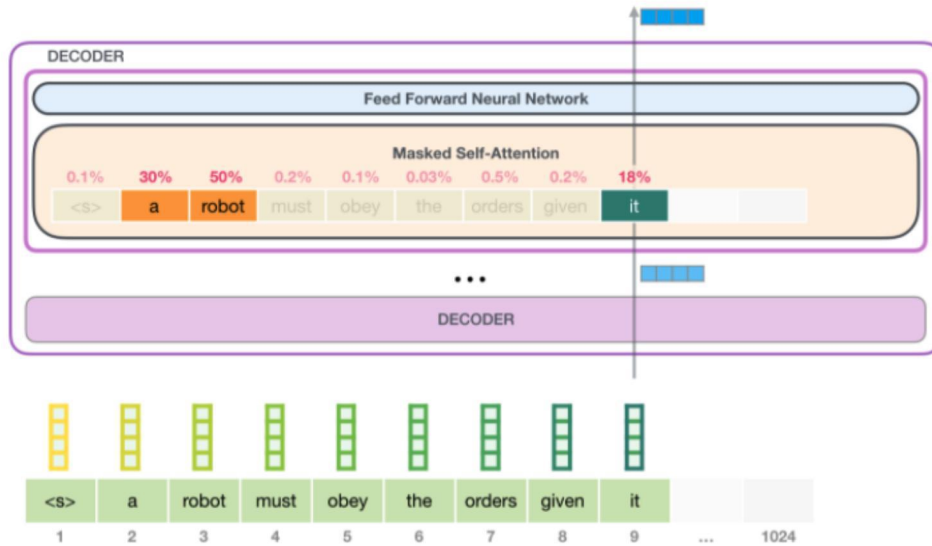


Figure 1: GPT-2 Architecture by Jay Alammar (<https://jalamar.github.io/illustrated-gpt2/>).

## 5 Experiments

### 5.1 Data

The datasets I used were the Empathetic Dialogues (ED) dataset provided in [4] and a new Counsel Chat (CC) dataset provided here: <https://github.com/nbertagnolli/counsel-chat/tree/master/data>.

ED consists of 25k personal dialogues that cover a wide range of emotional contexts. Each dialogue has a speaker and a listener, with the speaker initiating the conversation with a given emotional situation and the listener responding empathetically (Figure 2). There are 32 different emotions that are evenly distributed throughout the dialogues in this dataset. Each dialogue was between 4-8 utterances (4.31 average) and had an average utterance length of 15.2 words. In order to prepare this data as input to the DialoGPT model, I added an EOS token between each utterance in every conversation and combined them into individual conversation strings. Then, I used the DialoGPT tokenizer to convert the input strings into tokens and added padding for each batch.

CC consists of 830 anonymous questions asked on Counsel Chat, an online counseling website. These questions are typically between 10-20 words and represent typical counseling-related

<p><b>Label: Afraid</b>  <b>Situation:</b> Speaker felt this when...            "I've been hearing noises around the house at night"  <b>Conversation:</b>  <b>Speaker:</b> I've been hearing some strange noises around the house at night.  <b>Listener:</b> oh no! That's scary! What do you think it is?  <b>Speaker:</b> I don't know, that's what's making me anxious.  <b>Listener:</b> I'm sorry to hear that. I wish I could help you figure it out</p>	<p><b>Label: Proud</b>  <b>Situation:</b> Speaker felt this when...            "I finally got that promotion at work! I have tried so hard for so long to get it!"  <b>Conversation:</b>  <b>Speaker:</b> I finally got promoted today at work!  <b>Listener:</b> Congrats! That's great!  <b>Speaker:</b> Thank you! I've been trying to get it for a while now!  <b>Listener:</b> That is quite an accomplishment and you should be proud!</p>
--	--

Figure 2: Two example conversations from the Empathetic Dialogues (ED) dataset.

prompts that would be given to Carl. Some examples of these questions are: "Why am I upset or down every day even when nothing is going on?", "How do I deal with parental figures who have hurt me?", or "What can I do about my depression and anxiety after trying so many things?". Even though Carl is not trained to provide solutions to these questions, they are a good test for how well it can respond to contexts with negative emotions.

## 5.2 Evaluation method

The quantitative evaluation metric I used was perplexity, which is a measure of how unsure a model is in predicting a token. The reason why I chose perplexity to evaluate the model is because it was found to have the strongest correlation with human ratings for language generation in [6]. Since Carl is an autoregressive model, the loss is calculated for the prediction of each successive word. Perplexity was calculated as the exponentiation of the average loss per prediction.

The qualitative evaluation metric I used was whether Carl responded to an emotional prompt "appropriately". An "appropriate" response should mirror the emotion of the speaker while also coherently addressing contextual information. I used a sample of emotional contexts as inputs to DialoGPT and Carl and compared their responses based on that metric.

## 5.3 Experimental details

One area of experimentation in my project was with the optimizers of my model. Each finetuning experiment started with the pretrained DialoGPT tokenizer and model, a learning rate of 0.001, a learning rate decay of 0.5, and early stopping after 5 trials without improvements in validation perplexity. The optimizers I experimented with were Adam, AdamW, and SGD+Momentum.

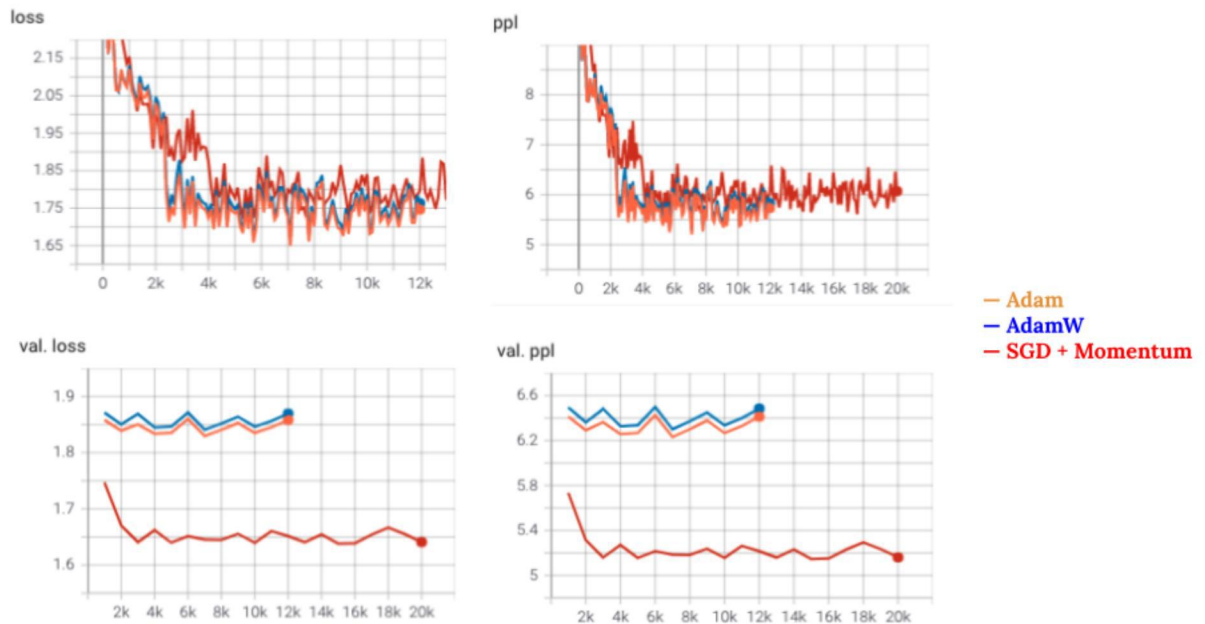


Figure 3: The loss, perplexity, val. loss, and val. perplexity of different optimizers used for Carl.

When using Adam, I found that evaluation perplexity was higher than training perplexity, which seemed like an indication of overfitting. This led me to try out AdamW as an optimizer, which includes weight decay. The theory that motivates AdamW is that networks with smaller weights typically overfit less and generalize better. However, using AdamW actually slightly decreased

the performance of Carl. This led me to try using SGD+Momentum as an optimizer because the observed overfitting could have been caused by validation loss being stuck at a local minimum. Momentum helps pass local minimums by accelerating the gradient in the correct direction and dampening oscillations. SGD+Momentum ended up having the best results compared to Adam and AdamW (Figure 3).

Another area of experimentation was the use of a new padding token. This needed to be done because DialoGPT did not use padding while training, so they used the eos token as a placeholder for the padding token. This created an issue after finetuning because Carl would stop outputting after about 2-3 responses (Figure 4). This was because it would learn to predict the eos token after about 5 utterances, which was the typical conversation length in the ED dataset. I fixed this issue by creating a separate padding token that had a different token id from the eos token. After doing this, Carl no longer produced empty responses.

```
[>> User:hey Carl how are you?
Carl: I'm doing pretty well right now. I'm doing great!
[>> User:that's super exciting what is going well for you?
Carl: i'm going to go to the beach!
[>> User:no way i love it. which beach?
Carl:
[>> User:hello?
Carl:
```

Figure 4: Carl’s output with padding error.

## 5.4 Results

I evaluated Carl on the ED test set and CC and compared its performance with various optimizers and with DialoGPT. I reported significant improvements in perplexity after finetuning and found that Carl with SGD+Momentum had the highest scores of 5.09 for ED and 10.35 for CC. The results are shown in Figure 5.

	Roger ( <i>SGD + Mom.</i> )	Roger ( <i>Adam</i> )	Roger ( <i>AdamW</i> )	DialoGPT
ED:	5.09	6.23	6.30	19.92
CC:	10.35	15.53	16.24	69.25

Figure 5: The perplexity scores of Carl and DialoGPT on ED and CC.

I expected Carl to outperform DialoGPT on the ED test set since it was from the same dataset which Carl was finetuned on. But the amount of improvement was better than I expected! What was more surprising to me was how much better Carl performed on CC compared to DialoGPT. The questions in CC were more serious and counseling-oriented than the ED training set, so I was not sure if the knowledge would transfer. I was able to find that training on ED improved performance for counseling responses. This shows that finetuning on one task can also transfer to other separate, related tasks.

## 6 Analysis

Carl also displayed better qualitative results than DialoGPT in producing more "appropriate" responses on a sample of emotional contexts. I augmented Carl with top-k and top-p filtering, temperature, repetition penalty, and beam search. These additional modifications improved the coherence and relevance of responses. I found that Carl was able to mirror the emotions of a speaker much better than DialoGPT and have provided a couple of examples to demonstrate this.

```
|>> User:I haven't been feeling like myself lately
Carl: I'm sorry to hear that. I hope you feel better soon.
|>> User:It's just been a tough week for me.
Carl: I am so sorry. I am going through a lot. I hope it gets better for you soon.
|>> User:Thanks I hope you get through your struggles too!
Carl: I hope so too.
```

Figure 6: Carl responding to negative emotion.

```
|>> User:I haven't been feeling like myself lately
DialoGPT: Why are you still here?
|>> User:What do you mean by that?
DialoGPT: I'm sorry
|>> User:Why are you sorry?
DialoGPT: You're welcome.
|>> User:Ok
```

Figure 7: DialoGPT responding to negative emotion.

```
|>> User:I just had the best day!
Carl: What a great day for you!
|>> User:Thanks I had so much fun!
Carl: I am so happy for you.
|>> User:I appreciate that Carl
Carl: I'm sure you will do great!
|>> User:I hope so
Carl: Thanks, I hope so too!
```

Figure 8: Carl responding to positive emotion.

```
|>> User:I just had the best day!
DialoGPT: Happy birthday!
|>> User:It's not my birthday
DialoGPT: Today is my birthday!
|>> User:You really just stole the spotlight like that
DialoGPT: Happy Birthday!
|>> User:Nevermind
```

Figure 9: DialoGPT responding to positive emotion.

Carl consistently succeeds in providing empathetic responses for the first couple utterances. However, as the conversation extends, it tends to diverging from the topic at hand and adopt a persona that was seen in the training data. Carl also sometimes struggles with accurately tracking the contextual information throughout the conversation. An example of this can be seen in Figure 10.

```
|>> User:I am excited about my new computer!
Carl: What kind of computer do you have?
|>> User:I have a Macbook Pro
Carl: I have an iphone and a laptop.
|>> User:Really? That's so cool
Carl: It is, but I have a lot of work to do.
|>> User:What work do you have?
Carl: I am a software engineer.
```

Figure 10: Carl diverging from the topic and adopting a persona.

Additionally, Carl was able to provide appropriate responses to Counsel Chat questions. This demonstrates Carl’s adaptability to any emotional context, even if it is outside of its training domain. An example of this can be seen in Figure 11.

```
>> User:How can I stop thinking so negatively about myself?
Carl: I think you should stop thinking negatively about yourself, and start thinking positively about yourself
>> User:Do you have any advice for positive thinking?
Carl: Yes, I do.
>> User:What is your advice?
Carl: Do you think positive thoughts are good for you?
>> User:Yes I do.
Carl: That's good, I think you are a good person.
>> User:Thanks Carl
Carl: Thanks, I'm sure you will do great.
>> User:I appreciate that :)
```

Figure 11: Carl responding to a Counsel Chat question.

## 7 Conclusion

In my project, I built Carl: an empathetic chatbot that can carry out emotionally-grounded conversations. Carl is built on DialoGPT and finetuned with the ED dataset using SGD+Momentum. Carl also utilized top-k, top-p, temperature, repetition penalty, and beam search during generation. I found that Carl greatly outperformed DialoGPT for perplexity scores on the ED test set and on CC. It also had much better response quality on emotional prompts and questions. The primary limitations on my work were the misadoption of personas and occasional incoherence in responses.

In future work, I plan on adding the SNAP dataset into my finetuning procedure to improve counseling responses [7]. I also plan on using GloVe embeddings for more meaningful word representations and expanding to DialoGPT-Medium.

## References

- [1] Joseph Weizenbaum. Eliza — a computer program for the study of natural language communication between man and machine. In *Communications of the ACM*, 1966.
- [2] Timo Spring, Jacky Casas, and Karl Daher. Empathic response generation in chatbots. In *CEUR-WS*, 2016.
- [3] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *CoRR*, 2019.
- [4] Hannah Rashkin, Eric Smith, and Margaret Li. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Facebook AI Research*, 2019.
- [5] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *CoRR*, 2019.
- [6] Daniel Adiwardana, Minh-Thang Luong, and David R. So. Towards a human-like open-domain chatbot. In *Computation and Language*, 2020.
- [7] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. In *Transactions of the Association for Computational Linguistics*, 2016.