BlenderBot/Llama2: An Instruction-Tuned ChatBot for Mental Health

Prathamesh Bagal, Andrew Cheung, Shi Bin Lim

Introduction. Mental health is emerging as an increasingly critical challenge, affecting over 1 in 5 American adults and disproportionately impacting young adults at a higher prevalence[1]. People are feeling more isolated than ever despite having access to more avenues of connection with the rise of digital technology. This issue has been further exacerbated by the recent pandemic. As a result, there is a higher demand for mental healthcare amid a shortage of mental health providers. Psychologists are exploring more ways to meet this demand through technology, such as telemedicine or awareness programs, to get people the help they need. However, not everyone has access to local programs. This leads many to turn to online platforms such as Reddit to share their feelings and stories, hoping to find a connection in these virtual spaces. Our group wants to expand on the options for mental healthcare by creating a chatbot that can serve as a safe space for people.

We fine-tuned BlenderBot (Roller et al., 2021), an advanced conversational AI developed by Facebook, and Llama2 (Touvron et al., 2023), an open-source chat model optimized for dialogue use cases, to engage in mental health conversations with users. Our instructions ensure that the chatbot only provides helpful responses, without providing or encouraging any harmful, illegal, or dangerous content. Additionally, it is trained to seek clarification if the question is not factually coherent and cautioned to not spread any false information. Three versions were trained using online datasets for both chat models. These were then evaluated together with their Vanilla versions, based on their ability to provide helpful responses, akin to a psychologist's response, and compared using the BLEU score metric.

It is important to note that the models that we trained are not a real replacement for therapy. There are many hard topics that only a medical professional can tackle. We do not endorse users to use these chatbots with the expectation of mental health benefits and do not take any responsibility for their actions. We are purely interested in identifying whether large language models (LLMs) are capable of mimicking speech patterns provided by psychologists and whether data obtained through virtual spaces, like Reddit, could be useful in improving their performance.

Background. LLMs are trained on vast amounts of text and are capable of performing language-related tasks. These models have been pre-trained on a large corpus of data. Instruction-tuning is a tactic to guide a model's learning process on a specific task (Dey et al., 2024), and in this case, to act as a helpful mental health chatbot. By providing additional data for the chatbot during this fine-tuning process, the model will learn to mimic the speech patterns of the most recent data provided to them.

Data. We fine-tuned a model using a combination of three datasets. Two are Hugging Face datasets, referenced as phr_mental_therapy_dataset (PHR), a conversational dataset for fine-tuning chatbot models as a mental therapy assistant providing safe answers, and mental_health_counseling_conversation (Counseling), that contains prompts and answers that were obtained from real licensed psychologists. The last dataset was obtained through the Reddit API, in subreddits where people discuss their mental health issues or are looking for advice.

Counseling is a collection of questions ('context') and answers ('responses') from two online counseling and therapy platforms. The questions cover a range of mental health problems and the answers are provided by professional psychologists. The dataset contains 3.51k rows. There are repeated questions, with answers provided by different psychologists.

PHR has been synthetically generated by using GPT3.5-turbo. It is formatted as a dialog between a user and a chatbot. Each back-and-forth is counted as one training sample: there are multiple training samples per

one row. It was designed as a dataset to be used for training for mental health conversations.

Reddit posts and comments were scraped using a custom code that obtains the top 25 posts of all time from the following subreddits where users discuss a wide range of topics that may affect their mental health: r/offmychest, r/advice, r/mentalhealth, r/confessions, r/self, r/AITAH, r/AmITheAsshole, r/depression, r/anxiety, r/relationships, and r/family. The content from these posts is trained as user inputs and up to 15 comments (that meet a minimum threshold of 500 characters) are used as responses. The rationale behind this threshold is that longer comments usually add more perspective and provide more qualitative responses than shorter ones.

Counseling data	➢ Questions and answers provided by qualified psychologists ➢ 2 MB ➢ 3.51k rows ➢ 300,000 tokens
PHR synthetic data	➢ Synthetically generated conversations using GPT3.5-turbo ➢ 211 MB ➢ 99.1k rows ➢ 250,000 tokens
Reddit data	➢ ~15 comments from top 25 posts of all time for each subreddit ➢ 6.6 MB ➢ 1735 rows ➢ 1,500,000 tokens

Table 1. Datasets, description, size.

We trained 3 models based on the following structure in Table 2. The Vanilla model is not fine-tuned, it is the base transformer model obtained from Hugging Face. Model 1 is trained on the first 2000 rows provided in the Counseling dataset. Model 2 is trained on the first 2000 rows of Counseling and the first 2000 rows of PHR (synthetic data). Model 3 is trained on the first 2000 rows of Counseling and all 1735 rows of the Reddit data.

Model 1: Vanilla	
Model 2: Counseling	(first 2000 rows)
Model 3: Counseling + PHR	(first 2000 rows) (first 2000 rows)
Model 4: Counseling + Reddit	(first 2000 rows) (all 1735 rows)

Table 2. Model for evaluation.

The last 500 rows of the Counseling dataset will be used as the gold standard for testing. Each model will produce an output for every corresponding 'context' entry in the test dataset. The BLEU score will be calculated by comparing the outputs with the gold standard responses. It is expected that the model trained purely on the Counseling data will produce the best results. We want to see how additionally training the models on the synthetically generated dataset or the Reddit data can enhance the quality of responses generated by an LLM.

Method. The BlenderBot and Llama2 chatbots were both instruction-tuned to provide helpful responses without encouraging harmful/dangerous behavior. For models 3 and 4, we've decided to supplement the Counseling data with either synthetic PHR data or Reddit data. For the Llama2 model, instructions were fed under the following format:

<s> [INST] <<SYS>> You are a helpful and joyous mental therapy assistant. Always answer as helpfully and cheerfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. <</SYS>> **[USER INPUT]** [/INST] **[CHATBOT RESPONSE]** <s>

The PHR synthetic data was provided in this format. The Reddit data was collected and formatted such that the "user input" is the Reddit post and the "chatbot response" is the Reddit

comment. Comments were selected off criteria if they were above a lengthy threshold (>500 characters) because short answers were not valuable. Reddit posts were not limited, and improvements to this model could be made to set a minimum threshold and a maximum threshold. We discovered that Reddit posts could be unexpectedly lengthy, and we did not account that "user input" would normally not be of such great length.

The formatted input/response data was tokenized **[1. Syntax]**. Based on a given vocabulary of the specific tokenizer used, the data is broken up based on that vocabulary with known vector embeddings. For Llama2, the AutoTokenizer was used to tokenize the input. For BlenderBot, the BlenderBot small 90M Tokenizer was used. During the training loops, the weights of the model are updated based on the data that was seen **[2. Semantics]**. The weights affect the model's calculation on determining the probability of the next word given the past history. The trained LLM models **3. Transformer LLMs]** were instruction-tuned and uploaded to HuggingFace. Using a test data set composed of the next 500 rows from the unseen Counseling data, the generated output was compared to gold responses provided by the licensed psychologists using BLEU score evaluation **[4. Applications]**. BLEU score is a metric that determines how similar the generated output is compared to the psychologist's response by looking at the precision based on n-gram matches.

Model Details:

Name: BlenderBot Small 90M
Developed by: Meta
Release Date: August 2022
Type: Transformer Model
Description: An instruction-tuned model optimized for replicating therapist responses.

Intended Use:

To provide abstractive summarization of short-form creative texts by extracting underlying meanings and capturing the main ideas succinctly.

Factors:

Specifically designed to act as a conversational mental health chatbot .

Metrics:

BLEU Score: Measures the similarity between the model generated responses and the actual therapist responses.

Training Data:

- ➢ Questions and answers provided by qualified psychologists (2000 rows)
- ➢ Synthetically generated conversations using GPT3.5-turbo (2000 rows)

Evaluation Data:

- ➢ Questions and answers provided by qualified psychologists (500 rows)

Ethical Considerations:

Designed as a class project. Not intended to be used in real life scenarios. This is NOT a therapist.

Model Card for BlenderBot on Counselor data

Evaluation/Results. The expected ranking of the models in terms of performance from worst to best is described as follows: Vanilla model, Counseling + Reddit, Counseling, and Counseling + PHR. The randomness and personalized stories obtained from the comment section of Reddit posts would likely be difficult for the model to learn, though it may provide some insights into human connection. We hypothesize that synthetic data could be a great boost in learning, and if so, could be used for future models to learn more about psychology.

	BlenderBot	Llama2
Vanilla	0.016	0.168
Counseling	0.116	**0.189**
Counseling + PHR	0.112	0.176
Counseling + Reddit	**0.117**	N/A

Table 3. Average BLEU score (3-gram) of each model.

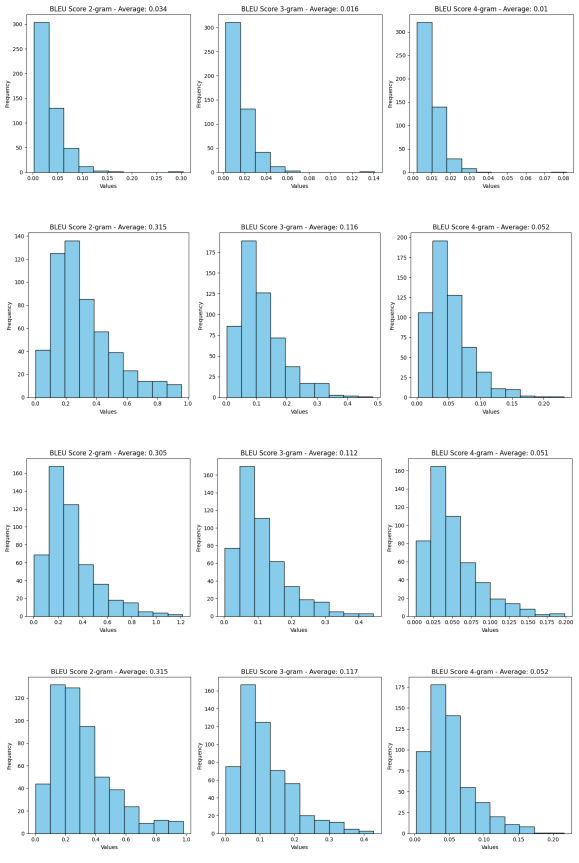

Figure 1: Histogram of distribution of BLEU scores (2-,3-,4-grams) for BlenderBot (Vanilla, Counseling, Counseling + PHR, Counseling + Reddit)

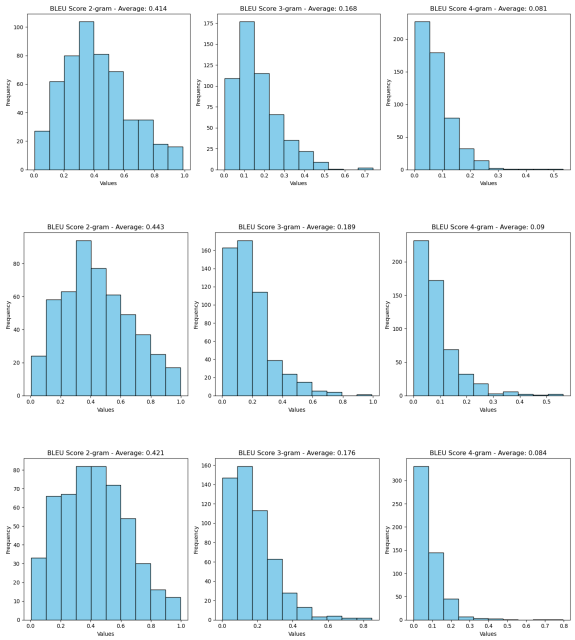

Figure 2. Histogram of distribution of BLEU scores (2-,3-,4-grams) for BlenderBot (Vanilla, Counseling, Counseling + PHR, Counseling + Reddit)

The Vanilla model for BlenderBot and Llama2 both performed worse than the fine-tuned models (Table 3, Figure 1, Figure 2). Blenderbot performed severely worse at 0.016. Llama2 performed significantly better than its BlenderBot counterpart for all models trained. Llama2 could not be run for the Counseling + Reddit data, the model was taking too long to run and the GPU was limited.

Conclusion. We hoped to contribute to the field of NLP and mental health by producing models that could mimic psychologist responses in potentially life-saving conversations. Although the models were not as good as expected, they did provide useful insight on what kind of text could be useful in training a model about mental health. From the evaluation we learn that adding Reddit data does slightly better than the model trained on Counselor data only. This could be explored further to see how the performance varies when we add more non-Counselor data. The synthetic data also performed decently on the evaluation for its model. This could be used as a base study to explore if adding synthetically generated data would be useful to train such mental health dialog systems.

References

Kim, S., Cha, J., Kim, D., & Park, E. (2023). Understanding mental health issues in different subdomains of social networking services: Computational Analysis of text-based Reddit posts. *Journal of Medical Internet Research*, *25*. https://doi.org/10.2196/49074

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., & Weston, J. (2021). Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. https://doi.org/10.18653/v1/2021.eacl-main.24

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://doi.org/https://doi.org/10.48550/arXiv.2307.09288

U.S. Department of Health and Human Services. (n.d.). *Mental illness*. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/mental-illness