Can Large Language Models (LLMs) Replicate Human Counselor Responses?

Seungjun (Josh) Kim
kims17@uci.edu

Jina Chun
chunyj@uci.edu

Simon Guo
yutong4@uci.edu
University of California, Irvine, Informatics

Abstract

Over 50% of the global population grapples with mental health illnesses or disorders. Despite earlier endeavors to create Artificial Intelligence (AI)-powered chatbots for the treatment of mental health issues, the algorithms and frameworks employed fell short in replicating the nuanced dimensions of counseling as practiced by human counselors. This project aims to explore the potential of AI and Large Language models in providing counseling support, with the ultimate objective of enhancing the accessibility of counseling services.

1 Introduction

More than 50% of the world's population experience mental health illness or disorders (e.g. depression) at some point in their lifetime (CDC 2023). Yet, people often do not have access to counseling services or mental therapy due to various factors including financial constraints and unavailability of such services in the area they live in. For example, the accessibility of these services is limited for people, especially youth, who are uninsured (Cummings et al. 2017). Previous literature suggests that this lack of accessibility to mental health services may result in increased suicide risk (Hester 2017, Ku et al.2021). While there exists some previous studies that attempted to develop chatbots powered by Artificial Intelligence (AI) for treating mental health issues, the algorithms or frameworks used were not sophisticated enough to replicate the subtle aspects of counseling performed by human counselors (Kamitat et al. 2018, Oh et al. 2017, Denecke et al 2020). We aim to address the following research question in this paper – Do the counseling texts generated by LLMs share the traits and structures that are touted as hallmarks of good counselor conversations? Through answering this research question, we explore the feasibility and effectiveness of texts generated by Large Language Models (LLMs) such as ChatGPT as a means to counsel and provide emotional support for individuals experiencing stress, depression and anxiety in particular. The use of AI to support or even partially replace counseling will enable us to overcome the prevailing barriers for mental health counseling and improve the accessibility of counseling services by providing a NLP generative model for AI-powered mental counseling support.

2 Methods

2.1 Data Collection

To be able to compare human counselor responses against the LLM generated responses, a dataset with question-answer pairs related to mental health counseling was required. For this purpose, the CounselChat dataset open-sourced on Github was used (Nbertagnolli 2023). CounselChat is an online platform where patients who need therapy or counseling can find therapists and counselors near them and also be able to filter them by specialty or type of mental health issue they are experiencing. The dataset used had 2,129 observations in total and 12 unique columns which include the question title, question text, answer text, topic, upvotes, and number of views. We focused only on the 3 types of mental health issues – anxiety, stress, and depression which left us with 586 observations in total. As for the LLM of choice for generating AI counseling texts, we chose Dolly, the LLM released by Databricks (Conover et al. 2023). Among the various versions of Dolly, we chose the lightest model with 3 billion parameters. We chose Dolly because of its relative lightness of the model which allowed us to run it locally without relying on external computational resources. In addition, Dolly has been proudly publicized as a token of LLM democratization with the little amount of corpus used and the open source nature of it (e.g., code and specific implementation released to the public, no barriers to use). We were curious as to whether even the least computationally expensive and completely open sourced LLM will be just as effective and valuable for serving as a mental health counseling tool.

2.2 Literature Review

A literature review on the hallmarks of good counselor conversations was conducted. Althoff, Clark, and Leskovec did a large-scale analysis on data retrieved from text-based counseling service platforms where people in crisis (e.g., depression, self-harm, anxiety, etc.) engage in therapeutic conversations with counselors. The findings revealed that good counselors have lengthier average message length, have more variation in their language across conversations which indicate their ability to adapt more, and spend more time on problem solving than introducing the problem than less successful counselors (Althoff 2016). Another study discovered that high-quality counseling conversations contained higher numbers of words on average per turn, had more positive sentiment on average, and scored higher in both the Linguistic Style Matching (LMS) and Linguistic Style Coordination (LSC) scores. The last trait, in particular, implied that good counselors had the tendency to mirror the language of their clients as high-quality interactions showed higher levels of linguistic alignment (Perez-Rosas 2019).

2.3 Analysis

First, analysis on the length and number of words between responses from human counselors and AI was done. This was based on findings from previous literature that good counselors tend to respond verbosely than less successful counselors.The second analysis was analyzing sentiments of the human counselor responses and the LLM-generated responses. The sentiment analysis was conducted using a Python natural language analysis tool *Stanza* to compare the sentiment scoring of the human and LLM-generated responses to observe whether LLM generated responses resemble those generated by good counselors in the context of mental health counseling (Peng et al. 2020).
Third, cosine similarity analysis was applied to measure how closely the responses produced by the LLM (e.g., Dolly) match those produced by human counselors using the scikit-learn library. In terms of the technologies and softwares for data generation and analysis, Python (version 3.11.4) and its relevant packages (e.g., pandas, numpy, sklearn, stanza, transformers) were used. Github was used to collaboratively manage versions of files and scripts.

3 Methods

3.1 Exploratory Data Analysis (EDA)

The final filtered CounselChat dataset used consisted of 586 observations and 12 columns. The features included the question title, question text, answer text, topic, upvotes, and number of views. 223 unique questions regarding depression, anxiety and stress were asked and answered by 183 unique counselors in the data. As shown in Figures 1 and 2, The number of views and upvotes that the counselor answers received both had long tails to the right with most of the data concentrated towards the lower end of the distribution and were moderately correlated with each other ($\rho = 0.286$).

3.2 Analysis on response text length and number of words

A major characteristic of good counselors mentioned by previous literature is that they tend to use language that is lengthier on average than that of less successful counselors. The number of words per response was also higher (Althoff 2016). The length of the responses and the number of words between the responses from human counselors and AI were compared against each other. In 27.30% of the instances, the AI generated responses were longer than those from the human counselors. In terms of word counts, 30.20% of AI generated responses contained more words than those from human counselors. The text lengths and word counts were also visualized into histograms to aid our understanding of how the distributions of these two statistics are different between the human counselors and AI. Figure [X] and Figure [Y] illustrate the distributions of text lengths and word counts. While all the distributions display long tails towards the right for each statistic, it can be seen that the majority of the data points from the AI generated responses lie leftward to that of the human counselor responses. These percentages and visualizations indicated that AI generated responses were generally shorter and included less words than those from human counselors. Another pattern to note was that the AI maintained more consistency in the length and number of words it used to generate responses for patients. The standard deviations in Table 1 and the relatively more uniform height of each bar in the histograms for AI generated responses reflect this information.

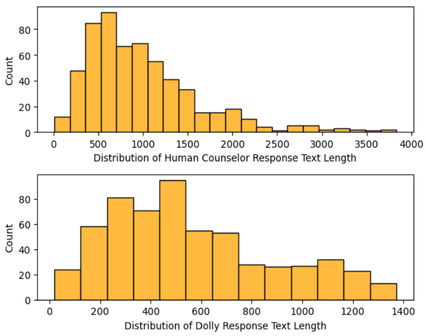

Figure 3. Distributions of Response Text Length.

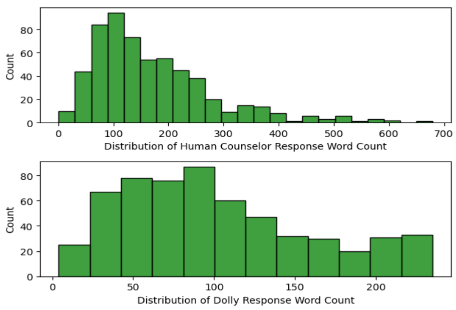

Figure 4. Distributions of Response Word Count.

	Text Length	Word Count
Human Counselor	639.50	111.55
AI (Dolly)	323.33	58.14

Table 1. Standard deviations of responses from human counselors and AI for each statistic.

3.3 Analysis on Sentiment

The sentiment analysis scores were computed for each of the responses instead of observing them at the sentence level. Specifically, a single sentiment score value was calculated for each of the responses by the human counselors and the LLM model. Table 2 below presents a distribution of responses categorized as negative, neutral, and positive for both the human counselors and the LLM model's Dolly-generated responses. The result highlights a similarity between the sentiment scores by human and AI-generated responses. This may be an indication of AI-generated answers resembling the characteristics of the human counselor’s response. Further, observing the raw data with the generated sentiment labels, most of the responses by the human and Dolly models had equal sentiment scores for a given question.

On the other hand, the number of responses categorized as having positive sentiment was the lowest among the three sentiment categories. The high number of negative sentiment classifications may be caused by the prevalence of words associated with a negative connotation in the responses. For instance, terms and words such as ‘nervous’, ‘anxiety’, ‘bad experience’, and ‘bully’ were repeatedly used in the responses generated by both the human counselors and AI. The frequent occurrences of words that are associated with negative connotations may lead to an increased likelihood of classifying the response as negative, resulting in a higher number of negative sentiment classifications regardless of overall tone, attitude, or overall intent by the human counselor and AI-generated responses.

	Negative	Neutral	Positive
Human Counselor	453	77	56
AI (Dolly)	469	74	43

Table 2. Number of responses with negative, neutral, and positive sentiment scores for human counselors and AI.

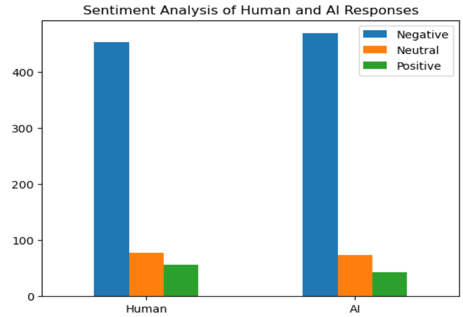

Figure 5. Bar graph representation of number of responses with negative, neutral, and positive sentiment scores for human counselors and AI.

3.4. Analysis on Similarity

The cosine similarity function from the *scikit-learn* package is applied on each pair of human and Dolly responses to the same patient question converted to term frequency–inverse document frequency (tf-idf) vector form. The results are detailed in Figure 6 and Table 3.

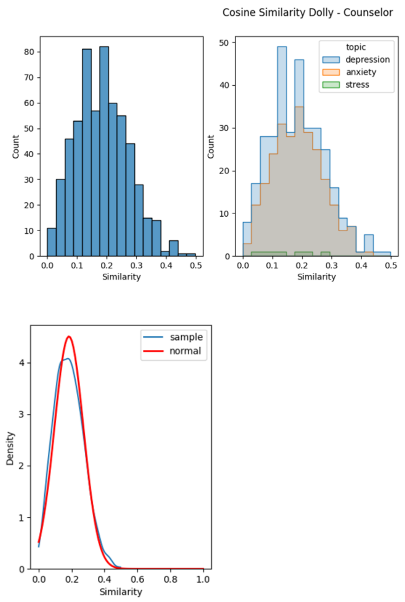

Figure 6. Distribution of Cosine Similarity Scores between Counselor and Dolly Responses.

Avg	Std	Min	25%	50%	75%	Max
0.184	0.089	0.000	0.120	0.179	0.243	0.498

Table 3. Summary of Cosine Similarity Scores between Counselor and Dolly Responses

As we can see from Table 3, the mean and median similarity scores are 0.183841 and 0.179327 respectively, which, given the possible range of score from 0 to 1, indicates a significant level of deviation between Dolly and human counselors' responses. As the cosine similarity scores are calculated from term frequency representations of the responses, the results of this analysis suggests that in response to patient questions, Dolly employs a word usage pattern distinct from that of human counselors. One aspect of this difference that can be easily quantified is that the human counselors' responses are much richer in vocabulary, adding up to 6641 distinct words used versus Dolly's 3958 distinct words. Note that this is not to say that Dolly's responses sound robotic. At least some level of human-like fluency can be discerned by reading through the Dolly sample responses.

Additionally, from Figure 6, we observe that the density of similarity scores resembles the normal distribution with the same mean and standard deviation, and this pattern holds across the topics of anxiety and depression, while the topic of stress does not contain enough samples to display a clear pattern.

3.5. Qualitative Analysis

While Dolly was LLM of choice for our analyses, we also briefly explored the responses generated by ChatGPT (OpenAI 2023), the iconic LLM from OpenAI that has now become widely known to and used by even those who are not in the technology industry, and compared them against the responses generated by Dolly and the human counselor responses for the same inquiry texts. 10 cases from the data were randomly selected for this comparison. Those selected case IDs were 371, 512, 188, 342, 505, 73, 267, 423, 37, and 112. Overall, several qualitative differences among responses from human counselors, Dolly, and ChatGPT were identified. First, ChatGPT responses were generally longer than those from Dolly. Analyses from previous sections have already shown that human counselor responses, on average, were lengthier than those from Dolly. In the case of ChatGPT, the responses were almost the same length or sometimes longer than those from human counselors. In terms of structure, ChatGPT often generated responses that were composed of multiple paragraphs as opposed to a big chunk of text usually generated by Dolly. Each paragraph corresponded to a certain theme and those different themes flowed well as a whole when we browsed through the different paragraphs altogether. Responses from these 3 venues not only differed in their style and length but also differed in their content. Dolly focused more on analytical aspects including questions on why a certain state or condition the patient is experiencing has arisen. On the other hand, responses from ChatGPT were more emotional with a focus on empathy and consolation. For instance, the responses to the 371th case in Table 4 illustrate these differences.

4 Discussion

This study sought to explore whether the recently released LLMs are effective enough in aiding human counselors or even partially replacing them to counsel patients struggling with mental health problems with a focus on depression, anxiety, and stress. In particular, Dolly was selected as the LLM of choice and Dolly was prompted to generate responses to patient questions from the CounselChat data.

Overall, Dolly generated responses seemed to be falling short in every aspect of our analyses compared to human counselor responses. They were, on average, shorter in length and contained less words than human counselor responses. They were also linguistically not similar to human counselors with an average cosine similarity score of 0.184 as cosine similarity scores closer to 0 indicate stronger dissimilarity. Regarding the breakdowns of different sentiments, responses from human counselors and Dolly generally shared a similar distribution pattern but responses from Dolly contained slightly higher proportions of negative sentiment.

It is worth noting that while these findings may imply that Dolly generated responses are not good enough yet to replicate human counselor responses and behavior, they do not necessarily suggest that LLMs cannot add value to mental health counseling. The true effectiveness of AI generated responses can be gauged through control trial experiments on real human subjects, in this case, patients suffering from depression, anxiety, and stress. This point is further elaborated in the following paragraphs discussing limitations of this study.

Nevertheless, strengths of using AI for mental health counseling and directions for improvement were also identified. This study revealed that AI responses are more consistent than those from human counselors in terms of their text length and number of words they contain. This is most likely due to the fact that therapists and counselors all have what they believe are most effective counseling methods and styles which lead to higher variability in various aspects of texts such as the text length. However, variability in responses across different counselors can mean more options available for patients to choose from. This means that variability and consistency in responses will be valuable for different cases. Therefore, the mental health counseling community should further investigate which area or application would benefit the most from this consistency of AI responses.

Moreover, AI for mental health counseling can be improved by transfer learning, a technique of training general purpose LLM with specific training data for the purpose of the user, thereby creating a customized LLM model built off of the original LLM (Golovanov et al. 2019). For example, Dolly could be trained on an additional corpus of human counselor conversations that are touted as effective and successful counseling which will reinforce Dolly's ability to generate responses to mental health related questions in a way that better align with various hallmarks of good counseling conversations.

This study comes with limitations. First, the number of data points amounted to only several hundred which may be enough for a pilot study but not adequate for a generalizable study with far-reaching implications. Researchers should collect a bigger dataset containing mental health related question and answer pairs, perform similar analyses, and examine if the results from this study remain the same.

Second, the findings are limited to Dolly, the LLM of choice in our study, and more generalizable insights will be unearthed only when the same analyses are conducted on the responses generated by a list of LLMs instead of one. Preferably, those responses will be compared against one another, not just against those from human counselors to help researchers better understand the strengths and weaknesses of different LLMs for mental health counseling. Some qualitative analysis between responses generated by Dolly and ChatGPT in this study already shed light to syntactical and content-wise disparities between those two different LLMs.

In addition, the analyses performed in this study could be made more granular by analyzing across 3 different types of mental health conditions–depression, anxiety and stress. While keeping only observations on one particular mental health condition from the CounelChat data would have allowed the study to be more focused, that would have resulted in an extremely small dataset where the results from the analysis would have become unnecessarily insignificant due to the small sample size.

Furthermore, future work may want to include interview studies that recruit actual participants experiencing anxiety, stress or depression and gauge if their mental and emotional state improves after being exposed to the AI generated counseling texts. This will allow researchers to gain qualitative insights to understanding how the AI generated counseling texts affect depressed, stressed or anxious patients. It may be advised to mask the fact that the texts were generated by AI because participants being aware of this fact may influence their perceptions towards the text themselves and thus may lead to misleading results. Researchers can also conduct a control trial experiment where the patients experiencing mental health problems are randomly assigned to control and intervention groups to see if there is statistically significant result in terms of the patients' ability to distinguish AI generated counseling texts from those of human counselors.

5 Conclusion

The primary objective of our project was to explore the potential of AI and Large Language models (LLMs) in providing counseling support, with the ultimate goal of enhancing the accessibility of counseling services. To assess the performance of an LLM, we conducted three analyses to compare its responses with those generated by human counselors.

The results of our experiment revealed that the responses generated by the LLM, named Dolly, exhibited poor performance across all aspects of our analyses when compared to the responses from human counselors. Nonetheless, it is worth noting that AI responses demonstrated a higher level of consistency in terms of text length and word count, unlike the variable responses from human counselors. In conclusion, we recommend future research endeavors that involve the inclusion of additional training data, such as interview studies with actual participants experiencing anxiety, stress, or depression. By assessing whether exposure to AI-generated counseling texts improves the mental and emotional state of these individuals, researchers can gain qualitative insights into the impact of AI-generated counseling texts on depressed, stressed, or anxious patients.

This expansion of research will contribute to a deeper understanding of how AI-generated counseling texts affect the well-being of individuals struggling with mental health issues. It holds significant implications for the field, shedding light on the potential benefits and limitations of AI in providing counseling support and paving the way for further advancements in the accessibility and effectiveness of counseling services.

6 References

About mental health. 2023. *Centers for Disease Control and Prevention*. https://www.cdc.gov/mentalhealth/learn/index.htm#:~:text=More%20than%2050%25%20will%20be,some%20point%20in%20their%20lifetime.&text=1%20in%205%20Americans%20will,illness%20in%20a%20given%20year.

Benson S. Ku, Jianheng Li, Cathy Lally, Michael T. Compton, and Benjamin G. Druss. Associations between mental health shortage areas and county-level suicide rates among adults aged 25 and older in the USA, 2010 to 2018. 2021. *General Hospital Psychiatry*, Volume 70: 44-50. ISSN 0163-8343.

Janet Cummings, Hefei Wen, and Benjamin G. Druss. Improving Access to Mental Health Services for Youth in the United States. *JAMA.* 2013;309(6):553–554.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko and Ho-Jin Choi, A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. 2017*. 18th IEEE International Conference on Mobile Data Management (MDM)*, Daejeon, Korea (South), pages 371-375.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin..Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks. 2023. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Nbertagnolli. counsel-chat. 2023. https://github.com/nbertagnolli/counsel-chat.

OpenAI. ChatGPT. 2023. https://chat.openai.com/chat.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Ronald D.Hester. 2017. Lack of access to mental health services contributing to the high suicide rates among veterans. *Int J Ment Health Syst* 11, 47.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-Scale Transfer Learning for Natural Language Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

Takeshi Kamita, Tatsuya Ito, Atsuko Matsumoto, Tsunetsugu Munakata, and Tomoo Inoue, A Chatbot System for Mental Healthcare Based on SAT Counseling Method. 2019. *Mobile Information Systems*, vol. 2019, Article ID 9517321, 2019, pages 11.

Tim Althoff, Kevin Clark, and Jure Leskovec; Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. 2016. *Transactions of the Association for Computational Linguistics* 2016; 4 463–476.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea.What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. 2019. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Appendices

Patient Question	I've gone to a couple therapy sessions so far and still everytime I walk in I get nervous and shaky. Is this normal? Should I still be feeling like this?
Human Counselor Response within the ChatCounsel Platform	Hi there, Thank you for your question. It's absolutely ok to feel nervous going to therapy. I have experienced anxiety going to see my own therapist. There can be a few reasons why you might feel this way. First, it is always unnerving to see a person who is a stranger and to share intimate things with that stranger. You mentioned it has only been a couple of sessions. Your anxiety might subside as you continue to see your therapist and grow more comfortable with him/her. Another reason why you might feel particularly nervous is perhaps you are not comfortable sharing things. As children, we might have bad experiences opening up to others. Someone might bully or ridicule us, and that experience can stay with us for a long time, making it extra scary to share our intimate feelings and thoughts with others. If this the case, as you continue with your sessions and have positive interactions with your therapists, this anxiety will subside with time and you will relearn that it is safe to share. And the last thought why you might feel the shakes is perhaps you know that you need to talk about some past experiences or memories that are uncomfortable and difficult. Regardless of the reason, it might help you to bring up your nervousness in your session and share how you feel with your therapist. Having an open dialogue about your anxiety with your therapist can help you resolve some of that anxiety and built greater trust with your therapist.
Dolly Response	There are many factors that contribute to nervousness or anxiety. From your anxiety symptoms going away after a session to your therapist not offering you a refund for not feeling any improvement after a few sessions. Your first step is to check in with yourself. Ask "why" am I feeling anxious? Are my relationships with my therapist and if so why are those relationships lacking trust? Ask yourself "how" can I heal my anxiety and truly have an experience that helps me heal? Ask yourself these questions and then consult the wisdom of your therapist and make a treatment plan that works best for you.
ChatGPT Response	"It's completely normal to feel nervous or shaky before therapy sessions, especially if you're still in the early stages of therapy. Many individuals experience anxiety or apprehension when starting therapy, as it involves opening up about personal thoughts and emotions. Your feelings of nervousness are valid, and it's important to remember that therapy can be an unfamiliar and vulnerable experience for many people.

	Over time, as you build a trusting relationship with your therapist and become more comfortable with the therapeutic process, these feelings of nervousness may gradually diminish. It's essential to communicate your concerns with your therapist, as they can provide support and reassurance tailored to your specific needs. Remember, therapy is a collaborative journey, and it takes time to develop a sense of ease and familiarity. Your therapist is there to help you navigate these emotions and create a safe space for healing and growth. With patience and persistence, you may find that your nervousness decreases as you continue attending therapy sessions and engage in the therapeutic process." It's crucial to note that while this response provides general guidance, each individual's experience may vary. A qualified therapist or counselor should assess the patient's specific circumstances to provide the most appropriate advice and support.

Table 4. Patient Question, Human Counselor Response, Dolly Response, and ChatGPT Response for Case ID = 371.

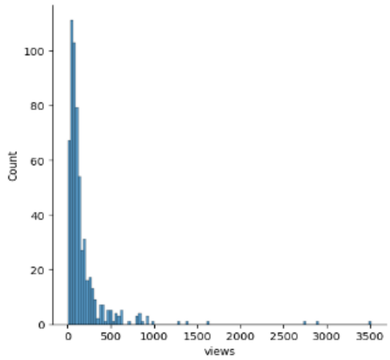

Figure 1. Distribution of Views.

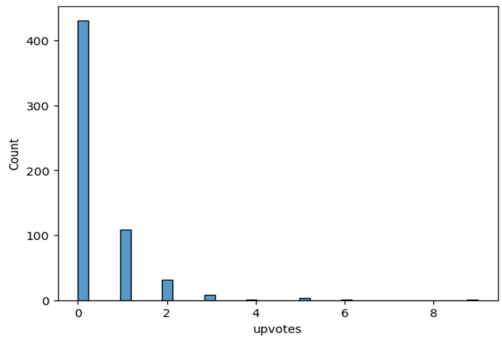

Figure 2. Distribution of Upvotes.