Report NIH AIM AHEAD

Inclusive Mental Health Counseling: Automated LLM Solutions on Social Media

Prepared by

Yifan Peng, Co-Principal Investigator

Ying Ding, Principal Investigator

David Hodge, Co-Principal Investigator

Yan Leng, Co-Principal Investigator

S. Craig Watkins, Co-Principal Investigator

September 16, 2024

**Overview:**

Mental health issues such as depression, anxiety, and stress are on the rise globally, yet access to effective therapeutic interventions remains limited. This study explores the potential of Large Language Models (LLMs) to provide scalable and accessible mental health support. Current LLMs often lack the empathy and sensitivity required for effective therapy. To address this, we developed an automated system using LLMs to generate therapeutic responses.

Our approach involves improving training data quantity and diversity with Generating Learning with Adaptive Nodes (GLAN) to create a comprehensive synthetic dataset covering various mental health concerns. We fine-tuned and applied DPO training to an open-source LLM (LLaMA-3) using this dataset. The model’s therapeutic responses were evaluated against established psychotherapeutic benchmarks and compared with those of GPT-4.

Results indicate that the trained LLaMA-3 model meets or exceeds current AI standards in creativity, directedness, perspective change, affirmations, sensitivity, and empathy. This study highlights the potential of AI in improving mental health care and provides a new method to address limited training data for future research and development.

1. **Introduction**

In recent years, mental health has emerged as a critical global issue, with increasing numbers of individuals experiencing conditions such as depression, anxiety, and stress (Collins et al., 2011) (Wainberg et al., 2017). The demand for effective therapeutic interventions has never been higher, yet access to mental health care remains limited for many. This gap presents an opportunity for innovative solutions, particularly through the use of Artificial Intelligence (AI). Large Language Models (LLMs) have shown promise in providing scalable and accessible mental health support. However, current AI models often fall short in generating responses that are truly empathetic, sensitive, and effective in a therapeutic context. Addressing these limitations is essential to harnessing the full potential of AI in mental health care.

Zhiyu et al., 2023 proposed the use of LLMs for computational psychotherapy by solving task of cognitive distortion detection using Diagnosis-of-Thought (DoT) prompting. This approach highlighted the use of instruction tuning in LLMs for subjectivity assessment, contrastive reasoning, and schema analysis for psychotherapy applications. Siyuan et al., 2023 used the EmpathyDialogues (ED) dataset to compare the responses generated by different LLMs using three empathy related metrics to show that LLMs respond in an empathic manner but still lacks the reliability desired in empathetic conversational systems. In Chain-of-Empathy (CoE) prompting (Yoon et al., 2023), various psychotherapy approaches were utilized to induce LLMs to reason about human emotional states. It was found Cognitive Behavioral Therapy (CBT) based CoE reasoning resulted in the most balanced generation of empathetic responses whereas LLMs without any reasoning generated exploratory responses. June et al., 2023 showed that open-source LLM fine-tuned on real conversations between consulting clients and professional psychologists achieved the performance level of ChatGPT on an evaluation benchmark consisting of seven psychotherapy factors. However, their reliance on recording real conversations between a patient and counsellor presents challenges in context of privacy, accessibility and security. Zhonghua et al., 2023 proposed an emotional support chat bot by firstly generating an emotional support dialogue dataset using ChatGPT and then fine-tuned LLaMA using this dataset. However, the sole reliance on generating the dialogue dataset using ChatGPT makes this approach susceptible to hallucination and implicit bias contained within the data generation model itself.

Existing methods for generating therapeutic responses using AI face several significant challenges. Firstly, the available training data is often too limited and small, failing to encompass the diverse range of mental health issues and therapeutic scenarios necessary for effective counseling. This lack of comprehensive data restricts the model’s ability to generalize and respond appropriately to various mental health concerns. Secondly, vanilla GPT models (OpenAI. et al.), while powerful, lack the specific information and nuances required for high-quality therapeutic practices. These models are designed for general-purpose use and do not inherently possess the sensitivity, empathy, and specialized knowledge needed for mental health counseling. Additionally, existing models often struggle with maintaining consistency in therapeutic tone and approach, leading to responses that may be inappropriate or unhelpful. Furthermore, the absence of fine-tuning on domain-specific data means that these models are not optimized to prioritize therapeutic effectiveness, resulting in responses that may lack the depth and personalization required for meaningful mental health support. Addressing these challenges is crucial for developing AI systems that can truly assist in mental health care.

1. **Research Aim**

The primary objective of this study is to develop an automated system leveraging Large Language Models (LLMs) to generate therapeutic responses for individuals struggling with mental health issues. This system aims to provide scalable, accessible, and effective mental health support by producing responses that are empathetic, sensitive, and aligned with therapeutic best practices.

To achieve this, we focus on several specific aims. First, we aim to improve the diversity and quantity of limited training data by utilizing GLAN (Haoran et al., 2024) to create a comprehensive synthetic dataset that encompasses a wide range of mental health concerns. This approach ensures that the training data is not limited to specific disorders but includes a broad spectrum of mental health issues. Second, we implement fine-tuning and DPO training (Rafael et al., 2023) on an open-source LLM (AI@Meta et al., 2024) using the generated dataset. This process refines the model’s ability to generate responses suitable for mental health counseling, ensuring they are both effective and empathetic. Finally, we evaluate the efficacy of the therapeutic responses generated by the trained model using established psycho-therapeutic benchmarks (Althoff et al., 2016) (Lambert et al., 2015), comparing the model’s responses with those generated by GPT-4 (OpenAI et al., 2023) to ensure that the new system meets or exceeds the standards of current AI models in terms of creativity, directedness, perspective change, affirmations, sensitivity, and empathy.

This report is organized into several key sections. The Methodology section outlines our approach, including the GLAN technique for data generation, fine-tuning, and DPO training of the LLaMA-3 model, and the evaluation metrics used. Implementation Details cover the technical aspects, such as dataset specifics, model configuration, and training parameters. The Results section presents the performance of the trained LLaMA-3 model, comparing it with GPT-4 using psychotherapeutic benchmarks. Finally, the Discussion section analyzes our findings, exploring implications, strengths, limitations, potential applications, and future research directions.

1. **Methodology**

To automate the LLM for the generation of the therapeutic responses, our approach will focus on generating sample data from a seed dataset which will then be used to fine-tune an open-source LLM (AI@Meta et al., 2024) and use DPO training (Rafael et al., 2023) on top of it. The fully trained model will then be able to produce the inclusive therapeutic responses for social media posts by people struggling with ill mental health. To evaluate the performance of our model, we will compare the responses generated by the trained model with the responses generated by GPT-4 (OpenAI et al., 2023) using evaluation benchmarks necessitating a sound conversation between a psychologist and mentally ill patient.

* 1. **Step 1: Generating Learning with Adaptive Nodes (GLAN)**

Generative Learning with Adaptive Nodes (GLAN) (Haoran et al., 2024) is an innovative technique designed to enhance the diversity of synthetic training data, thereby improving performance across various tasks. This approach is particularly valuable in scenarios where available training data is limited to a specific domain or fails to provide a comprehensive representation of the subject matter. For instance, in mental health research, if the existing dataset predominantly focuses on Obsessive-Compulsive Disorder (OCD), GLAN can be utilized to generate data encompassing a broader spectrum of disorders, such as Post-Traumatic Stress Disorder (PTSD).

GLAN operates by constructing a hierarchical tree of human knowledge. The process begins with the creation or adoption of a list of general topics intended to represent the entirety of human knowledge. Subsequently, each topic is recursively subdivided into its constituent sub-topics. In the original implementation, the terminal nodes of this tree are designated as "subjects," for which comprehensive syllabi are generated.

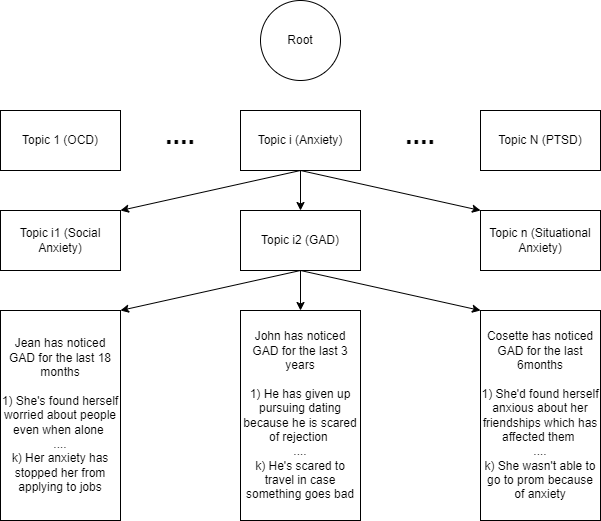


Figure 1: Generating data using GLAN

The original GLAN paper utilized this syllabus-based approach to generate homework instructions and corresponding answers using a Large Language Model (LLM). Our implementation adapts this methodology for mental health research. We initiated the process with a curated list of seed mental health topics to generate the knowledge tree. The terminal nodes (subjects) in our adaptation represent individual case studies, comprising an overview of the primary mental health concern and specific examples of its impact on the individual's life.

* 1. **Step 2: Fine-Tuning/DPO Training of open-source LLM**

Any open-source LLM trained on general data will not generate suitable therapeutic responses for mental health counseling by default. To refine the output of any such LLM, we need to fine-tune them on any dataset which depicts the practices adopted in mental health counseling. So, we use the data generated by GLAN (Haoran et al., 2024) for terminal nodes (subjects) since it contains questions as well as better/worse responses similar to what is exercised in psychological counseling. We then implement fine-tuning and DPO training (Rafael et al., 2023) on an open-source LLM (AI@Meta et al., 2024) to ensure that we can generate therapeutic responses which align with the therapy techniques implicitly mentioned in the data.

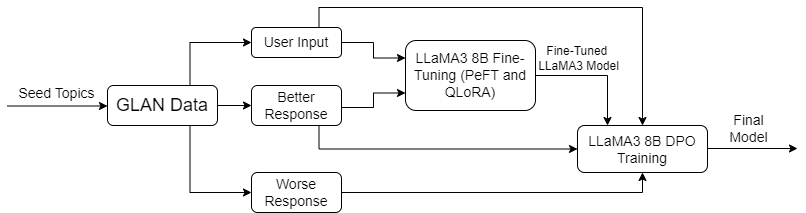
We use LLaMA-3 (AI@Meta et al., 2024) with 8 billion parameters as our open-source LLM, and implement fine-tuning using performance-efficient fine-tuning (PEFT) and QLora with 4-bit quantization. We use the instruction model of LLaMA-3 8B, so we needed to format our training data into a chat template consisting of a question and the better response as extracted from terminal nodes in the GLAN generated mental health dataset.

Figure 2: Supervised Fine-Tuning and DPO Training for LlaMA3 8B

After fine-tuning, the resulting model is used to implement direct preference optimization (DPO) training which further improves the fine-tuned model to learn the difference between better responses and worst responses. Again, to maximize efficiency, we used PEFT and QLoRA with 4-bit quantization to implement the DPO training with the chat template consisting of a question, better response, and worse response. The final model gives a therapeutic response to any input from a person sharing mental health problems.

* 1. **Step 3: Evaluation**

To evaluate the efficacy of the therapeutic responses generated by trained LLaMA3 8B (AI@Meta et al., 2024), we extract two evaluation benchmarks from psychological counseling literature. In the first literature (Althoff et al., 2016), the following psychotherapeutic factors are used to analyze the effectiveness of counseling conversations:

* **Creativity**: How unique this response is to the user’s situation
* **Directedness**: How direct the responses is in bringing up and discussing solutions to the problems
* **Perspective Change**: How much the response is focusing on being positive, thinking about others, or looking towards the future
* **Affirmations**: How much the response is affirming the patient’s experience

In another literature (Lambert et al., 2015), an evidence-based guide to practice effective psychotherapy counseling highlighted two important psycho-therapeutic factors:

* **Sensitivity**: How warm, accepting, and understanding the response is to the patient
* **Empathy**: How much empathetic the response is for the patient

To evaluate the effectiveness of the responses generated by the trained LLaMA3 8B, we ask GPT-4 (OpenAI et al., 2023) to generate the responses to the same set of questions. This helps us compare the responses by LLaMA3 8B with GPT-4 for each of the psycho-therapeutic factors highlighted in the two evaluation benchmarks.

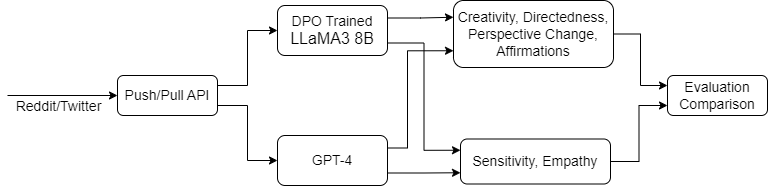


Figure 3: Evaluation/Comparison using GPT-4 and DPO trained LlaMA-3 8B

1. **Implementation**
   1. **CounselChat Dataset**

CounselChat (Bertagnolli et al., 2020), a platform where therapists respond to questions posed by clients. Users have the option to "like" responses that they find particularly helpful or insightful. To specify the seed topics for data generation, we used the categories of mental health problems specified on this platform which include the following:

A pie chart with text on it

Description automatically generated

* Depression
* Relationships
* Family
* Romantic
* Trauma
* Anger-Management
* Addiction
* Sexuality
* Behavioral issues

The overall structure of data looks like the following:

A screenshot of a computer screen

Description automatically generated

* 1. **Reddit Dataset**

Data from Reddit will be used to understand the context and expression of mental health problems as they appear in social media discourse. PullPush Reddit API allows us to search comments and submissions by keywords related to mental health challenges. This data will help evaluate our LLM’s performance in a real-world scenario and adapt our models to the informal and diverse ways in which users express mental health concerns on social media.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 4: Search comments and submissions by keywords

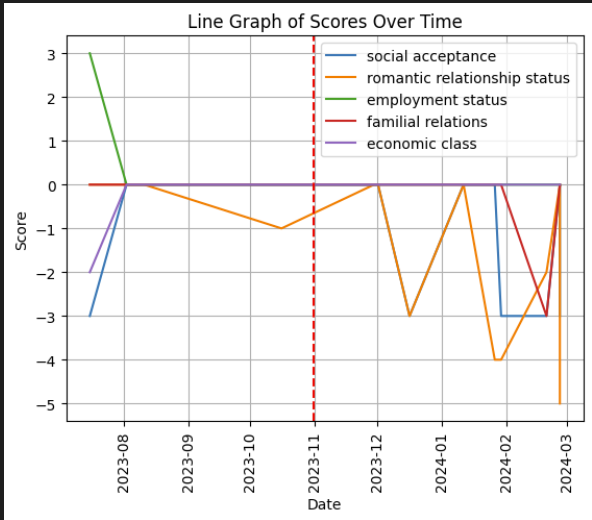
We also identified how various social determinants like gender, employment, and social acceptance impact mental health, including the current post and the history. This gave us an idea on whether we can use this dataset for evaluation purposes or not.

Figure 5: Scores on a scale of [-5,5] based on the adversity (<0)

* 1. **GLAN Data Generation**

Our implementation to generate GLAN (Haoran et al., 2024) based synthetic dataset leverages existing data from CounselChat (Bertagnolli et al., 2020). The process involves two primary steps:

1. For each subject, we prompt the LLM (GPT-3.5-Turbo) (OpenAI et al., 2023) with a sample post from CounselChat, instructing it to generate a new, subject-specific post.
2. In a separate request, we task the LLM (GPT-3.5-Turbo) (OpenAI et al., 2023) with generating both high-quality and low-quality responses to the generated question, simulating varied therapist interventions.

This approach allows for the creation of a diverse, synthetic dataset that encompasses a wide range of mental health concerns and potential therapeutic responses. Using this approach, we generated a total of 2048 samples of the dataset including questions, better responses, and worse responses.

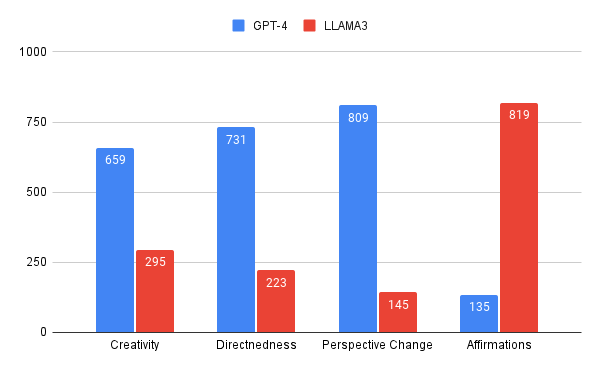
* 1. **Model Fine-Tuning/DPO Training**

We used the open-source model Meta-Llama-3-8B-Instruct (AI@Meta et al., 2024) from the official HuggingFace repository of Meta (meta-llama) to implement the fine-tuning due its suitability for chat-like interface as desired in mental health counseling applications. To fit the model within a single NVIDIA RTX A6000 GPU, we implemented parameter-efficient fine-tuning (r=64, alpha=16, dropout=0.1) and 4-bit quantization using QLoRA. Out of 2048 samples, 90\% were used as training data and 10\% were used as validation data during fine-tuning. The final training was implemented using a train/eval batch size of 4 for a total of 3 epochs which took ~2 hours to complete the fine-tuning process.

After the fine-tuning, the same parameters were used for PEFT and QLoRA to implement DPO training (Rafael et al., 2023) on the fine-tuned model where we used “better” and “worse” responses along with the relevant questions as the training data. For DPO, train/eval batch was set to 1 for a single epoch along with the following training parameters: loss=sigmoid, beta=0.1. The DPO training procedure took around ~4 hours to complete.

1. **Results**

To evaluate the efficacy of the therapeutic responses generated from the trained LLaMA3 8B model (AI@Meta et al., 2024) and compare it against the responses generated from GPT-4 (OpenAI et al., 2023), we use the CounselChat (Bertagnolli et al., 2020) dataset as the evaluation dataset and employ the psychotherapeutic factors highlighted in two evaluation benchmarks (Althoff et al., 2016) (Lambert et al., 2015).

Firstly, we extract all the unique questions from the CounselChat dataset which are 954 in total. For each of these questions, we generated the response from the LLaMA3 trained model as well as the GPT-4 using standard prompting. The question and resulting responses from both the models (LLaMA3, GPT-4) are then passed to GPT-4 along with the psychotherapy factors, and GPT-4 is prompted to rank both responses individually in the context of each psychotherapy factor.

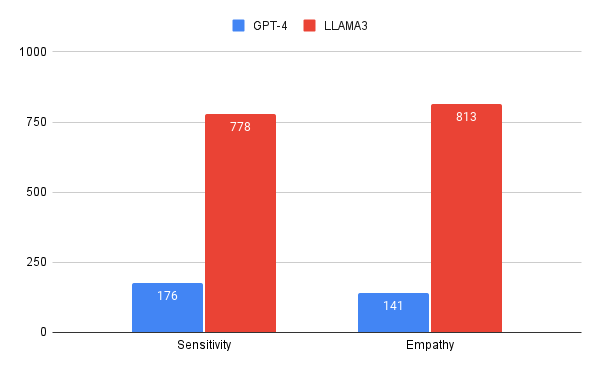
Figure 1: First Evaluation Benchmark (Althoff et al., 2016)

Figure 2: First Evaluation Benchmark (Lambert et al., 2015)

In the first evaluation benchmark, our trained LLaMA3 8B model has improved on the affirmations, but it lacks in other metrics i.e., creativity, directedness and perspective change. However, our model shows improved performance on both the factors for the second evaluation benchmark. Overall, if we calculate the cumulative top ranks of both the models for each of the six psychotherapy factors, our trained LLAMA3 8B achieves the top rank in 3073 questions whereas GPT-4 does so for 2651 questions. Thus, the trained LLAMA3 8B performs overall 16\% better than GPT-4 on the combined psychotherapy factors for both the evaluation benchmarks. These results exhibit the psychotherapeutic relevance of the data generated using GLAN even though we prompted GPT-3.5-Turbo to generate the dataset using which we fine-tuned LLaMA3 8B model and then subsequently implemented DPO training (Rafael et al., 2023).

1. **Conclusion and Future Work**

In conclusion, this study demonstrates the potential of leveraging LLM to generate therapeutic responses for individuals struggling with mental health issues. By employing the GLAN (Haoran et al., 2024) technique to create a diverse and comprehensive synthetic dataset, and refining the model through fine-tuning and DPO training (Rafael et al., 2023), we have developed a system capable of producing empathetic, sensitive, and effective therapeutic responses. Our evaluation, using established psychotherapeutic benchmarks, shows that the trained LLaMA-3 (AI@Meta et al., 2024) model meets or exceeds the standards set by current AI models like GPT-4 (OpenAI et al., 2023). This work highlights the promise of AI-driven solutions in enhancing access to mental health support and underscores the importance of continued research and development in this critical field.

For future work, to generate the dataset using GLAN, we used GPT-3.5-Turbo which resulted in less intelligent responses and low word count as compared to the responses generated via GPT-4. Moreover, the prompt to generate the data for fine-tuning/DPO training of LLaMA3 8B and getting raw responses from GPT-4 for relevant comparison did not involve evaluation metrics which we later use to evaluate/compare the responses received from both the models which resulted in unfair data generation and subsequent evaluation. By incorporating the aforementioned challenges within our framework and generating more training data using GLAN (~10000 samples), we intend for a more robust and fair evaluation of our trained model in our future works and experiments.

1. **References**

[1] Wainberg, Milton L., et al. "Challenges and opportunities in global mental health: a research-to-practice perspective." Current psychiatry reports 19 (2017): 1-10.

[2] Collins, Pamela Y., Vikram Patel, Sarah S. Joestl, Dana March, Thomas R. Insel, Abdallah S. Daar, Isabel A. Bordin et al. "Grand challenges in global mental health." Nature 475, no. 7354 (2011): 27-30.

[3] Benton, T.D., Boyd, R.C. and Njoroge, W.F., 2021. Addressing the global crisis of child and adolescent mental health. JAMA pediatrics, 175(11), pp.1108-1110.

[4] Patel, V., Minas, H., Cohen, A. and Prince, M. eds., 2013. Global mental health: Principles and practice. Oxford University Press, USA.

[5] Stade, E.C., Stirman, S.W., Ungar, L.H., Boland, C.L., Schwartz, H.A., Yaden, D.B., Sedoc, J., DeRubeis, R.J., Willer, R. and Eichstaedt, J.C., 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Mental Health Research, 3(1), p.12.

[6] Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y.H., Zhou, P., Moran, L.V., Ananiadou, S. and Beam, A., 2024. Large language models in mental health care: a scoping review. arXiv preprint arXiv:2401.02984.

[7] Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y. and Wang, Z., 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. arXiv preprint arXiv:2307.11991.

[8] Saeidnia, H.R., Hashemi Fotami, S.G., Lund, B. and Ghiasi, N., 2024. Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. Social Sciences (2076-0760), 13(7).

[9] Graham, S., Depp, C., Lee, E.E., Nebeker, C., Tu, X., Kim, H.C. and Jeste, D.V., 2019. Artificial intelligence for mental health and mental illnesses: an overview. Current psychiatry reports, 21, pp.1-18.

[10] Siyuan, B.L, and Aravind, S.R, 2023. Harnessing Large Language Models' Empathetic Response Generation Capabilities for Online Mental Health Counselling Support. arXiv preprint arXiv:2310.08017

[11] Yoon, K.L., Inju, L., Minjung, S., Seoyeon, B., Sowon, H., 2023. Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models, Korean Journal of Cognitive Science. 2024, Vol. 35 Issue 1, p23-48. 26p

[12] June, M.L., Donghao, L., He, C., Tianhe, R., Zeyi, L., Jiamin, W. 2023. ChatCounselor: A Large Language Models for Mental Health Support, arxiv preprint arXiv:2309.15461

[13] Zhonghua, Z., Lizi, L., Yang, D., Liqiang, N. 2023. Building Emotional Support Chatbots in the Era of LLMs, arxiv preprint arXiv:2308.11584

[14] Qiang, Z., Jason, N., Yusuke, M., 2023. Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models, arxiv preprint arXiv:2305.17878

[15] Zhiyu, C., Yujie, L., William, Y.W., 2023. Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting, arxiv preprint arXiv:2310.07146

[16] Haoran, L., Qingxiu, D., Zhengyang, T., Chaojun, W., Xingxing, Z., Haoyang, H., Shaohan, H., Xiaolong, H., Zeqiang, H., Dongdong, Z., Yuxian, G., Xin, C., Xun, W., Si-Qing, C., Li, D., Wei, L., Zhifang, S., Benyou, W., Wai, L., Furu, W., 2024. Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models, arxiv preprint arXiv:2402.13064

[17] Althoff T., Clark, K., and Leskovec, J., 2016, Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. Transactions of the Association for Computational Linguistics, 4, 463.

[18] Lambert, M. J., 2015, The therapeutic alliance: An evidence-based guide to practice. Psychotherapy Research, 26(2), pp. 259–261. doi: 10.1080/10503307.2015.1031200.

[19] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

[20] AI@Meta, Llama 3 Model Card, 2024, https://github.com/meta-llama/llama3/blob/main/MODEL\_CARD.md

[21] Bertagnolli, N., 2020, Counsel chat: Bootstrapping high-quality therapy data, Towards Data Science. https://towardsdatascience. com/counsel-chat~…

[22] Rafael, R., Archit, S., Eric, M., Stefano, E., Christopher, D.M., Chelsea, F., 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Thirty-seventh Conference on Neural Information Processing Systems