Synthesized Strategy for Mental Health Support

Stanford CS224N Custom Project

Evelyn Song
Stanford School of Medicine
evsong@stanford.edu

David Yuan
Stanford University
davidy02@stanford.edu

Abstract

Addressing the growing global burden of mental health disorders necessitates innovative and accessible support solutions. In this study, we present a novel approach to mental health support utilizing advanced Natural Language Processing techniques. By synthesizing existing datasets and implementing innovative methodological strategies, we refine and enhance the capabilities of language models for empathetic and supportive dialogue generation. Our contributions include the development of a tailored dataset with preserved features such as strategy tags, alongside a pioneering methodology involving explicit strategy sampling during model training. Despite computational limitations, our experiments demonstrate promising results, showcasing significant improvements in model performance compared to existing approaches. We envision our strategies as catalysts for future advancements in mental health support technology, with potential applications extending beyond current limitations to achieve human-level performance. Through continued collaboration and research, we strive to revolutionize the landscape of mental health support, empowering individuals worldwide to access the assistance they need to thrive.

1 Introduction

The prevalence of mental health issues, encompassing conditions like depression and anxiety disorders, has emerged as a significant societal challenge in recent years. Studies estimate that around 20% of the global population grapples with a form of mental health disorder (Holmes et al., 2018). Despite the growing demand for support, a substantial portion of those affected struggle to access mental health services due to various barriers, including financial constraints and limited availability (Olfson, 2016). In response to this pressing need, there has been a surge in the development of therapy chatbots, leveraging advancements in language models to offer innovative mental health support solutions. These chatbots present a cost-effective and easily accessible alternative to traditional therapy, positioning them as a promising initial intervention tool for addressing mental health concerns.

However, existing Natural Language Processing (NLP) techniques encounter challenges in effectively addressing the complexities inherent in mental health support. Many current models are tailored to specific aspects of mental health care, such as promoting positive reframing or enhancing empathy. While these focused approaches are valuable, they often fail to capture the multifaceted nature of mental health counseling, which demands a comprehensive understanding and a nuanced response to each individual's unique condition.

Our research endeavors to fill this gap in the literature by refining and expanding upon existing datasets while innovating methodological approaches to develop a more holistic and effective model for mental health support. Specifically, we embarked on the synthesis of well-known datasets in the field, preserving their characteristic features such as strategy tags. Additionally, we propose a novel methodology that involves fine-tuning a language model with explicit strategy sampling, while concurrently training another model to sample strategies. This innovative approach aims to enhance the model's ability to generate empathetic and supportive responses tailored to the diverse needs of individuals seeking mental health support.

2 Related Work

2.1 Fostering Empathy

The fusion of NLP and mental health support has catalyzed notable advancements in recent years. One pioneering study by Sharma et al. (2020) introduced a computational methodology for quantitatively assessing empathy in text-based mental health interactions. Emphasizing empathy's pivotal role in effective care, they employed a multi-task RoBERTa-based model to detect empathic expressions within online dialogues. Their findings underscore the potential for enhancing empathy in mental health support through training and feedback mechanisms for peer supporters. This work signifies the critical influence of empathy in elevating the quality of assistance rendered by chatbots and AI-driven platforms.

2.2 Emotional Support

In a distinct yet complementary effort, researchers from Tsinghua University introduced the Emotional Support Conversation (ESC) task, inspired by Helping Skills Theory. Their objective was to imbue dialogue systems with the ability to offer emotional support Liu et al. (2021). They crafted the Emotional Support Conversation dataset (ESConv) with detailed annotations that shed light on the support strategies employed between help-seekers and supporters. The construction of this dataset involved meticulous training for supporters and stringent quality control throughout the data collection process. Their findings accentuate the necessity of structured emotional support and the strategic application of diverse support mechanisms for creating more empathetic dialogue systems.

2.3 Positive Reframing

Ziems et al. (2022) introduces the concept of positive reframing, distinct from sentiment reversal, aiming to generate text that shifts a negative viewpoint to a positive one while preserving the original message's intent. This involves a complex text style transfer task, challenging due to the necessity of maintaining semantic integrity. To advance research in this domain, the authors present

POSITIVE PSYCHOLOGY FRAMES, a benchmark comprising 8,349 paired sentences annotated with 12,755 instances across six strategies derived from positive psychology theories. This dataset facilitates the exploration of the task's feasibility and the development of models capable of executing positive reframing without altering the fundamental meaning of the text. The paper evaluates various advanced text style transfer models against this dataset, highlighting significant challenges in achieving meaningful positive reframes and suggesting future research directions to improve psychological well-being and cognitive performance through linguistically positive adjustments.

2.4 Language Models

Language models play a pivotal role in natural language processing (NLP) tasks, serving as the foundation for various applications, including machine translation, text generation, and sentiment analysis. These models aim to understand and generate human-like text by learning the statistical patterns and structures present in large corpora of text data. Recent advancements in deep learning techniques have led to the development of large language models that exhibit impressive capabilities in understanding and generating natural language.

Large language models, characterized by their vast size and extensive training data, have garnered significant attention in recent years due to their remarkable performance across a wide range of NLP tasks. Models such as OpenAI's GPT-4 (OpenAI et al., 2024) have demonstrated state-of-the-art results in tasks such as language understanding, text generation, and question answering. These models leverage transformer architectures (Vaswani et al., 2017), which enable them to capture long-range dependencies and contextual information effectively.

Given the computational resources required to train and fine-tune large language models, researchers often rely on pre-trained models that have been trained on vast amounts of text data. Fine-tuning these pre-trained models on domain-specific datasets allows researchers to adapt the models to specific tasks while leveraging the knowledge learned during pre-training.

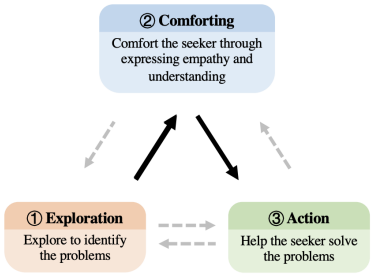
In this study, we leverage large language models for text generation and classification tasks, specifically utilizing DialoGPT as a text generation model and BERT as a classification model. DialoGPT, a variant of the GPT series developed by OpenAI, excels in generating contextually coherent responses in conversational settings (Zhang et al., 2019). On the other hand, BERT, developed by Google, is renowned for its effectiveness in various NLP tasks, including text classification, question answering, and named entity recognition (Devlin et al., 2018).

Our choice of using DialoGPT and BERT in this study is motivated by their proven performance in their respective tasks and their availability as pre-trained models. Given the computational constraints associated with training large language models, we opt to utilize pre-trained models and fine-tune them on domain-specific datasets to achieve our research objectives.

3 Approach

3.1 Mental Health Support Framework

Advances in cognitive-behavioral therapy (CBT) and psychotherapy research within clinical psychology have elucidated the nuanced processes underlying effective therapeutic interactions, particularly in sentiment analysis and intervention stages pertinent to the development of mental health support systems, such as chatbots. Drawing on a Unifying Theory for mental health assistance, as proposed by leading scholars in the field (Hill, 2009; DiGiuseppe et al., 2019), our methodology categorizes the training texts for a NLP mental health chatbot into three distinct stages: Comforting, Exploration, and Action, as illustrated in Figure 1. This classification is rooted in an integrative approach that leverages client-centered, psychoanalytic, and cognitive-behavioral theories to facilitate a comprehensive exploration of helping skills. By aligning the chatbot's interaction framework with these empirically supported stages, we aim to emulate the dynamic and adaptive process of therapeutic change, acknowledging the critical interplay of affect, cognition, and behavior in fostering client growth and resolution (Hill, 2009). This methodological foundation not only underpins the chatbot's design but also ensures that its interventions are theoretically grounded and tailored to effectively support users through their journey of self-exploration and problem-solving.



Strategies	Stages			Examples	Lexical Features
Question				*Can you talk more about your feelings at that time?*	do you (15.0), are you (13.8), how (13.7), what (12.3), do (11.5)
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*	is that (8.2), so you (8.2), it sounds (7.1), correct (7.1), so (6.6)
Reflection of Feelings				*I understand how anxious you are.*	can tell (7.4), understand how (5.8), are feeling (5.1), tell (5.1), understand (4.9)
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*	my (15.3), was (10.5), me (10.2), had (9.7), myself (7.8)
Affirmation and Reassurance				*You've done your best and I believe you will get it!*	its (5.7), thats (5.6), will (5.4), through this (5.1), you will (4.7)
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*	maybe (7.3), if (6.5), have you (6.4), talk to (5.8), suggest (5.8)
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*	there are (4.4), will (3.8), available (3.7), seen (3.3), possible (3.3)
Others				*I am glad to help you!*	welcome (9.6), hope (9.6), glad (7.3), thank (7.0), hope you (6.9)

Figure 1: Annotation Dimensions of the Emotional Support Framework.

3.2 Dataset

Our study utilizes a comprehensive dataset comprising two distinct components, each serving a unique purpose in analyzing peer-to-peer mental health support mechanisms.

1. *Mental Health Subreddits*: This dataset is compiled from 55 subreddit communities focused on various aspects of mental health (Sharma and De Choudhury, 2018). It comprises 1.6 million threads and 8 million interactions, reflecting the diverse and vibrant discussions within the mental health community on Reddit (Sharma et al., 2020).

2. *Emotional Support Conversation Dataset (ESConv)*: Developed by researchers at Tsinghua University (Liu et al., 2021), the ESConv dataset offers a richly annotated collection of dialogues within a seeker and supporter framework. Unlike the previous datasets characterized by single-turn interactions or brief conversations, ESConv features longer dialogues that more closely simulate real-life therapeutic conversations.

Datasets 1 has previously been annotated by researchers Sharma et al. (2020) to identify three critical stages of empathy within the context of online mental health support: emotional reactions, interpretations, and explorations. In parallel, Dataset 2 has been annotated by (Liu et al., 2021). to capture various dimensions of support, including exploration, comforting, and action planning.

Given the fact that the different datasets compose of different sets of therapy strategies for each interaction, we managed to find a method to unify those strategy tags so as to combine the two datasets. Using the Unifying Theory framework for mental health support discussed above, we have leveraged and synthesized the datasets 1 based on the corresponding strategy map, as shown below. Our ground-truth corpus contains 28,505 conversations between a seeker and a supporter with annotated labels from trained crowdworkers, derived from Sharma et al. (2020) and Liu et al. (2021). A small proportion of data fail to fall into any category, which we annotated as "Others".

Table 1: Dimensions Of Support and Their Components

Category	**Components**	**Count**
Comforting	Emotional Reactions, Reflection of Feelings, Affirmation and Reassurance, Self-Disclosure	10,429
Exploration	Explorations, Question, Restatement or Paraphrasing, Reflection of Feelings, Self-Disclosure	10,445
Action	Interpretations, Self-Disclosure, Affirmation and Reassurance, Providing Suggestions, Information	11,674
Others	Others	3,661

3.3 Models

Our choice of models centers around DialoGPT, a state-of-the-art conversational model known for its exceptional ability to engage in meaningful dialogue while maintaining a manageable size. This model strikes a balance between complexity and efficiency, making it well-suited for our application in mental health support.

To refine DialoGPT for our specific task, we leverage the synthesized dataset mentioned previously. Through fine-tuning, we tailor DialoGPT to better understand and respond to the nuances of mental health conversations. This fine-tuning process ensures that the model is not only proficient in general dialogue but also adept at providing empathetic and supportive responses in the context of mental health therapy.

To validate the efficacy of our synthesized dataset, we compare the performance of the fine-tuned model trained on our dataset with that trained on the original ESConv dataset. This comparative analysis allows us to assess the impact of our dataset on the model's ability to generate appropriate and empathetic responses in mental health support scenarios.

In addition to fine-tuning, we introduce the feature of strategy sampling to further enhance the model's effectiveness. Strategy sampling involves training the DialoGPT models with an explicit strategy tag and incorporating a BERT classifier model that predicts the current strategy. During inference, we sample the strategy using the classifier based on the context provided by the user. This allows us to generate responses tailored to the specific emotional and therapeutic needs of the individual seeking support.

We compare different strategy sampling methods, including no sampling (pure LM), random sampling (classifier is random), and a fine-tuned classifier sampler. This comparative analysis helps us identify the most effective strategy sampling approach for enhancing the model's performance in mental health support dialogues.

3.4 Training

Our training process involved fine-tuning four separate DialoGPT models to cater to different experimental conditions. Firstly, we fine-tuned one model using the ESConv dataset and another model using our synthesized dataset. Additionally, we fine-tuned two more models by incorporating a strategy tag into the training process, one using the ESConv dataset and the other using our dataset. Each fine-tuning session consisted of training on top of the pre-trained model for four epochs, a duration dictated by our computational constraints.

In addition to fine-tuning the DialoGPT models, we trained a BERT classifier model using our dataset with four classes corresponding to different dialogue strategies. Despite facing computational limitations and time constraints, we trained the classifier model for three epochs. The result model achieved an accuracy of 45

The rationale behind training multiple DialoGPT models lies in the need to compare the effectiveness of different datasets and training strategies. By fine-tuning separate models using the ESConv dataset and our dataset, we can assess the impact of dataset choice on the model's performance. Furthermore, incorporating a strategy tag into the training process allows us to explore the effectiveness of explicitly modeling dialogue strategies in enhancing the model's conversational capabilities.

Training the BERT classifier model complements the fine-tuning of the DialoGPT models by providing a mechanism for predicting dialogue strategies. Despite facing constraints in computational resources and time, the achieved accuracy and F1 score indicate the model's potential utility in guiding the selection of dialogue strategies during inference. This preliminary training phase lays the groundwork for further analysis and experimentation to optimize the model's performance and explore its full potential in mental health support applications.

3.5 Evaluation

Our evaluation methodology encompasses both automatic evaluation metrics and human evaluations. Automatic evaluations serve as a baseline sanity check, providing quantitative measures of the model's performance. In particular, we focus primarily on the BLEU metric, which calculates the overlap between the generated responses and reference responses based on n-gram matching. This

metric provides a straightforward way to assess the syntactic similarity between the model-generated responses and the ground truth, serving as a useful benchmark for evaluating the model's language generation capabilities.

Human evaluations, on the other hand, offer deeper insights into the model's effectiveness in real-world scenarios. In this process, participants are presented with responses generated by different models in response to the same prompts. They are then asked to compare the responses and choose the one that they find more helpful and supportive. This human-centric approach allows us to capture the nuances of human judgment and preference, providing qualitative feedback on the model's ability to provide empathetic and supportive responses in mental health support scenarios.

Through human evaluations, we gain a better understanding of how well the model aligns with the needs and expectations of individuals seeking mental health support. By comparing the responses generated by different models, participants can assess the model's ability to effectively address their emotional and therapeutic needs. This qualitative assessment complements the quantitative metrics obtained from automatic evaluations, providing a comprehensive evaluation of the model's performance from both technical and user-centric perspectives.

4 Experiments and Analysis

4.1 Dataset and Strategy Tags

In our experiment, we fine-tuned four separate DialoGPT models using different combinations of datasets and training strategies: ESConv dataset, our synthesized dataset, and the incorporation of an explicit strategy tag. The performance of each model was evaluated using the BLEU metric, which measures the similarity between the model-generated responses and reference responses.

The BLEU scores of the trained models are summarized in Table 4.1. It is evident that incorporating our synthesized dataset, especially when combined with the explicit strategy tag, leads to a notable improvement in BLEU score compared to the baseline pre-trained model and models trained solely on the ESConv dataset. Particularly, the model trained with our dataset and the explicit strategy tag achieved the highest BLEU score of 0.300, indicating superior performance in generating responses that closely match reference responses.

Moreover, Table 6 provides a comparison of our model against other configurations in a human evaluation setting, highlighting the percentage of wins and losses in pairwise comparisons. Remarkably, our model trained with our dataset and the explicit strategy tag outperforms other configurations in a significant percentage of comparisons, demonstrating its effectiveness in generating more coherent and contextually relevant responses.

The results underscore the importance of dataset selection and the incorporation of explicit strategy tags in training dialogue generation models. Our synthesized dataset proves to be advantageous over existing datasets, likely due to its tailored focus on mental health support scenarios. Additionally, explicitly modeling dialogue strategies enhances the model's ability to generate responses that align with the intended therapeutic objectives, leading to improved performance.

Table 2: BLEU Scores of DialoGPT Models

Model	**Fine-tuned**	**Our Dataset**	**Strategy Tag**	**BLEU Score**
Pre-trained Baseline				0.073
ESConv	✓			0.226
Our Dataset	✓	✓		0.229
ESConv + Strategy Tag	✓		✓	0.293
Our Dataset + Strategy Tag	✓	✓	✓	**0.300**

Table 3: Comparison of **DialoGPT model trained with our dataset and explicit strategy tag** against other models (Human evaluation)

vs.	**ESConv**	**Our Dataset**	**ESConv + Strategy Tag**
Win(%)	68	65	63
Loss(%)	32	35	37

4.2 Strategy Sampling

In this subsection, we explore the impact of different strategy sampling methods on the performance of our models in generating empathetic responses tailored to mental health support scenarios. We compare the effectiveness of three sampling methods: no sampling, random sampling, and classifier sampling.

Table 4.2 presents the BLEU scores of our models trained with different sampling methods. Remarkably, incorporating classifier sampling leads to the highest BLEU score across both datasets. This suggests that explicitly modeling dialogue strategies during inference improves the model's ability to generate responses that closely match reference responses.

Furthermore, Table 5 provides insights into the human evaluation of models trained with different sampling methods. The percentage of wins and losses in pairwise comparisons reveals that classifier sampling consistently outperforms random sampling and no sampling. Participants consistently preferred the responses generated by models trained with classifier sampling, highlighting the effectiveness of this sampling method in producing more helpful and supportive responses.

Overall, these findings underscore the importance of strategy sampling in enhancing the performance of dialogue generation models for mental health support. By explicitly incorporating dialogue strategies into the inference process, we can ensure that the model generates responses that are aligned with the intended therapeutic objectives, ultimately improving the quality of support provided to individuals in need.

Table 4: BLEU scores of our models with different sampling methods

Model	**Our Dataset**	**Random**	**Classifier**	**BLEU Score**
ESConv,No Sampling				0.293
ESConv,Random		✓		0.271
ESConv,Classifier			✓	0.313
Our Dataset,No Sampling	✓	✓		0.300
Our Dataset,Random	✓			0.278
Our Dataset,Classifier	✓		✓	**0.331**

Table 5: Comparison of the models trained with different sampling methods (Human evaluation)

vs.	**No Sampling**		**Random Sampling**	
%	Win	Loss	Win	Loss
Random Sampling	46	54		
Classifier Sampling	**75**	25	**77**	23

4.3 Strategy Sampling on Large Language Models

In this experiment, we explore the effectiveness of in-context learning with GPT-3.5 for generating supportive responses in mental health dialogues. Specifically, we compare the responses generated by a GPT-3.5 agent assuming the role of a supporter with and without the guidance of our classifier for prompting explicit strategies. Additionally, we evaluate the performance of another popular method, Chain-Of-Thoughts (COT) prompts, in generating supportive responses.

To conduct the experiment, we tasked the GPT-3.5 agent with completing sampled conversations from the validation set, simulating real-world interactions with individuals seeking mental health support. For each conversation, we provided no additional prompts, prompts generated using the COT method, and prompts generated by our classifier sampling strategy. We then collected human evaluations to assess the quality and effectiveness of the generated responses.

Overall, given the inherent strength of GPT-3.5 as a language model, the responses generated were of good quality across all conditions. However, upon closer examination, we observed distinct differences in the nature and depth of the responses based on the prompting method used.

The responses generated with our classifier-sampled strategy prompts tended to be longer and more in-depth, delving into the nuances of the conversation topics and offering comprehensive support. This finding aligns with our hypothesis that explicitly guiding the model with predefined strategies

would lead to more tailored and substantive responses, enhancing the overall quality of the support provided.

Conversely, the COT method yielded slightly less favorable results, with the model often providing overly comprehensive but shallow responses. While the model demonstrated a breadth of knowledge and understanding, it struggled to maintain coherence and depth in its responses, potentially due to the lack of focused guidance provided by the prompts.

To quantify the preferences of human evaluators, we conducted pairwise comparisons between the responses generated with and without our classifier sampling strategy. As shown in the table, the results revealed a preference for the responses generated with the classifier sampling strategy, with a majority favoring the longer and more substantive responses produced by this method.

In summary, while all methods produced satisfactory responses, our classifier sampling strategy emerged as the preferred approach for guiding GPT-3.5 in generating supportive responses in mental health dialogues. By providing explicit strategies, we enable the model to produce more tailored and comprehensive support, ultimately enhancing the overall effectiveness of the interaction.

Table 6: Comparison of **GPT-3.5 model with classifier strategy prompting** against no additional prompting and COT (Human evaluation)

vs.	**Vanilla**	**Chain-Of-Thoughts**
Win(%)	55	65
Loss(%)	45	35

5 Discussion and Conclusion

In this paper, we have addressed the pressing need for more effective mental health support solutions by refining existing datasets and proposing innovative methodological approaches. Our contributions include the synthesis of well-known datasets in the field, while preserving their characteristic features such as strategy tags. Additionally, we have introduced a novel methodology involving fine-tuning language models with explicit strategy sampling, aimed at enhancing the generation of empathetic and supportive responses tailored to individual mental health needs.

However, our work is not without limitations. Due to computational and time constraints, we conducted our experiments on a smaller scale, focusing primarily on DialoGPT models. We recognize that our strategies hold the potential for even greater performance gains when applied to larger language models with extensive pretrained knowledge. We anticipate that future research utilizing such models could achieve performance levels approaching human-level responses in mental health support dialogues.

Furthermore, while we synthesized a selection of datasets to train our models, we acknowledge that this dataset compilation is not exhaustive. The limited number of datasets may constrain the generalizability of our findings to the broader landscape of mental health support. To address this limitation, we advocate for increased collaboration with medical institutions and research laboratories to gather more diverse and extensive datasets. By leveraging a wider range of data sources, future research can build more robust and comprehensive models for mental health support.

Moving forward, we envision a future where technology plays an increasingly integral role in providing effective and accessible mental health support. By continuing to refine our models and expand our dataset repositories, we aim to empower individuals worldwide to access the support and resources they need to lead healthier and more fulfilling lives. With continued dedication and collaboration, we remain committed to advancing technology-driven solutions for mental health support, ultimately contributing to a brighter and more inclusive future for all.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

R. DiGiuseppe, R. Venezia, and R. Gotterbarn. 2019. What is cognitive behavior therapy? In S. G. Little and A. Akin-Little, editors, *Behavioral interventions in schools: Evidence-based positive strategies*, 2 edition, pages 325–350. American Psychological Association.

Clara E. Hill. 2009. *Helping Skills: Facilitating, Exploration, Insight, and Action*, 3 edition. American Psychological Association.

Emily A Holmes, Ata Ghaderi, Catherine J Harmer, Paul G Ramchandani, Pim Cuijpers, Anthony P Morrison, Jonathan P Roiser, Claudi L H Bockting, Rory C O'Connor, Roz Shafran, Michelle L Moulds, and Michelle G Craske. 2018. The lancet psychiatry commission on psychological treatments research in tomorrow's science. *Lancet Psychiatry*, 5(3):237–286.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*, v1.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs (Millwood)*, 35(6):983–990.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu,

Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, v1.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Atlanta, GA, USA. ACM. Paper No.: 641.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. ArXiv:2204.02952v1.