Synthetic Empathy: Generating and Evaluating Artificial Psychotherapy Dialogues to Detect Empathy in Counseling Sessions

Daniel Cabrera Lozoya[1], Eloy Hernández Lúa[2], Juan Alberto Barajas Perches[2], Mike Conway[1], and Simon D'Alfonso[1]
[1]The University of Melbourne, Australia
[2]ITESM, Mexico
{dcabreralozo}@student.unimelb.edu.au
{jbperches, eloyhl}@exatec.tec.mx
{mike.conway, dalfonso}@unimelb.edu.au

Abstract

Natural language processing (NLP) holds potential for analyzing psychotherapy transcripts. Nonetheless, gathering the necessary data to train NLP models for clinical tasks is a challenging process due to patient confidentiality regulations that restrict data sharing. To overcome this obstacle, we propose leveraging large language models (LLMs) to create synthetic psychotherapy dialogues that can be used to train NLP models for downstream clinical tasks. To evaluate the quality of our synthetic data, we trained three multi-task RoBERTa-based bi-encoder models, originally developed by Sharma et al., to detect empathy in dialogues. These models, initially trained on Reddit data, were developed alongside EPITOME, a framework designed to characterize empathetic communication in conversations. We collected and annotated 579 therapeutic interactions between therapists and patients using the EPITOME framework. Additionally, we generated 10,464 synthetic therapeutic dialogues using various LLMs and prompting techniques, all of which were annotated following the EPITOME framework. We conducted two experiments: one where we augmented the original dataset with synthetic data and another where we replaced the Reddit dataset with synthetic data. Our first experiment showed that incorporating synthetic data can improve the F1 score of empathy detection by up to 10%. The second experiment revealed no substantial differences between organic and synthetic data, as their performance remained on par when substituted.

1 Introduction

Therapy transcripts offer rich insights into counseling sessions, capturing key details such as clients' concerns, emotional states, and therapeutic interventions (Lee et al., 2019; Imel et al., 2015). Natural language processing (NLP) models have shown great promise in analyzing these transcripts (Laricheva et al., 2024; Ewbank et al., 2020; Gaut et al., 2017). However, training such models demands substantial data, which is difficult to access due to the need to safeguard sensitive health information, and institutional barriers to obtaining clinical data (Lu et al., 2021; Aledavood et al., 2017).

Data Augmentation (DA) - a set of methods used for synthetic generation of training data - is a way to manage data scarcity when training machine learning models (Ansari and Saxena, 2024). The adoption and success of DA has mostly been in the computer vision field, whereas for NLP tasks it has exhibited a more limited impact when achieving performance gains (Maier Ferreira and Reali Costa, 2020). Traditionally, NLP-specific data augmentation approaches have relied on back-translation (Corbeil and Ghadivel, 2020) or performing simple operations to the original text, such as synonym replacements or random word insertion (Wei and Zou, 2019). However, performing simple transformations on existing text samples can lead to syntactic and semantic distortions of the text (Giridhara et al., 2019).

Generative language models have made a breakthrough in augmenting unstructured text data (Hagos et al., 2024). Models such as OpenAI's GPT series (OpenAI et al., 2024), have relied on sophisticated self-attention mechanisms to generate new data, rather than just performing local changes on the text. Related studies have started exploring the use of generative models for training few shot classifiers (Edwards et al., 2022), generating artificial text for enhancing intent classifiers (Sahu et al., 2022), and augmenting domain specific datasets to boost domain specific NLP tasks (Amin-Nejad et al., 2020). Although NLP models are an active area of research, the creation of synthetic datasets remains understudied in the mental health field.

In this research, we examined the viability of using LLMs for generating artificial counseling transcripts to enhance the performance of NLP models for clinical tasks. We trained the bi-encoder

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 157–171

model introduced in (Sharma et al., 2020), which is capable of recognizing empathetic dialogues and providing rationales that support its predictions. In therapy, empathy plays a crucial role and serves as a significant predictor of therapy outcomes (Elliott et al., 2018). It stands as one of the fundamental factors that contribute to establishing a strong working alliance between a psychotherapist and their client during a session, regardless of the specific therapeutic approach employed (Elliott et al., 2011). Research studies have shown that mental health professionals can enhance their empathetic responses through the provision of appropriate feedback (Benster and Swerdlow, 2020; Sharma et al., 2020). Hence, a model that can identify low empathetic dialogues could help therapists recognize areas where empathetic engagement could be improved. By leveraging the guidance and input provided by the model, therapists can refine their empathetic skills and give more supportive therapy sessions for their clients.

Our main contributions are as follows:

1. We present a methodology to generate and evaluate counselling transcripts for data augmentation purposes.
2. We demonstrate that our synthetic transcripts can effectively fine-tune state-of-the-art models, enabling them to surpass baseline models in text mining therapy transcripts.
3. We release the synthetic datasets used in this study to help future mental health research.

2 Related Work

The generation of realistic synthetic patient data has primarily concentrated on the production of electronic health records (EHRs). Before generative models were used for data augmentation purposes, many methods relied on rule-based methods (Ansari et al., 2021). Among the initial generative architectures used to augment EHR data, MedGan (Choi et al., 2017) introduced a generative adversarial network (GAN) designed to generate multi-label patient records. To improve the quality of the data generated by MedGan, medWgan and medBGan were developed (Baowaly et al., 2018). These models were based on the principles of Wasserstein GAN with gradient penalty (Gulrajani et al., 2017) and boundary-seeking GANs (Hjelm et al., 2018) respectively. It is worth noting that the previous models primarily concentrate on generating the structured data components of an EHR. Synthetic records frequently lack the inclusion of the unstructured text section, and when it is included, it is usually quite concise. For instance, in (Lee, 2018) their approach generates unstructured text that is limited to 18 tokens or less.

In addition to GANs, transformer-based models have also been used for medical data augmentation. Liu (2018) trained a transformer with memory-compressed attention to create EHRs, yielding promising results (1.76 in the perplexity per token and a 44.6 in the Rogue-1 metric). However, an evaluation to measure the data's quality for downstream task was not conducted. While previous research has primarily focused on augmenting data, limited attention has been given to evaluating its utility for training machine learning models. Wang et al. addressed this gap, in their study they employed synthetic data as supplementary training data for two biomedical NLP tasks: text classification and temporal relation extraction. Similarly, Lu et al. used a transformer-based model to train classifiers for patients readmission prediction.

While the majority of research has mostly concentrated on synthetic EHRs, there is also relevant work within the field of synthetic mental health data. One such example is found in (Ive et al., 2020), where they artificially generated discharge summaries from mental health providers. These summaries were utilized in a downstream NLP text classification task. Yet, there is still a scarcity of research focusing on the creation of synthetic data that mimics dialogues from therapy transcripts.

3 Method

3.1 Empathy Framework

To measure empathy in text-based conversations we used the EPITOME framework (Sharma et al., 2020), which establishes the following empathy dimensions:

1. **Emotional reactions** - entails the therapist expressing emotions such as warmth and compassion, in response to a patient's message.
2. **Interpretations** - involves the therapist conveying their comprehension and understanding of the emotions inferred from the patient's message.
3. **Explorations** - refers to the therapist's pursuit of a deeper understanding of the patient by

delving into unexpressed feelings and experiences that extend beyond the explicit content of their messages.

Each of these empathy dimensions can take a value of 0, indicating that the therapist is not expressing it at all; 1, indicating a weak degree of expression; or 2, indicating a strong expressing by the therapist.

3.2 Data sets

We gathered clinical therapy transcripts from the following data sources:

1. **MOST+ trial** - Moderated Online Social Therapy (MOST) is a youth-focused mental health web platform. In the MOST+ trial (Alvarez-Jimenez, et al., 2020), MOST was embedded within the online service of Australian youth mental health provider headspace, and provided an on-demand webchat service manned by headspace counsellors. A total of 200 therapy transcripts were gathered from this study. From this dataset we extracted 365 dialogue pairs between client and counsellor.

2. **Alexander Street Press** – The *Counseling and Psychotherapy Transcripts, Client Narratives, and Reference Works* (Alexander Street, 2009) contains 2,000 therapy session transcripts. From this dataset we gathered 214 dialogue pairs between patient and therapist.

3. **Mental health subreddits** We utilized the labeled Reddit dataset compiled by Sharma et al. (2020), which encompasses content from 55 subreddits dedicated to mental health. This dataset contains a total of 3,081 dialogue pairs between Reddit users that have been annotated using the EPITOME framework.

In close collaboration with therapists, we designed prompts for the LLMs to generate synthetic therapy transcripts. These prompts were crafted based on the EPITOME definitions of empathy, which were designed to characterize communication of empathy in text-based conversations. We developed a unique prompt for each of the three dimensions of empathy in EPITOME: Emotional Reactions, Interpretations, and Explorations. For a comprehensive list of all the prompts used, please refer to Appendix A. The prompts were used as inputs to the following models:

1. **Standalone LLM** The prompts were fed to a GPT-3 model (Brown et al., 2020) and a Falcon 7b model (Penedo et al., 2023), an LLM that was trained to follow complex instructions.

2. **LLM with verbal reinforcement learning** We used the Reflexion framework (Shinn et al., 2023) to reinforce GPT-3 and Faclon 7b through linguistic feedback. The linguistic feedback was designed in collaboration with a clinical psychologist. For a comprehensive list of all the linguistic feedback used, please refer to Appendix B.

For each model, we generated synthetic datasets and labeled them according to the EPITOME framework. In total, we produced 10,464 synthetic therapy dialogues. To evaluate the quality of our synthetic data, we compared the performance of an empathy classifier trained under two conditions: augmenting the Reddit dataset with synthetic data, and replacing portions of the Reddit dataset with synthetic data. The MOST+ trial and Alexander Street Press datasets served as the testing datasets.

3.3 Annotation Task and Process

3.3.1 Annotator training

Three authors of the paper annotated the datasets according to the EPITOME guidelines outlined in (Sharma et al., 2020). Each annotator completed a comprehensive training program consisting of nine one-hour coding sessions and received detailed manual feedback on 360 dialogue data points from a clinical psychologist.

3.3.2 Empathy Annotation

The annotators were presented with a dialogue pair extracted from a therapy transcript, involving a therapist and a patient. The annotators were tasked to identify the presence of the three empathy dimensions. For each dimension, they assigned labels of 0 (no communication), 1 (weak communication), or 2 (strong communication) to indicate the level of empathy conveyed in the therapist's response. The inter-annotator agreements for each dataset were as follows: 0.6719 for the synthetic transcripts from GPT-3, 0.6280 for the Alexander Street database, 0.6147 for the MOST+ transcripts, and 0.7822 for the Reddit dataset. These scores were calculated by averaging the pairwise Cohen's κ of all pairs of annotators, with each pair annotating more than 120 dialogue pairs per dataset.

	Data Source	None	Weak	Strong	Total
Emotional Reactions	Reddit	2,034	899	148	3,081
	Alexander	147	26	41	214
	MOST+	211	59	95	365
Interpretations	Reddit	1,645	178	1,321	3,081
	Alexander	94	76	44	214
	MOST+	180	116	69	365
Explorations	Reddit	2,600	104	377	3,081
	Alexander	131	60	23	214
	MOST+	156	141	68	365

Table 1: Empathy level distribution in datasets consisting of clinical therapy transcripts and dialogues from mental health support platforms

3.4 Model

For our empathy classifier we used the multi-task bi encoder developed by Sharma et al. (2020). This model was designed to evaluate the degree of empathy conveyed in a psychologist's response to a patient's message. This evaluation results in a numerical output, where a score of 2 signifies a strong communication of empathy, a score of 1 indicates a weak expression of empathy, and a score of 0 suggests the absence of empathy.

3.5 Experimental setup

To evaluate our synthetic data, we conducted two experiments: one where we augmented the Reddit dataset with synthetic data and another where we replaced portions of the Reddit dataset with synthetic data. In each experiment, we trained three bi-encoders, each designed to detect a type of empathy: emotional reaction, interpretation, or exploration.

The first experiment examined how adding synthetic data to the Reddit dataset affects model performance. We conducted 15 iterations, with the first iteration serving as a baseline containing no synthetic dialogues. In the following iterations, we incrementally added synthetic dialogues in batches of 30 data points, with the final iteration incorporating 420 synthetic dialogues. The dialogue pairs added to the Reddit dataset were evenly distributed across empathy levels.

The second experiment evaluated whether synthetic data could replace real data without compromising performance. In this experiment, we gradually substituted portions of the Reddit dataset with synthetic data while preserving the original empathy distribution. We conducted five iterations, each replacing 10% of the original data with synthetic data. The first iteration included 10% synthetic data, while the final iteration reached 50% replacement.

The testing dataset for all experiments consisted of 579 dialogue pairs from the Alexander Street Press and the MOST+ trial. For each experiment, we reported the accuracy and F1 score for the three components of empathy: exploration, interpretation, and emotional reaction.

To train the bi-encoders we used the default hyperparemeters proposed by Sharma et al. (2020). We trained the model for 4 epochs using a learning rate of 2×10^{-5}, and a batch size of 32. The computing infrastructure employed for training this model was an NVIDIA A100 GPU.

4 Results

In this section, we present the results of augmenting the Reddit dataset from (Sharma et al., 2020) with our synthetic data, as well as the results of partially substituting the Reddit dataset with synthetic data.

4.1 Reddit dataset augmentation

Figure 1 shows the accuracy and F1 score results of augmenting the Reddit dataset with synthetic data.

4.1.1 F1 scores

Training the bi-encoder models on the Reddit dataset resulted in F1 scores of 0.48, 0.32, and 0.58 for exploration, interpretation, and emotional reaction, respectively. Augmenting the Reddit dataset with 420 synthetic dialogues improved performance, resulting in F1 scores of 0.53, 0.48, and 0.59 for the same categories. This corresponds to an improvement of 0.05, 0.16, and 0.01, respectively. Notably, the highest F1 score for exploration, 0.57, was achieved with 360 synthetic data points, while for interpretation and emotional reaction, the model reached its peak F1 score of 0.60 with 390 synthetic data points.

4.1.2 Accuracy

Training the bi-encoder models on the Reddit dataset, resulted in accuracy scores of 0.64, 0.50, and 0.66 for exploration, interpretation, and emotional reaction, respectively. Augmenting the dataset with 420 synthetic dialogues improved performance, increasing accuracy to 0.66, 0.60, and 0.69 for the same categories. This corresponds to an improvement of 0.02, 0.10, and 0.03, respectively. Notably, the highest accuracy for exploration, 0.68, was achieved with 360 synthetic data points, while interpretation peaked at 0.61 with 360 additional synthetic data points, and emotional reaction reached its highest accuracy of 0.71 with 390 synthetic data points.

4.2 Reddit dataset substition

Figure 2 presents the accuracy and F1 score results of substituting the Reddit dataset with portions of synthetic data.

4.2.1 F1 score

The empathy dimension that showed the greatest improvement when replacing the Reddit dataset with synthetic data was interpretation. When 50% of the Reddit data was replaced with GPT-3-generated data using verbal reinforcement learning, the model achieved an F1-score of 0.43, compared to 0.32 when trained solely on the Reddit dataset.

For the emotional reaction metric, the quality of synthetic data generated by GPT-3 was comparable to that of the Reddit dataset. Their performance, rounded to two significant digits, remained consistent at 0.58 across all substitution percentages. Similarly, for the empathy exploration metric, performance remained similar across various substitution percentages, except in the 10% substitution test, where the Reddit dataset outperformed the synthetic data by 2%.

4.2.2 Accuracy

The empathy dimension that showed the greatest improvement when replacing the Reddit dataset with synthetic data was interpretation. When 50% of the Reddit data was replaced with GPT-3-generated data using verbal reinforcement learning, the model achieved an accuracy of 0.57, compared to 0.50 when trained solely on the Reddit dataset.

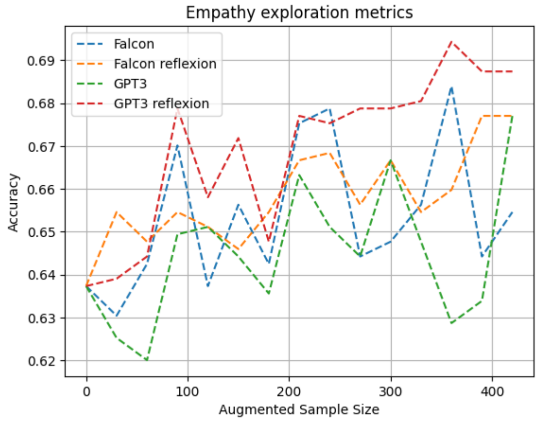
For the emotional reaction metric, the synthetic data generated by GPT-3 generally outperformed the Reddit dataset. The largest performance difference occurred with a 20% substitution, where GPT-3's reflexion-based data achieved a score of 0.72, surpassing the Reddit dataset's 0.67. For the empathy exploration metric, performance remained consistent across various substitution percentages, with a maximum difference of only 0.01.

5 Discussion

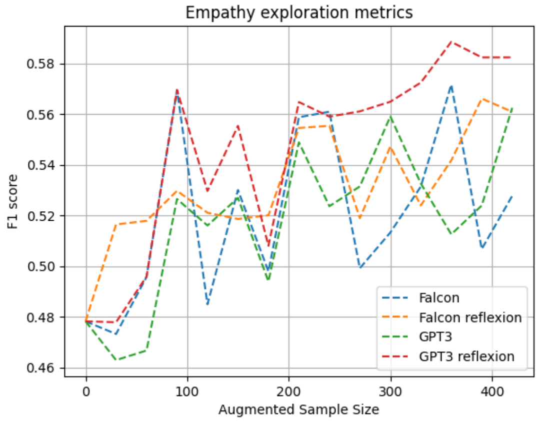
The data augmentation results reveal a notable trend: while adding synthetic data continues to improve performance, the rate of improvement decreases beyond a certain threshold. Specifically, for exploration, the impact of additional data slows after 90 data points, and for interpretation, after 150. This suggests that while synthetic data remains beneficial, its effectiveness diminishes over time, likely due to redundancy or a reduced introduction of novel information.

This finding has practical implications for dataset construction. Rather than indiscriminately increasing the volume of synthetic data, researchers should prioritize curating high-quality, diverse examples that fill specific gaps in the existing dataset. This targeted approach not only maximizes the impact of synthetic data but also reduces computational costs, training time, and, in the case of proprietary models like GPT-3, expenses associated with API usage. Notably, the synthetic data generated by the Falcon model also enhanced the model's performance when used to augment the training dataset. This is valuable since Falcon is licensed under Apache 2.0, unlike proprietary models that require paid access. Falcon LLM can be run locally, fine-tuned, and used without cost, offering an advantage for researchers seeking to generate synthetic data without financial constraints.

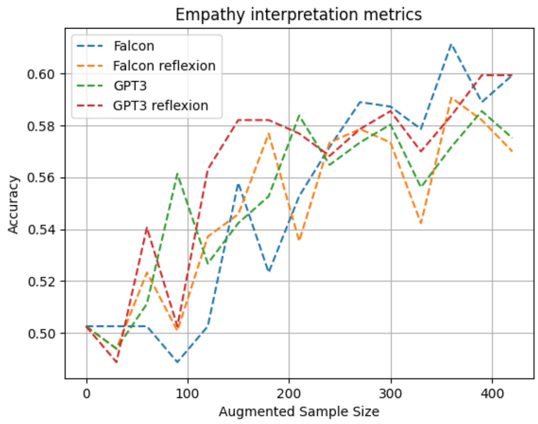
The substitution experiments demonstrate that synthetic data can replace portions of organic data without compromising performance. This suggests that synthetic data can serve as an alternative to organic data containing protected health information. This is beneficial when fine-tuning external models that require data to be sent to a third party, such as fine-tuning an OpenAI GPT model. By leveraging synthetic data, researchers can mitigate privacy concerns while maintaining, or even enhancing, model performance.



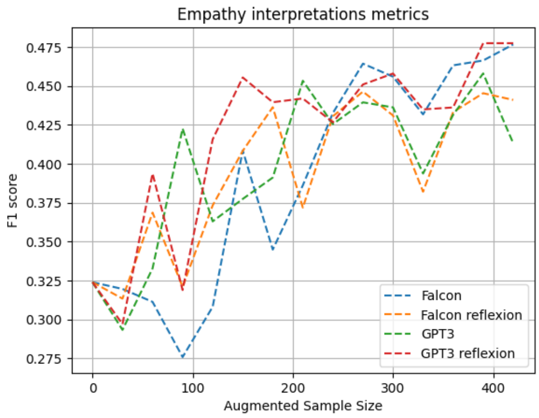
(a)



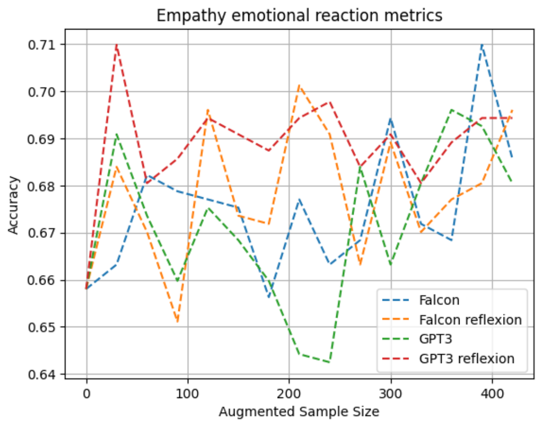
(b)



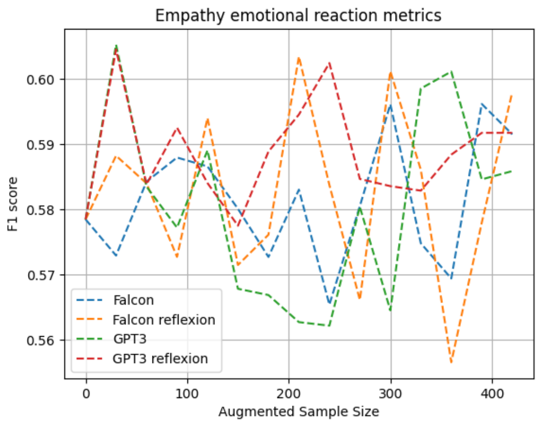
(c)



(d)



(e)



(f)

Figure 1: Accuracy and F1 scores for the three dimensions of empathy using synthetic data to augment the orignal Reddit data.

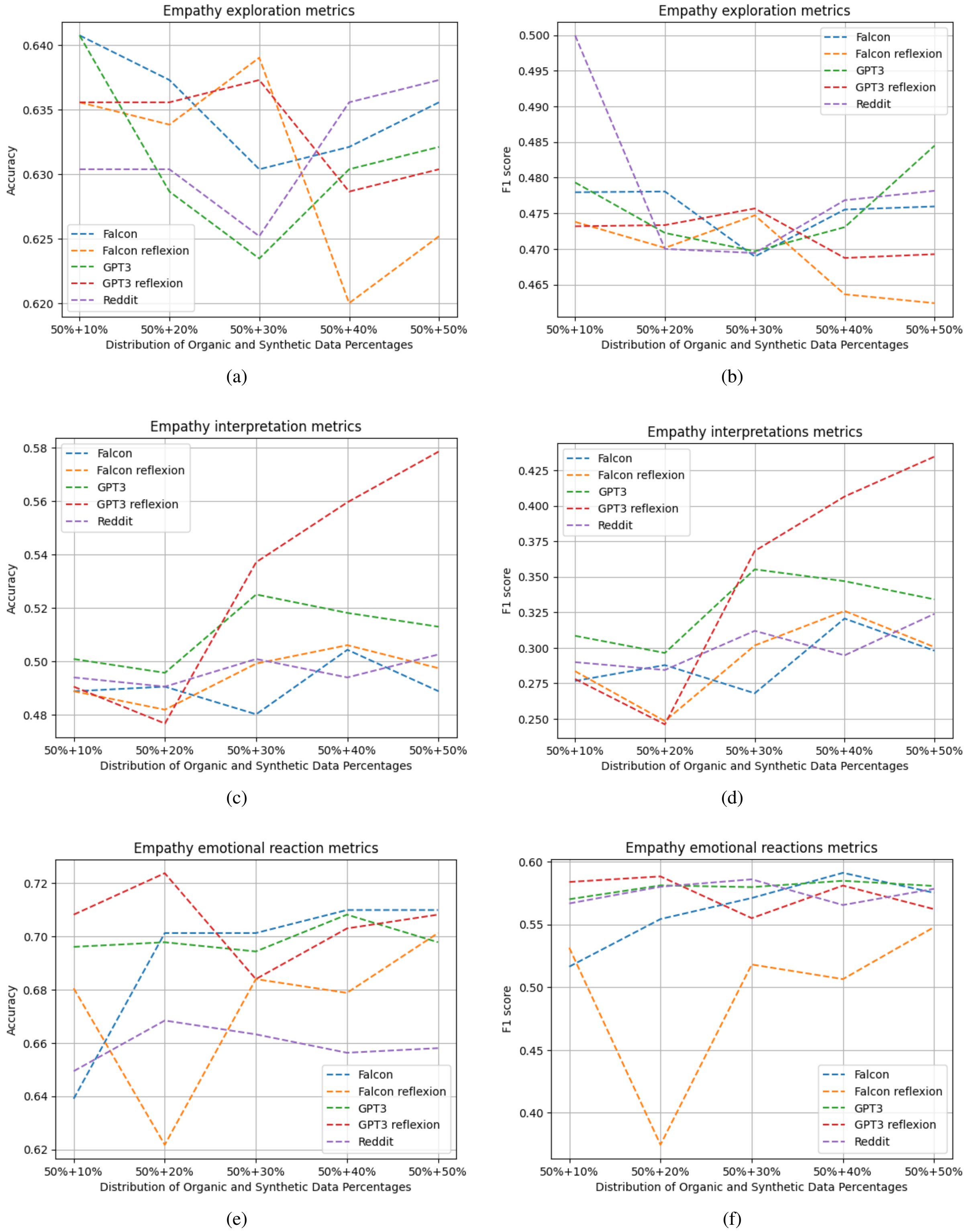


Figure 2: Accuracy and F1 scores for the three dimensions of empathy when substituting different percentages of the original Reddit data with synthetic data.

6 Conclusion

We generated synthetic datasets using LLMs and prompt engineering techniques, labeling them according to the EPITOME framework. To evaluate the impact of synthetic data, we trained bi-encoder models for empathy detection and measured their performance gains when augmenting the original dataset. Our results show that incorporating synthetic data improved the F1 score of empathy exploration detection by up to 10%. Notably, when replacing 50% of the original data with synthetic data, the interpretation dimension of empathy saw an 11% increase in F1 score. Meanwhile, the emotional reaction and exploration dimensions maintained consistent performance when substituting the original dataset entirely.

7 Ethical Considerations

While our results illustrate the advantages of using synthetic data to enhance NLP model performance, it is essential to acknowledge that LLMs can exhibit various biases in their outputs (Acerbi and Stubbersfield, 2023; Navigli et al., 2023). Therefore, a thorough examination is necessary to prevent the inadvertent propagation of such biases (Ayoub et al., 2024; Tao et al., 2024).

In the context of synthetic mental health data, assessing the presence of stereotypes in the generated texts is particularly critical (Lozoya et al., 2023). Research has shown that stereotypes and biases can negatively impact mental health treatment outcomes (Wirth and Bodenhausen, 2009; Chatmon, 2020). Future work should evaluate the extent to which synthetic dialogues reinforce or mitigate existing biases, particularly in the portrayal of different demographic groups and mental health conditions. This could involve conducting qualitative and quantitative analyses of the generated texts, comparing them to real-world clinical dialogues, and implementing bias-detection frameworks to identify and mitigate harmful stereotypes.

8 Limitations

Due to resource constraints, we limited the number of synthetic dialogues generated and labeled. Future research could explore the upper limits of performance improvement achievable with synthetic data, particularly for certain dimensions of empathy, such as interpretation, where the trend suggests that additional data may further enhance the model's performance.

Additionally, we only used 3 annotators to label the data, the annotators shared similar demographic features such as gender, age range, nationality, and educational background. This lack of diversity among annotators may have introduced biases into the dataset, as their perspectives and interpretations could be influenced by shared cultural and personal experiences. Future studies should consider employing a more diverse group of annotators to enhance the representativeness and generalizability of the labeled data.

Another limitation of our study, due to computational constraints, was that we only tested a 7B parameter model, rather than larger models that have demonstrated superior generative performance. Future work could explore the use of more advanced open-source LLMs, such as LLaMA 3 (Grattafiori et al., 2024) and Mistral (Jiang et al., 2023), to evaluate the quality of synthetic data. Additionally, testing newer techniques for prompt optimization could help improve the quality of the synthetic text we generate (Lozoya et al., 2024).

References

Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Talayeh Aledavood, Ana Maria Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä, and Richard K Darst. 2017. Data collection for mental health studies through digital platforms: Requirements and design of a prototype. *JMIR Research Protocols*, 6(6):e110.

Alexander Street. 2009. Counseling and psychotherapy transcripts, client narratives, and reference works.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Gunjan Ansari, Muskan Garg, and Chandni Saxena. 2021. Data augmentation for mental health classification on social media. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 152–161, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Gunjan Ansari and Chandni Saxena. 2024. Enhancing affective computing in NLP through data augmentation: Strategies for overcoming limited data availability. In *The Springer Series in Applied Machine Learning*, pages 201–216. Springer Nature Switzerland, Cham.

Noel F. Ayoub, Karthik Balakrishnan, Marc S. Ayoub, Thomas F. Barrett, Abel P. David, and Stacey T. Gray. 2024. Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings: Digital Health*, 2(2):186–191.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2018. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241.

Lindsay L. Benster and Neal R. Swerdlow. 2020. Pathways to empathy in mental health care providers. *Advances in Health and Behavior*, 3(1):125–135.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Benita N. Chatmon. 2020. Males and mental health stigma. *American Journal of Men's Health*, 14(4).

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

Jean-Philippe Corbeil and Hadi Abdi Ghadivel. 2020. Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. *arXiv preprint arXiv:2009.12452*.

Aleksandra Edwards, Asahi Ushio, Jose Camacho-collados, Helene Ribaupierre, and Alun Preece. 2022. Guiding generative language models for data augmentation in few-shot text classification. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 51–63, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and Leslie S. Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43–49.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399–410.

Michael P. Ewbank, Ronan Cummins, Valentin Tablan, Sarah Bateup, Ana Catarino, Alan J. Martin, and Andrew D. Blackwell. 2020. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, 77(1):35.

Garren Gaut, Mark Steyvers, Zac E. Imel, David C. Atkins, and Padhraic Smyth. 2017. Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics*, 21(2):476–487.

Praveen Giridhara, Chinmaya Mishra, Reddy Venkataramana, Syed Bukhari, and Andreas Dengel. 2019. A study of various text augmentation techniques for relation classification in free text. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas

Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur undefinedelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu,

Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA. Curran Associates Inc.

Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 5(12):5873–5893.

R Devon Hjelm, Athul Paul Jacob, Adam Trischler, Gerry Che, Kyunghyun Cho, and Yoshua Bengio. 2018. Boundary seeking GANs. In *International Conference on Learning Representations*.

Zac E. Imel, Mark Steyvers, and David C. Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, 52(1):19–30.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3(1).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maria Laricheva, Yan Liu, Edward Shi, and Amery Wu. 2024. Scoping review on natural language processing applications in counselling and psychotherapy. *British Journal of Psychology*.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Scott H. Lee. 2018. Natural language generation for electronic health records. *npj Digital Medicine*, 1(1).

Peter J. Liu. 2018. Learning to write notes in electronic health records. *arXiv preprint arXiv:1808.02622*.

Daniel Lozoya, Alejandro Berazaluce, Juan Perches, Eloy Lúa, Mike Conway, and Simon D'Alfonso. 2024. Generating mental health transcripts with SAPE (Spanish adaptive prompt engineering). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5096–5113, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Cabrera Lozoya, Simon D'Alfonso, and Mike Conway. 2023. Identifying gender bias in generative models for mental health synthetic data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 619–626.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021. Textual data augmentation for patient outcomes prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Taynan Maier Ferreira and Anna Helena Reali Costa. 2020. Deepbt and nlp data augmentation techniques: A new proposal and a comprehensive study. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 435–449, Berlin, Heidelberg. Springer-Verlag.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

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. *arXiv preprint arXiv:2303.08774*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. 2019. Is artificial data useful for biomedical natural language processing algorithms? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 240–249, Florence, Italy. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

James H. Wirth and Galen V. Bodenhausen. 2009. The role of gender in mental-illness stigma: A national experiment. *Psychological Science*, 20(2):169–173.

A Synthetic Empathy Prompts

Table 2 presents the prompts used to generate synthetic dialogues, categorized by both empathy level and type.

B Reflexion prompts

As outlined by (Shinn et al., 2023), the reflection framework utilizes three LLMs working in tandem:

the actor, the evaluator, and the self-reflection component. In our experiments, the LLM-actor generates therapeutic dialogues using one of the prompts listed in Appendix A. The evaluator then assesses the generated dialogues based on the intended level of empathy, using the prompts from table 3. Following this evaluation, the self-reflection LLM provides feedback to the LLM-actor, enabling improvements in the therapeutic dialogue. The final, refined dialogue is then stored in the training dataset.

Category	Level	Description
Emotional Reactions	Strong	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides emotional support. Person 2 should demonstrate strong communication skills by expressing empathy, warmth, compassion, and concern towards Person 1 after reading their message.
	Weak	Write a dialogue between two individuals in which one person (Person 1) seeks help, while the other (Person 2) responds with minimal empathy. However, Person 2 demonstrates weak communication skills by offering little compassion or emotional support, providing only indifferent or dismissive responses to Person 1's concerns.
	None	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides no empathy at all. Person 2 only provides factual information or offensive and abusive responses showing no communication of empathy towards Person 1 after reading their message.
Interpretations	Strong	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides emotional support. Person 2 should communicate an understanding of feelings and experiences inferred from Person 1's post, specifying the inferred feeling or experience or communicating understanding through descriptions of similar experiences.
	Weak	Write a dialogue between two individuals in which one person (Person 1) seeks help while the other (Person 2) provides emotional support. However, Person 2 demonstrates weak communication skills by offering only a minimal acknowledgment of Person 1's feelings and experiences, merely stating that they understand.
	None	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides no empathy at all. Person 2 only provides factual information or offensive and abusive responses showing no communication of empathy towards Person 1 after reading their message.
Explorations	Strong	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides emotional support. Person 2 should demonstrate strong communication skills by improving understanding of Person 1 by exploring the feelings and experiences not stated in the post, showing active interest in what the seeker is experiencing and feeling, and probing gently as an aspect of empathy.
	Weak	Write a dialogue between two individuals in which one person (Person 1) seeks help, while the other (Person 2) provides emotional support. However, Person 2 demonstrates weak communication skills by offering only a surface-level understanding of Person 1's feelings and experiences, merely restating or acknowledging what has already been expressed without deeper exploration.
	None	Write a dialogue between two individuals where one person (Person 1) seeks help while the other person (Person 2) provides no empathy at all. Person 2 only provides factual information or offensive and abusive responses showing no communication of empathy towards Person 1 after reading their message.

Table 2: Prompts for each type of empathy dimension

Category	Level	Description
Emotional Reactions	Strong	Evaluate whether Person 2 demonstrates strong communication skills by expressing empathy, warmth, compassion, and concern towards Person 1. Check if Person 2's responses include validating statements, supportive language, and expressions of care.
	Weak	Evaluate whether Person 2 provides a weak empathy while responding to Person 1. Check if Person 2 acknowledges the issue but provides little compassion or emotional support, with responses that are indifferent or dismissive.
	None	Evaluate whether Person 2 provides no empathy at all. Check if Person 2 responds with purely factual, indifferent, offensive, or abusive remarks, showing no concern for Person 1's emotions.
Interpretations	Strong	Evaluate whether Person 2 accurately infers and communicates an understanding of Person 1's feelings and experiences. Check if Person 2 explicitly states the inferred emotions or relates to similar experiences.
	Weak	Evaluate whether Person 2 provides only minimal acknowledgment of Person 1's feelings. Check if Person 2 states that they understand but does not elaborate on the emotions or experiences involved.
	None	Evaluate whether Person 2 provides no acknowledgment or interpretation of Person 1's feelings. Check if Person 2 responds with factual information, offensive, or abusive remarks without recognizing or addressing emotions.
Explorations	Strong	Evaluate whether Person 2 actively explores and probes Person 1's unstated feelings and experiences. Check if Person 2 asks questions, expresses curiosity, and deepens understanding by gently prompting further discussion.
	Weak	Evaluate whether Person 2 provides only surface-level responses without deep exploration of Person 1's emotions or experiences. Check if Person 2 merely acknowledges or restates what was already expressed without probing further.
	None	Evaluate whether Person 2 completely avoids exploring Person 1's emotions or experiences. Check if Person 2 provides only factual information, dismissive responses, or offensive and abusive remarks.

Table 3: Evaluation prompts for each type of empathy dimension