CBTLlama: Fine Tuning Large Language Models For Identifying Thought Distortions

David Schiff
davidschiff100@gmail.com

April 29, 2024

Abstract

This paper presents a novel approach to enhancing Cognitive Behavioral Therapy (CBT) by leveraging large language models, specifically, Meta's LLaMA-3 8B model, fine-tuned on CBT thought distortion assessments generated by the AI assistant Claude. The proposed method aims to democratize and enhance cognitive behavioral therapy for individuals by leveraging the capabilities of large language models. We automate the creation of diverse and tailored CBT scenarios and fine-tune the large language model for assessing and challenging cognitive distortions, this approach has the potential to improve the accessibility and effectiveness of CBT. The paper discusses the methodology, dataset generation, model fine-tuning, and evaluation of the generated thought distortion assessments. Our model and training data [1] [2] are available on huggingface https://huggingface.co/epsilon3/cbt-llama3-8b-finetuned

1 Introduction

Cognitive Behavioral Therapy (CBT) is a widely used and effective treatment for various mental health conditions, focusing on identifying and challenging distorted thoughts and beliefs. However, learning to analyze and challenge realistic thought distortion scenarios for each individual can be time-consuming and take many sessions for someone new to CBT. Therefore, an AI assistant trained to recognize and challenge cognitive distortions could be very benefitial. In this paper we create a novel way of generating realistic scenarios based on virtual personas generated by Claude. These scenarios include a variety of negative emotions to make the model generalize to different possible complex emotional scenarios. We hope this model is an inspiration for future methods of assisting therapists in treating their patients using similar methods.

2 Previous Work

Recent studies on fine-tuning large language models (LLMs) for mental health applications demonstrate significant progress in this interdisciplinary field. Stade et al. (2024) propose a framework for responsible development and evaluation of behavioral healthcare advancements using LLMs, emphasizing the necessity for task-specific model adjustments [7]. Furthermore, Ji et al. (2021) introduced ”Mentalbert,” a publicly available language model specifically pretrained for mental healthcare applications, highlighting the potential of LLMs in tasks related to mental disorders [8]. Flemotomos et al. (2018) developed models for analyzing therapist and patient interactions during cognitive behavioral therapy (CBT) sessions, suggesting that automated evaluation of therapy quality is feasible [9]. This research was extended by Flemotomos et al. (2021), who improved these models to perform automated quality assessments of CBT sessions, showing that contextualized language representations could significantly enhance therapy quality monitoring [10]. Jiang et al. (2024) explored AI-enhanced CBT, using deep learning to extract cognitive pathways from social media texts, thus illustrating how LLMs can offer insights into cognitive processes relevant to mental health [11]. Recently, Na (2024) introduced CBT-LLM, a novel large language model fine-tuned for Chinese psychological health QA based on CBT principles [12]. To the best of our knowledge, no model that analyzes cognitive distortions and challenges them in English using AI-generated synthetic data has yet been developed.

3 Methodology

3.1 Data Generation

3.1.1 Triggering Events

In this work, synthetic ”triggering” events were generated by Claude using a novel technique leveraging demographic and emotional ”seeds”. Each seed contains the following information: **Sex, Age, Occupation, Relationship Status, Negative Emotions**. Negative emotions are derived from the following list and are used to generate realistic scenarios such that each 4 tuple (Sex, Age, Occupation, Relationship Status) is paired with one negative emotion out of this list.

- Anger
- Anxiety
- Bitterness
- Contempt
- Despair
- Disappointment
- Disgust
- Embarrassment
- Envy
- Fear
- Frustration
- Grief
- Guilt
- Hatred
- Helplessness
- Hopelessness
- Humiliation
- Insecurity
- Irritation

- Jealousy
- Loneliness
- Melancholy
- Misery
- Neglect
- Panic
- Paranoia
- Rage
- Regret
- Rejection
- Remorse
- Resentment
- Sadness
- Shame
- Sorrow
- Spite
- Stress
- Suffering
- Unhappiness
- Vengefulness
- Worry

Though some emotions stated in this list are similar, they are essentially a different token when read by the model and thus give the model the ability to create a similar situation but slightly different - disrupting the determinism slightly. This is proven by the fact that in the whole dataset there are only two scenarios identical word by word. Claude's created scenarios actually managed to capture multiple emotions in the scenarios generated.

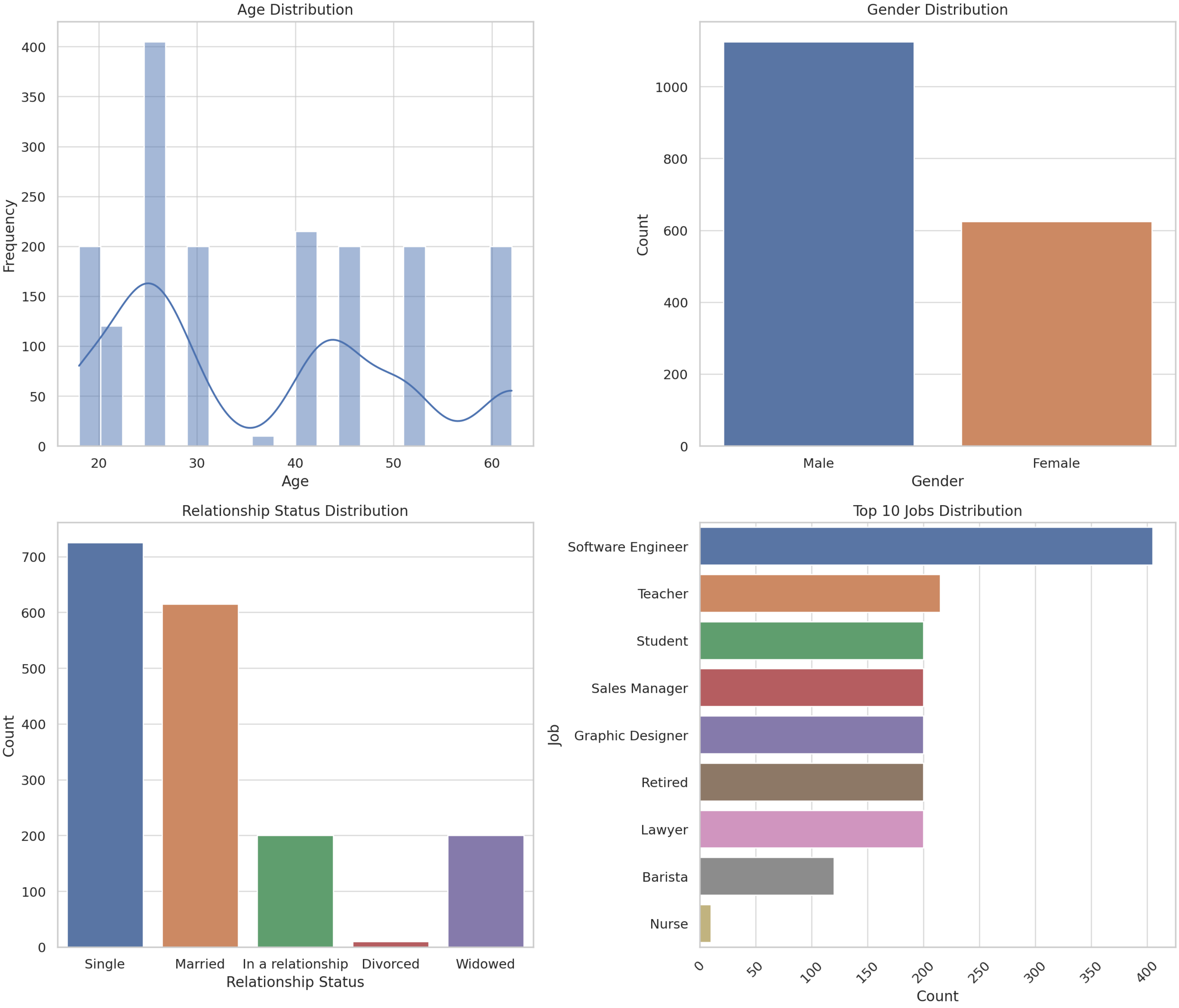

Figure 1: Synthetic Data Statistics

3.1.2 Cognitive Distortion Assessment and Challenge

Following the generation of the realistic scenarios

- My coworker took credit for my work during a team meeting. I felt so angry and betrayed when my coworker took credit for my hard work. It was incredibly frustrating to see them get praised for something I did.
- My friends made plans to hang out without inviting me. I feel left out and forgotten. It hurts to know that they didn't think to include me. I wonder if they don't like me anymore. Maybe I did something wrong or they just don't care about me.

We use Claude to analyze the event and the subjective assessment using a cognitive distortions work sheet [6]. We started by using Claude Opus but due to budget limitations, we used Claude Haiku with few shot prompting. The prompts used will be included in our git repository (see huggingface model repository).

Cognitive Distortions: Detailed Descriptions and Challenging Strategies

- **All-or-nothing thinking (also known as black-and-white thinking):**
  - *Description:* Seeing things in absolute, black-and-white categories, with no middle ground. If your performance falls short of perfect, you see yourself as a total failure.
  - *Challenge:* Recognize that reality exists on a continuum, and there are shades of gray between the extremes. Evaluate your performance on a scale of 0-100 instead of thinking in all-or-nothing terms.
- **Overgeneralization:**
  - *Description:* Viewing a single negative event as a never-ending pattern of defeat. Using words like ”always” or ”never” when thinking about negative events.
  - *Challenge:* Recognize that a single event does not necessarily define a pattern. Look for counter-examples that disprove the overgeneralization. Replace extreme words like ”always” and ”never” with more realistic and specific language.
- **Mental filter:**
  - *Description:* Picking out a single negative detail and dwelling on it exclusively, so that your vision of all reality becomes darkened, like a drop of ink that discolors an entire beaker of water.

  - *Challenge:* Consider the whole picture, not just the negative detail. Make a list of the positive aspects of the situation to balance out your perspective.

- **Discounting the positive:**
  - *Description:* Rejecting positive experiences by insisting they ”don’t count.” If you do a good job, you may tell yourself that it wasn’t good enough or that anyone could have done as well.
  - *Challenge:* Give yourself credit for your accomplishments, skills, and positive qualities. Recognize that dismissing the positive is an irrational bias and consciously challenge it.
- **Jumping to conclusions:**
  - *Description:* Interpreting things negatively when there are no facts to support your conclusion. This includes mind reading (assuming someone is reacting negatively to you) and fortune telling (predicting that things will turn out badly).
  - *Challenge:* Ask yourself, ”What is the evidence for and against my conclusion?” Consider alternative explanations and possibilities. Test your predictions to see if they come true.
- **Magnification (catastrophizing) or minimization:**
  - *Description:* Exaggerating the importance of your problems and shortcomings, or minimizing the importance of your desirable qualities. This is also called the ”binocular trick.”
  - *Challenge:* Put your situation into perspective. Ask yourself, ”How much will this matter in a year or five years?” Consider the realistic consequences rather than imagining the worst-case scenario.
- **Emotional reasoning:**
  - *Description:* Assuming that your negative emotions necessarily reflect the way things really are: ”I feel it, therefore it must be true.”
  - *Challenge:* Recognize that feelings are not facts. Just because you feel a certain way doesn’t mean the situation is as you imagine it to be. Look for objective evidence rather than basing your conclusions on feelings alone.
- **Should statements:**

    - *Description:* Criticizing yourself or other people with "shoulds," "shouldn'ts," "musts," "oughts," and "have-tos." This creates unrealistic expectations and pressure.
    - *Challenge:* Replace "should" statements with more realistic and flexible language, such as "I would prefer" or "It would be nice if." Recognize that "should" statements often lead to frustration and disappointment.

- **Labeling:**
    - *Description:* Using a single negative label to define yourself or others globally, such as "I'm a loser" or "He's an S.O.B."
    - *Challenge:* Recognize that you and others are complex individuals with many qualities, not just one. Avoid defining yourself or others based on a single characteristic or event. Use specific and realistic language instead of global labels.

- **Personalization and blame:**
    - *Description:* Holding yourself personally responsible for events that aren't entirely under your control or blaming others while overlooking how your own attitudes and behavior might contribute to a problem.
    - *Challenge:* Consider all the factors that contribute to a situation, not just your own role. Recognize that blaming yourself or others excessively is unproductive and often inaccurate. Focus on finding solutions and taking appropriate responsibility.

Finally, we fine tuned Llama3 8B on google colab using a single A100 GPU for 5 epochs. Our dataset consists of 621 samples of triggering events, their response, cognitive distortions and their challenges by Claude. We divide the dataset into training and testing, leaving 100 samples for test.

3.1.3 Training Methodology

We use the pre-trained model `unsloth/llama-3-8b-Instruct-bnb-4bit` with a maximum sequence length of 2048 tokens and 4-bit precision via `torch.float16`. Our model employs Low-Rank Adaptation (LoRA) on key transformer modules including query, key, value, and output projections, with a rank of 16 and an alpha value of 16. Gradient checkpointing is activated to optimize memory usage. The model is trained with a batch size of 2, using gradient accumulation over 4 steps, and a linear learning rate schedule starting from 0.0002. We utilize `adamw_8bit` for optimization, with a warm-up of 5 steps. Training runs for up to 5 epochs.

4 Results

After 5 epochs the loss on the train set was 0.27 test set was 0.8614, a significant improvement considering the initial train set loss was 1.78. Qualitative analysis of model results was done manually by us and there were significant improvements after fine tuning. Of course, a proper analysis done by actual CBT therapists should be done. Because the model results were not reviewed by any therapists, any usage of the model should be done with caution.

Model Output Example

It's completely natural and understandable to feel terrible, remorseful, and regretful in this situation. Letting your team down and causing them unnecessary stress due to procrastination is a difficult thing to cope with, and your emotions are valid. It's important to acknowledge and process these feelings as part of learning and growth.

Here are some potential cognitive distortions and challenges based on the situation you described:

- **Overgeneralization**: "I always let my team down and cause them stress."
  - *Challenge*: Recognize that this specific incident does not define a permanent pattern. Consider times when you have been responsible and met deadlines on time, which discredits the overgeneralization.
- **Labeling**: "I'm a terrible team member for letting this happen."
  - *Challenge*: Avoid defining yourself based on a single event. You are a complex individual with many qualities, and one mistake does not negate your overall value as a team member.
- **Magnification (catastrophizing)**: "This one mistake ruins everything and makes me a failure."
  - *Challenge*: Put the situation into perspective. While it was an important deadline, it does not negate all your other contributions and successes. Consider how much this specific incident will matter in the long run.
- **Should statements**: "I should have managed my time better" or "I shouldn't have let this happen."
  - *Challenge*: Replace these "should" statements with more realistic and compassionate language, such as "I wish I had managed

my time better" or "I'm sorry this situation occurred." Recognize that everyone makes mistakes and that being hard on yourself is not productive.

Remember, everyone faces challenges and makes mistakes from time to time. The most important thing is to learn from this experience, take responsibility for your actions, and find ways to prevent similar situations in the future. Be kind to yourself and focus on moving forward productively.

5 Discussion

This study demonstrates the feasibility of using large language models like LLaMA-3 8B for assessing realistic and personalized scenarios for Cognitive Behavioral Therapy (CBT). The integration of demographic data and emotional states to create triggering events for fine tuning represents a significant advancement in automating and personalizing mental health interventions. One of the notable findings was Claude's capability to generate distinct scenarios with minimal repetition, showcasing the potential for creating diverse treatment scenarios without therapist input.

However, the study also highlighted several challenges. The model's performance, while improved after fine-tuning, still requires rigorous validation by mental health professionals to ensure its practical efficacy. Because of this limitation, it is hard to say exactly how effective LLMs are in detecting and challenging cognitive distortions.

6 Conclusion

The findings from this study underscore the transformative potential of AI in mental health, particularly through the application of large language models in CBT. While promising, the application of AI in therapeutic contexts must proceed with cautious optimism. Future research should focus on extensive validations, ethical considerations, and continuous improvement of the models to better mimic therapeutic interactions. This research lays a foundational step towards more accessible and efficient mental health interventions, potentially democratizing mental health support and making it more available to those in need. The most important takeaway from this paper is the possibility of using LLMs to assist those in need of Cognitive Behavioural Therapy treatment. A patient could possibly use a large language model running on his own laptop (for privacy reasons) to train themselves to better understand and challenge cognitive distortions.

References

[1] Activating/Triggering Events Dataset `https://huggingface.co/datasets/epsilon3/cbt-triggering-events`

[2] Cognitive Distortions Analysis Dataset `https://huggingface.co/datasets/epsilon3/cbt-cognitive-distortions-analysis`

[3] Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research, 36*(5), 427-440. `http://doi.org/10.1007/s10608-012-9476-1`

[4] Leahy, R. L. (2020). *Emotional schema therapy: Distinctive features.* Routledge.

[5] Hu, H., & Yang, Y. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[6] Cognitive Behavioural Therapy Worksheet `https://arfamiliesfirst.com/wp-content/uploads/2013/05/Cognitive-Distortions.pdf`

[7] E.C. Stade, S.W. Stirman, L.H. Ungar, C.L. Boland (2024) Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation npj Mental Health Research, Nature `https://www.nature.com/articles/s44184-024-00056-z`

[8] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari (2021) Mentalbert: Publicly available pretrained language models for mental healthcare arXiv preprint arXiv:2110.15621 `https://arxiv.org/abs/2110.15621`

[9] N. Flemotomos, V.R. Martinez, J. Gibson, D.C. Atkins (2018) Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions Interspeech, ISCA `https://www.isca-archive.org/interspeech_2018/flemotomos18_interspeech.pdf`

[10] N. Flemotomos, V.R. Martinez, Z. Chen, T.A. Creed (2021) Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations PloS one, Public Library of Science `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258639`

[11] M. Jiang, Y.J. Yu, Q. Zhao, J. Li, C. Song, H. Qi (2024) AI-Enhanced Cognitive Behavioral Therapy: Deep Learning and Large Language Models for Extracting Cognitive Pathways from Social Media Texts arXiv preprint arXiv:2404.11449 `https://arxiv.org/abs/2404.11449`

[12] Hongbin Na (2024) CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering arXiv:2403.16008 `https://arxiv.org/abs/2403.16008`