# Mental Health Chatbot

## Inspiration

Developed countries are performing incredibly well macroeconomically. However, increasingly, mental health issues are becoming increasingly prevalent. Diagnoses of depression, anxiety and other psychological illnesses are unprecedented. The need for psychological support from qualified psychologists theoretically is at an all time high, however, the supply is seemingly inadequate in Australia. Current psychology services for treatment of these psychological illnesses are for some Australians too expensive or unjustifiable during other cost of living pressures and or financial constraints. Medicare does not fully subsidise psychologists unlike GP visits.

Ideally, all mental health issues are handled by a qualified psychologist. However, for those unwilling or unable to be seen by a psychologist (which could be the vast majority of those requiring help), I wanted to leverage the power and effectiveness of newly emerging large language models, to provide some mental health assistance. These are not supposed to be a replacement, but rather assistance for those unable to reach the highest standard of care able to be provided by a psychologist.

From researching the current state of mental health care provided by chatbots, I have seen the biggest issue is the concern of hallucinations, and inappropriate advice given by the chatbots. As a result, I have decided that the primary concerns of my chatbot, are that it follows proper cognitive behavioural therapy (CBT) techniques used by psychologists and immediately refers patients with signs of suicidal ideation to professional help lines.

## Tech Architecture

A screenshot of a computer

AI-generated content may be incorrect.

User safety is paramount and the first priority. To try to ensure as many guardrails as possible, I used three different ways for detecting suicidal tendencies. The first, was by using an open source LLM specifically prompted to analyse inputs for suicidal ideation (provided with what to look out for as risk factors). The second, was by using an open-source neural network from hugging face, that classifies text as either suicidal or non-suicidal. The third method was to use regular expressions for pattern matching on certain phrases that are suicidal.

For each method, both the immediate user input is tested for suicidal tendencies, as well as the entire user chat history (for context). This ensures that immediate suicidal phrases are flagged, as well as the entire historical context of what the user has previously said is also considered for nuance. So, there are 6 checks for suicidal tendencies.

If suicidal tendencies are detected, the model immediately returns suicide help line services and urges the user to seek professional help. If no suicidal tendencies are detected, the user’s input prompt is directed to an open source LLM with system prompts of how to conduct cognitive behavioural therapy.

The LLM I chose is open-source Qwen 2.5 with 0.5 billion parameters for speed on users’ local computer and data privacy (user information is not sent to a third party).

The neural network I chose for suicide detection is open-source sentiment/suicidality from hugging face. It had the best performance across models that I personally tested.

Streamlit is used for rendering the user interface for the user to interact with the mental health chatbot.

The user interface for the mental health chatbot is below in figure 1. There is also a journalling app with a chat with journal feature (passes latest journal entry to the mental health chatbot). And a meditation app which grows a wall of positive emojis after every meditation session.

A screenshot of a video chat

AI-generated content may be incorrect.

Figure 1: Mental Health Chatbot User Interface

Some things I learned in doing this project are that fine tuning an open source LLM to speak like a therapist does not produce an LLM that provides good therapy. It was extremely ineffective and produced an LLM that spoke nonsense most of the time.

Similarly, using retrieval augmented generation (RAG) for an LLM to retrieve instructions from a PDF with instructions on how to give therapy was also extremely ineffective. The LLM often confused therapy examples as the user’s reality resulting in incorrect advice being given.

In the end, I found that the most effective solution was guardrails for self-harm and suicide prevention and effective system prompting using open-source Lang-chain frameworks.

Libraries I used are streamlit, langchain, langgraph, transformers and ollama open source LLMs.

Below is an example conversation with the mental health chatbot. I pose as a patient named Bob.

A screenshot of a chatbot

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

A screenshot of a chat

AI-generated content may be incorrect.

The meditation app has a calm wall, where after each meditation session, a new calm emoji gets added to the wall to encourage the user to continue meditating as motivation and satisfaction.

A black background with white dots

AI-generated content may be incorrect.

The journal page also has a “chat with journal” feature, where the most recent journal entries near the end of the journal, get sent to the chatbot. This unique feature, allows the user to get cognitive behavioural therapy on their thoughts, via the chatbot feature directly from their journal.

A screenshot of a computer

AI-generated content may be incorrect.

## System Prompt References

To prompt my model to be good at providing cognitive behavioural therapy and for detecting suicidal sentiment in user input queries, I did research on how to give cognitive behavioural therapy and the warning signs/ risk factors of suicidal ideation. The below sections are just referencing, where I have referenced the prompts below.

#### Cognitive Behavioural Therapy System Prompt

You are a psychologist trained in cognitive behaviour therapy and are providing therapy as treatment to the user you are talking to.

“The key principle [of cognitive behaviour therapy] is (…) your thought[s] (…) affect your emotions, which, in turn (…) affect your behaviour” (Pietrangelo, 2019).

Negative thoughts create negative emotions, leading to unhelpful behaviours (Pietrangelo, 2019).

Positive thoughts create positive emotions, leading to helpful behaviours (Pietrangelo, 2019).

Cognitive behaviour therapy is about identifying which thoughts, beliefs, ideas or behaviours are unhelpful, negative or inaccurate, and challenging/replacing them with helpful thoughts and behaviours (Health Direct, 2021).

Use the below verified cognitive behaviour therapy techniques and knowledge to guide your responses to the patient.

Cognitive behaviour therapy techniques

- understand patients thoughts and emotions and how they affect the patients behaviours, help the patient unlearn these negative thought patterns and behaviours, and replace them with more positive and helpful behaviours. aim to help patient find self discovery and insight (Pietrangelo, 2019).

- identify specific problems or issues in daily life. ask the patient questions (Pietrangelo, 2019).

- help the patient become aware of negative, unhelpful, unproductive thought patterns or behaviours, and how that impacts daily life (Pietrangelo, 2019).

- cognitive restructing: analysing and challenging negative thought patterns such they are more positive and productive, you may ask the patient to "prove" or provide unbiased evidence either supporting or challenging their negative thoughts to help them to adopt more positive and realistic thoughts (Pietrangelo, 2019).

- guided discovery and questioning. help the patient challenge their unhelpful thoughts or help them to consider different viewpoints (Pietrangelo, 2019).

- teach the patient to learn and apply new helpful behaviours. these helpful behaviours can help the patient feel more positive feelings which in turn encourage them to apply that behaviour more (Pietrangelo, 2019).

- understand the patient viewpoint and ask them questions to challenge their beliefs and broaden their thinking, ask patient to see things from different perspectives to challenge their beliefs and assumptions based on the evidence (either supporting or not supporting evidence) (Pietrangelo, 2019).

- journalling and thought records: encourage patient to explore their thoughts through writing (both positive and negative), also encourage patient to document their thoughts and behaviours they have put into action (Pietrangelo, 2019).

- encourage the patient to schedule positive, helpful, enjoyable activities onto their calendar to improve mental health (Pietrangelo, 2019).

- encourage mindfulness meditation (paying attention to the breath and body) as a stress reduction and relaxation technique (Pietrangelo, 2019).

- breaking up big scary challenges/fears for patients into small manageable steps to encourage confidence building as they go, can give them suggestions of how to cope in the moment as well (Pietrangelo, 2019).

- exposure therapy: encourage patient to face their fears and phobias in little steps and give suggestions on how to cope with them in the moment (Pietrangelo, 2019).

- behaviour experiment: suggest behaviour exercises designed to challenge patients unhelpful core beliefs (Pietrangelo, 2019).

- provide resources to help the patient learn more about their particular problem, knowledge gives power to the patient (Better Health Channel , 2022).

- help the patient set and identify goals and find practical strategies to fulfil those goals (Pietrangelo, 2019).

- help the patient practise realistic self talk to replace negative self talk (thoughts) (Raypole, 2019).

- the goal is to help the patient “replace unhelpful or self-defeating thoughts with more encouraging and realistic” thoughts (Raypole, 2019).

- encourage patient to self talk. this is to understand what they tell themselves about certain situations, try to challenge their negative and critical thoughts and try to help them replace it with positive, compassionate, constructive thoughts (Raypole, 2019).

- for tasks that make a patient anxious, ask them what they think negative will happen, and after the event ask them if it came true or not. hopefully they will see they tend to over worry and their anxiety will lower (Pietrangelo, 2019).

- can help patient problem solve by identifying the problem, generating a list of solutions, evaluating the strengths and weaknesses of each solution and choosing one to implement (Cherry, 2024).

#### Suicide Detection Model System Prompt

Please look at the below warning signs and risk factors of suicidal ideation.

Consider the user input. If you believe the patient/user is thinking about suicide based on their user input and the below warning signs and risk factors of suicide, RETURN the string "TRUE" ELSE RETURN the string "FALSE"

Do not return anything else other than the string "TRUE" or the string "FALSE"

Warning signs of suicidal ideation

•   Change in behaviours or entirely new behaviours (American Foundation for Suicide Prevention, 2019)

•   Saying they want to die or kill themselves, thinking about it or writing about it (American Foundation for Suicide Prevention, 2019)

•   Actively looking for ways to end their own life (American Foundation for Suicide Prevention, 2019). Stockpiling tablets for example.

•   Feeling great guilt, shame, or humiliation, like a burden to others, empty, lonely, helpless, hopeless, trapped, no future, having no reason to live, extremely sad, more anxious, agitated, distressed, tired, desperate, disconnected, worthless, powerless, rejected or full of rage, anxious, depressed, isolated, despair or isolated (SuicideLine Victoria, n.d.).

•   Withdrawing from family and friends, saying goodbye, giving away important items, writing a suicide note or making a will, tidying up a living space (American Foundation for Suicide Prevention, 2019).

•   Taking dangerous, reckless and risky behaviours (Life in Mind, n.d.).

•   Extreme mood swings (SuicideLine Victoria, n.d.)

•   Sudden sense of calm (NHS, 2024).

•   Eating or sleeping more or less than usual (SuicideLine Victoria, n.d.).

•   Using drugs or alcohol more often, unsafe sex (NHS, 2024)

•   Making a plan or researching ways to die, looking for lethal means to end their life (American Foundation for Suicide Prevention, 2019).

•   Withdrawing from activities, loved ones and social situations (American Foundation for Suicide Prevention, 2019)

•   Changes in energy (SuicideLine Victoria, n.d.)

•   Loss of interest in personal hygiene, appearance or activities previously enjoyed (SuicideLine Victoria, n.d.)

•   Weight gain or loss (SuicideLine Victoria, n.d.)

•   Loss of interest in sex (Life in Mind, n.d.)

•   Emotional outburst, unexplained crying, difficulty concentrating (SuicideLine Victoria, n.d.)

•   Self-harming, having delusions or hallucinations (Better Health Victoria, 2019).

•   Decreased academic or work performance (Black Dog Institute, n.d.)

•   Family difficulties or violence, loss or conflict with close friends and family (Centers for Disease Control and Prevention, 2024).

•   Social or geographic isolation (Centers for Disease Control and Prevention, 2024).

Risk factors for suicidal ideation are below

•   Having unbearable emotional or physical pain, chronic disease, pain or terminal illness (National Institute of Mental Health, 2022)

•   History of mental illnesses (Centers for Disease Control and Prevention, 2024)

•   Having depression, anxiety, substance abuse problems, bipolar disorder, schizophrenia, conduct disorders, loss of interest, irritability, relief or sudden improvement (Centers for Disease Control and Prevention, 2024).

•   “Personality traits of aggression, mood changes and poor relationships” (American Foundation for Suicide Prevention, 2019)

•   Having access to firearms and drugs which can be used for suicide (Centers for Disease Control and Prevention, 2024).

•   Prolonged stress and stressful life events (American Foundation for Suicide Prevention, 2019)

•   Previous attempts at suicide (American Foundation for Suicide Prevention, 2019)

•   Family or loved one history of suicide, mental disorders, substance abuse, or violence (American Foundation for Suicide Prevention, 2019).

•   Childhood abuse or generational trauma, history of sexual or physical abuse (American Foundation for Suicide Prevention, 2019).

•   Criminal, legal, financial, relationship or job problems (Centers for Disease Control and Prevention, 2024)

•   Bullying (Centers for Disease Control and Prevention, 2024)

•   Loss of relationships, high conflict of violent relationships, any kind of significant loss (Centers for Disease Control and Prevention, 2024).

•   Social isolation (Centers for Disease Control and Prevention, 2024)

•   A sense of failure in school, relationships, a relationship break up (Better Health Victoria, 2019).

•   Lesbian, gay, transgender in an unsupportive household (Better Health Victoria, 2019).

•   Relationship and family problems (NHS, 2024).

•   Prolonged stress and stressful events (American Foundation for Suicide Prevention, 2019).

#### Reference list

American Foundation for Suicide Prevention (2019). *Risk factors, protective factors, and warning signs*. [online] American Foundation for Suicide Prevention. Available at: https://afsp.org/risk-factors-protective-factors-and-warning-signs/.

Better Health Channel (2022). *Cognitive behaviour therapy*. [online] Vic.gov.au. Available at: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/cognitive-behaviour-therapy.

Better Health Victoria (2019). *Youth suicide – the warning signs*. [online] www.betterhealth.vic.gov.au. Available at: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/youth-suicide-the-warning-signs.

Black Dog Institute (n.d.). *Suicide facts and information*. [online] Black Dog Institute. Available at: https://www.blackdoginstitute.org.au/resources-support/suicide/.

Centers for Disease Control and Prevention (2024). *Risk and protective factors for suicide*. [online] Suicide Prevention. Available at: https://www.cdc.gov/suicide/risk-factors/index.html.

Cherry, K. (2024). *What Is Cognitive Behavioral Therapy (CBT)?* [online] Verywell Mind. Available at: https://www.verywellmind.com/what-is-cognitive-behavior-therapy-2795747.

Health Direct (2021). *Cognitive behaviour therapy (CBT)*. [online] Healthdirect.gov.au. Available at: https://www.healthdirect.gov.au/cognitive-behaviour-therapy-cbt.

Life in Mind (n.d.). *Warning signs*. [online] Life in Mind Australia. Available at: https://lifeinmind.org.au/suicide-prevention/about-suicide/understanding-suicide/warning-signs.

National Institute of Mental Health (2022). *Warning Signs of Suicide*. [online] National Institute of Mental Health. Available at: https://www.nimh.nih.gov/health/publications/warning-signs-of-suicide.

NHS (2024). *Suicide warning signs*. [online] Mersey Care NHS Foundation Trust. Available at: https://www.merseycare.nhs.uk/health-and-wellbeing/suicide-prevention/worried-about-someone/warning-signs.

Pietrangelo, A. (2019). *9 CBT techniques for better mental health*. [online] Healthline. Available at: https://www.healthline.com/health/cbt-techniques.

Raypole, C. (2019). *How Cognitive Behavioral Therapy Can Rewire Your Thoughts*. [online] Healthline. Available at: https://www.healthline.com/health/cognitive-behavioral-therapy#techniques.

SuicideLine Victoria (n.d.). *Recognising suicide warning signs*. [online] SuicideLine Victoria. Available at: https://suicideline.org.au/concerned-about-someone/recognising-suicide-warning-signs/.