Project Proposal - PsycGPT

Building a professional mental health Generative Chatbot using self Retrieval-Augmented-Generation

1. Introduction

The project we propose is to build a chatbot named PsycGPT, which utilizes emerging AI techniques such as Generative pre-trained transformers (GPT) and self Retrieval Augmented Generation (self-RAG) (Gola, 2024). To the best of our knowledge, it is the first of its kind in RAG-driven mental service chatbot. It is designed to act as an alternative to some of the basic psychotherapy and social works. Our goal is to reduce the expensive therapy fees and increase accessibility to mental healthcare by implementing hierarchical diagnosis with the PsycGPT. The target group of this chatbot is native Chinese speakers with a tendency of mild and moderate mental health problems.

To efficiently construct the PsycGPT, a framework called LangChain is introduced, which is generally employed for large language model-powered apps. The PsycGPT has two chains to achieve its functions: diagnosis and therapy. The Diagnosis Chain interacts with users to provide a diagnosis, which is then sent to the Therapy Chain to provide corresponding therapy. Psychological knowledge will be used to guide these processes. Therefore, this project will be a great practice of combining computational methods with psychological science to advance social good. By realizing the goals mentioned above, we could probably make it more affordable for different classes to solve mental health problems with professional therapy. In addition, we have a detailed preparation plan (see Appendix III, IV, and V as well as the rest of the proposal) to make this project realistic.

2. Background

In Hong Kong, the psychiatrist-to-population ratio is 7.55 per 100,000 persons, which is significantly below the OECD average of 18.00. The local CP-to-population ratio is 8.15 per 100,000 persons, while the OECD average is 53.00 (OECD, 2021). The shortage of mental health resources has been the main challenge the government is facing. To reduce the burden on the mental health system, some AI chatbot platforms have been implemented (Wasil et al., 2022). However, as we surveyed these platforms, most of the existing Apps are for English speakers and lack a Chinese version. So we decided to train the PsycGPT with Chinese data to better serve our target group who are native Chinese speakers. Meanwhile, most of them are pre-programmed Apps and unable to interact with users flexibly as a real therapist. With the equipment of updated techniques like GPT and self-RAG, it is promising that our project can well satisfy the demands of mental health resources and reduce costs.

3. Computational Approach

Digitalization of psychotherapy seems to be a promising future. As pointed out by Wasil et al. (2022), there are thousands of smartphone apps for mental wellbeing including medication, journaling and interaction with chatbot. Specifically, we would like to draw

attention to chatbots for their potential in mental health care. Firstly, as shown by Lau et al. (2022), chatbot-delivered psychotherapy can significantly improve depressive symptoms in clinical settings. Secondly, these chatbots could be widely accessible and affordable through personal devices like smartphones, which reduces the cost of seeking support (Vierhile et al., 2017). Lastly, previous research has demonstrated that individuals are willing to disclose their emotional thoughts with machines (Ho et al., 2018). These findings highlight the potential of psychological AI chatbots as an alternative method for conducting large-scale early intervention for mental well-being.

However, chatbots based on traditional natural language processing have many limitations. For example, most of them cannot deal with variations and are poorly adapted to new topics. In other words, most traditional chatbots are trained on a small data set and rely heavily on **a set of predefined response options** by programmers, which results in unnatural responses, especially in more complex conversations. Fortunately, recent advances in transformer-based large language models (LLMs) have overcome these problems. One famous example is OpenAI's ChatGPT. Simply speaking, the transformer-based model can generate new responses that did not appear in training datasets, thus enabling a more flexible and natural manner when dealing with complicated conversations like psychotherapy.

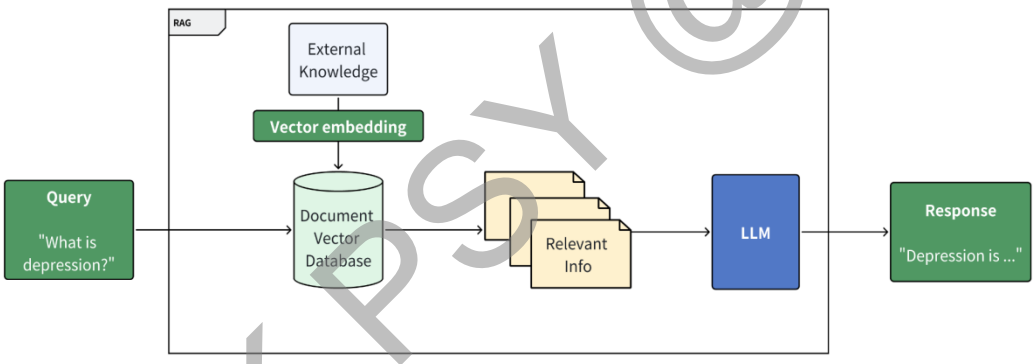

Figure 1: Basic architecture of Naive RAG

As discussed, we believe a transformer-based LLM that focuses on mental health care can greatly alleviate the burden on limited mental health care resources and actively contribute to the mental well-being of the public. Nevertheless, building such a model from scratch is not realistic in terms of time and budget. Therefore, we decided to build on top of existing pre-trained LLM and use self retrieval-augmented-generation to modify the model in the psychotherapy domain. This technique belongs to the broader category of retrieval augmented generation (RAG) that was first introduced by Lewis et al. (2020) as a means of enhancing LLM's output. As shown in Figure 1, RAG enables LLM to have an external knowledge source stored in its document vector database. Before proceeding to answer, LLM will perform a relevancy search that encodes the user's query into vector representations and compares it with a vector database to find the most relevant information. Then, LLM will provide user responses based on the external relevant information (see Appendix I for the full mathematical algorithm), thereby equipping LLM with abilities to handle knowledge-intensive tasks that are not trained before, such as psychotherapy.

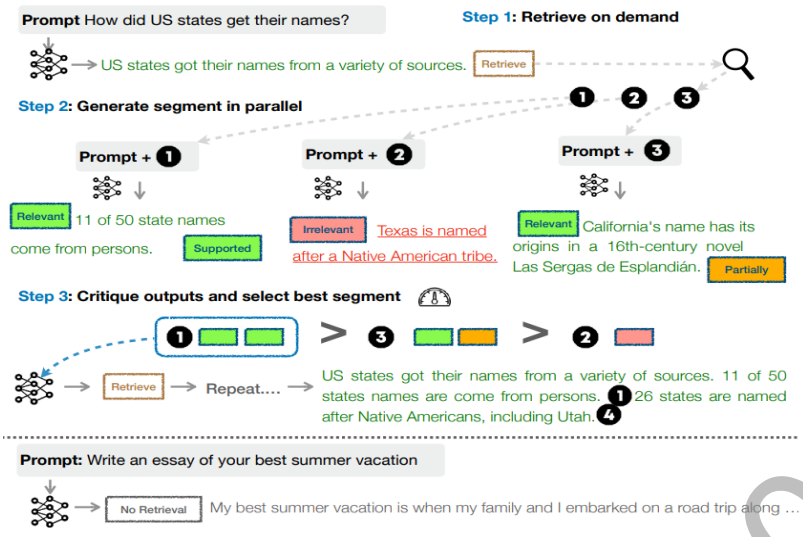

Figure 2: Advanced architecture of self-RAG

As for self-RAG, this comes from a recently published RAG method (Asai et al 2023). Compared with the original RAG, it adds multiple self-assessments to further improve the model's performance. As shown in Figure 2, self-RAG first determines whether a question needs to be retrieved from an external knowledge base (step 1: Retrieve on demand). If the current question can be answered by the original LLM model, the whole RAG process will be skipped. If RAG search is needed, the self-RAG algorithm will use different prompts to separate into segments for multiple searches and assess them based on two criteria: first, whether it is relevant to the question asked, and second, whether it supports potential output that could be generated. Each segment will get a quantitative score based on these criteria, and the information obtained from the best segment will be used for the final output (see Appendix II for full mathematics algorithm). Research has shown that models adopting self-RAG avoid generating useless responses, and because LLM requires that responses are grounded in retrieved evidence, it avoids outputting factually incorrect content. which is important for the quality and reliability of psychotherapy provided by it.

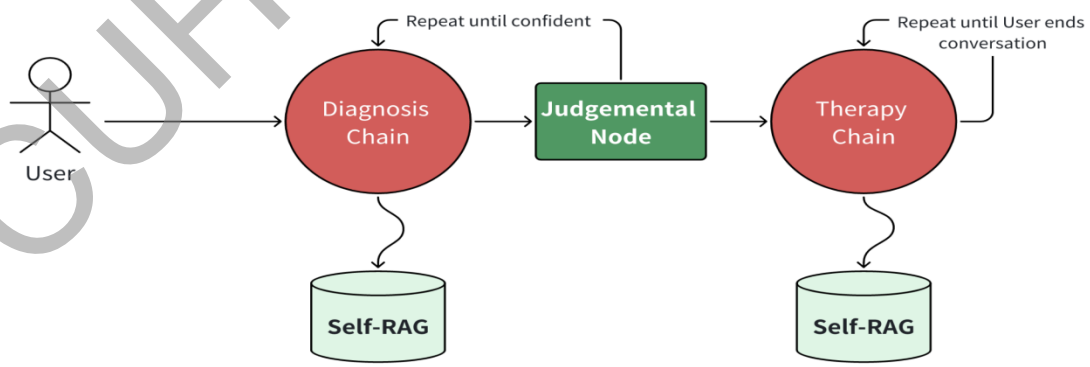

Figure 3: Proposed "Sequential-Chain" architecture of PsycGPT

Therefore, we decided to **implement self-RAG and feed external knowledge in psychotherapy to OpenAI's ChatGPT-4 model to create a new model called PsycGPT** using LangChain (Auffarth, 2023; Lim, 2023), an open-source platform for LLM

development. This platform enables us to build, evaluate, monitor and deploy LLMs. Specifically, PsycGPT is divided into two phases (or "Chains"): Diagnosis Chain and Therapy Chain. In the first stage, PsycGPT will determine the user's current tendency of psychological problems (depression, anxiety, etc.) through a continuous dialogue with the user. In the second stage, PsycGPT will choose the most suitable therapeutic approach and intervene accordingly. By using the sequential chain function provided by LangChain, we can connect the two stages and add a judgmental node in between to make sure that PsycGPT will only proceed to the next Therapy Chain if the diagnosis chain is confident enough to determine the user's tendency in certain psychological problems.

4. Psychological approach

Through the preceding discussion, it becomes clear that the most crucial aspect of creating PsycGPT by self-RAG lies in the quality of external knowledge. Therefore, the subsequent discussion will delve into why specific psychological knowledge is chosen.

In terms of the Diagnosis Chain, we utilize the International Classification of Diseases, Tenth Revision (ICD-10) as the diagnosis criteria. The ICD-10 is a standard classification system published by the World Health Organization for coding and classifying diseases and health-related problems. According to a survey of psychiatrists' attitudes towards mental disorders classification, for psychiatrists from about 44 countries, about 70% of respondents used ICD-10 in clinical work, while only 23% used DSM-5 as criteria (Reed et al., 2011). The PsycGPT will provide a reliable diagnosis based on the symptoms and criteria described in ICD-10.

For the Therapy Chain, a widely used therapy method called Cognitive behavioral therapy (CBT) is employed in our PsycGPT. A meta-analysis has demonstrated the efficiency of CBT on mental diseases like depression, bipolar disorder, schizophrenia and anxiety disorders (Hofmann et al., 2012). Moreover, some attempts have been made in this domain. Denecke et al. (2022) conducted a literature review to check the implementation of CBT in e-mental health apps and found it is reasonable to employ CBT techniques as electronic mental health interventions. With the great ability of Human-computer interaction, PsycGPT can perform better than traditional apps using predefined codes to conduct CBT. Thus, a wide range of CBT professional handbooks will be given as GPT's knowledge database. This includes LICBT (Low-Intensity CBT), BCBT(Brief CBT), CEBT (Cognitive Emotional Behavioral Therapy) and so on.

5. Expected outcome

5.1 Evaluation of the efficacy of PsycGPT

Since the goal of PsycGPT is to provide therapy to individuals with mild and moderate mental health problems, we will recruit some subjects with certain mental disorders to conduct controlled experiments. We plan to recruit subjects with mild tendencies to common psychological problems that exist in Hong Kong (depression, anxiety disorder and

bipolar disorder) to examine the efficacy of PsycGPT. For each disorder, subjects are assigned to two groups using block randomization. In terms of the control group, subjects will be provided with a normal version of ChatGPT to get therapy for 2 weeks. Then for the experiment group, they are required to use our PsycGPT to get the therapy for 2 weeks. Each subject needs to finish a self-report assessment for their mental well-being before and after the therapy. The 7-item Generalized Anxiety Disorder scale (GAD-7) is chosen for testing anxiety disorder (Catuara-Solarz et al., 2022). The Clinically Useful Depression Outcome Scale (CUDOS) is used for assessing depression remission (Zimmerman et al., 2004). The General Behavior Inventory (GBI) is for examining bipolar symptoms (Youngstorm et al., 2021). We will compare the results of the pretest and posttest to evaluate the efficacy of the PsycGPT. It is promising that the subjects in the experiment group will have better mental health and well-being after the treatment.

5.2 Social influence

As mentioned before, our project can benefit a lot in reducing the burden of the mental health care system and the possible costs of therapy fees. In Hong Kong, individuals can access mental health services in public hospitals for a relatively low price. Nevertheless, what they have to face are the long waiting lists and short consultation times (Tse et al., 2010). Our project can significantly relieve the existing dilemma as it just needs a small cost to run a platform and it is easy for a large amount of users to utilize the service at the same time. As the PsycGPT serves as the primary stage of the hierarchical diagnosis, the burden of the current mental health care system could be well eased.

Moreover, the implementation of PsycGPT could also reduce the stigma for individuals to use mental health services. It is evident that the mental disorder stigma can significantly impact on seeking mental health care (Corrigan et al., 2014). Through the online platform and human-computer interaction, the personal information and privacy can be well protected. Individuals who would feel discomfort and stigma to see a therapist can use the service at any time without the potential fear (D'Alfonso, 2020).

We will contact potential institutions like counseling centers of universities, psychiatric hospitals and NGOs for mental health to seek funds and further cooperation.

5.3 Uniqueness

The main difference between our PsycGPT and other chatbots is the emerging techniques we use to construct it. While the existing Apps are mostly based on predefined codes and dialogues, we employ the excellent large language model, ChatGPT, to realize a smooth and flexible human-computer interaction. Meanwhile, the RAG technique enables our chatbot to have external and professional knowledge in the psychological domain, which is quite different from the original ChatGPT model. We can also update the knowledge database whenever it is needed. Furthermore, it can reduce the hallucination issues which are common in LLM. The PsycGPT would have significantly higher performance when acting as a therapist.

REFERENCES

Key references:

Auffarth, B. (2023). Generative AI with Langchain: Build large language model (LLM) apps with python, chatgpt, and other llms. *Packt Publishing*.

Gola, A. (2024). Self-reflective rag with langgraph. *LangChain Blog*. https://blog.langchain.dev/agentic-rag-with-langgraph/

Lim, G. (2023). Langchain Crash course. *Greg Lim*.

Other references

Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv*. https://doi.org/10.48550/arXiv.2310.11511

Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2014). The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest, 15*(2), 37-70.

Catuara-Solarz, S., Skorulski, B., Estella-Aguerri, I., Avella-Garcia, C. B., Shepherd, S., Stott, E., ... & Dix, S. (2022). The Efficacy of "Foundations," a Digital Mental Health App to Improve Mental Well-being During COVID-19: Proof-of-Principle Randomized Controlled Trial. *JMIR mHealth and uHealth, 10*(7), e30976.

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology, 36*, 112-117.

Denecke, K., Schmid, N., & Nüssli, S. (2022). Implementation of cognitive behavioral therapy in e–mental health apps: Literature review. *Journal of medical.*

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), e19

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*, 427-440.

Lim, S. M., Shiau, C. W. C., Cheng, L. J., & Lau, Y. (2021). Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: A systematic review and meta-regression. *Behavior Therapy, 53*(2), 334-347.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih,W., Rocktäschel, T., Riedel, S., Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv.* https://doi.org/10.48550/arXiv.2005.11401

OECD. (2021). Fitter Minds, Fitter Jobs: From Awareness to Change in Integrated Mental Health, Skills and Work Policies. Mental Health and Work, *OECD Publishing, Paris*, https://doi.org/10.1787/a0815d0f-en.

Reed G. M., Correia J., Esparza P., Saxena S., Maj M. (2011). The WPA-WHO global survey of psychiatrists' attitudes towards mental disorders classification. *World Psychiatry, 10*, 118–131.

Tse, S., Cheung, E., Kan, A., Ng, R., & Yau, S. (2012). Recovery in Hong Kong: Service user participation in mental health services. *International Review of Psychiatry, 24*(1), 40-47.

Wasil, A. R., Palermo, E. H., Lorenzo-Luaces, L., & DeRubeis, R. J. (2022). Is there an app for that? A review of popular apps for depression, anxiety, and well-being. *Cognitive and Behavioral Practice, 29*(4), 883-901.

Youngstrom, E. A., Perez Algorta, G., Youngstrom, J. K., Frazier, T. W., & Findling, R. L. (2021). Evaluating and validating GBI mania and depression short forms for self-report of mood symptoms. *Journal of Clinical Child & Adolescent Psychology, 50*(5), 579-595.

Zimmerman, M., Posternak, M. A., & Chelminski, I. (2004). Using a self-report depression scale to identify remission in depressed outpatients. *American Journal of Psychiatry, 161*(10), 1911-1913.

Appendix I: Mathematical Principle of Naive RAG

1. **Vector embedding:**

For vector embedding of external knowledge into the model's own database. We use **OpenAI's word embedding model** that is trained using Matryoshka Representation Learning.

$$\min_{\{W^{(m)}\}} \frac{1}{N} \sum_{i \in [N]} \sum_{m \in M} c_m \cdot L\left(W^{(m)} \cdot F(x_i; \theta_F)_{1:m}; y_i\right)$$

Where:

- L is the multi-class softmax cross-entropy loss function.
- $c_m$are the importance scales for each nested dimension □.
- $W^{(m)}$ is the parameter of the linear classifier for each nested dimension.
- $F(x_i; \theta_F)_{1:m}$is the feature representation of input x_i up to dimension m
- y_i is the true label of input x_i
- N is the number of samples
- M is the set of nested dimensions

Simply speaking, a word embedding model trained using Matryoshka Representation learning can maintain accuracy while providing smaller embedding sizes, which leads to greater efficiency in terms of memory and computational resources. Moreover, it exhibits strong adaptability, capable of dynamically adjusting the dimensions of the embeddings according to the computational resource needs of different tasks. This made OpenAI's model particularly suitable for our case.

2. **Indexing:**

For matching the user query with the knowledge database, we will use simple **Cosine similarity**:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Where:

- A and B represent 2 different vectors
- $A \cdot B$ represent the dot product of 2 vectors
- $\|A\|$ and $\|B\|$ represent their Euclidean norm, or their length

Intuitively, Such an indexing algorithm will check which knowledge vector has the most similar detection in their vector space and find the most relevant information for LLM's reference.

Appendix II: Mathematical Principle of Self-RAG

1. Critic Model

The core of Self-RAG relies on an important model called the critic model. The function of this model is to predict whether a user query requires RAG search, and if so, predict whether the searched information is relevant or supportive. These prediction contents are called tokens (retrieve token, IsRel token, IsSup token). The retrieve token is for judging if RAG search is needed. IsRel token is for evaluating whether information is relevant or not. And IsSup token is for evaluating whether information is supportive of the chatbot's potential answer or not. The mathematical expressions of critic model are as follows:

$$\max_M E_{(x,y,r)\sim D_{\text{gen}}} \log p_M(y, r|x).$$

Where:

- $\max_M$ is the optimization goal, which is to maximize the following expectation with respect to the model
- $E_{(x,y,r)\sim D_{\text{gen}}}$ denotes the expected value over the distribution of data samples (x, y, r) drawn from a dataset D_{gen}, a training dataset prepared by the original paper. Here, x represents input data, y represents the target output, and r represents reflection tokens that are part of the output.
- is the natural logarithm of the probability p_M assigned by the model $log p_M(y, r|x)$ M to generating the output y and reflection tokens r, given the input x.

Then, the tree-decoding algorithm will use those critique tokens to make inferences and final evaluation score:

$$SS(\text{Critique}) = \sum_{G\in\{\text{ISREL, ISSUP, ISUSE}\}} w_G s_G$$

Where:

- $S(\text{Critique})$ is the score contributed by critique tokens, which is a weighted sum of scores for each critique group G. This score aims to adjust the segment score based on the critique tokens.
- w_G The weight for the critique group G, allowing customization of how much each critique aspect contributes to the overall score.
- s_{G_t}The score for the critique group G at segment t, which represents the model's assessment of the segment's quality in terms of the specific critique aspect

Then, $S(\text{Critique})$ will be used in the next stage of computation:

$$f(y_t, d, \text{Critique}) = p(y_t|x, d, y_{<t}) + S(\text{Critique})$$

Where:

- f(y_t, d, \text{Critique}) computes the score of a segment y_t given the passage d and the critique tokens. It aims to measure how well the segment y_t aligns with the desired criteria, including relevance, support, and usefulness.
- $p(y_t|x, d, y_{<t})$ is the probability of generating segment y_t given the input x, passage d, and all previous segments $y_{<t}$

Appendix III: Timeline

Period	Content
February 2024	-literature review for related topics (Done)
March 2024	-proposal writing and submission (Done) -construct basic functions of the PsycGPT -purchase API from OpenAI -train the model with psychological databases -design UI for the platform
April - June 2024	-seek for faculty member in clinical psychology as a supervisor -ethics review -seek for funds -subjects recruitment -efficacy examination
July 2024	-register a domain name and purchase web hosting service -upload the platform to the server -seek potential cooperation with school counseling centers or NGOs in both Hong Kong and Mainland China -promotion on social media
August 2024	-formal launch and operation -seeking for potential sponsors
September 2024 -	-daily Maintenance

Appendix IV: Budget Plan

Content	HKD	Remarks
API credit	5,000	based on usage of the platform: 10 USD/ 1M tokens for input and 30 USD/ 1M tokens for output
Domain and server	2,000	the price for one year
Subjects recruitment	3,000	pay 100 HKD for each subject and recruit 30 subjects in total
Promotion-related fees/reserve	~1000	Act as a general surplus reserve in the potential use of promotion (like organizing activities to promote)
Total budget	11,000	

Appendix V: Common Q&As

Q1: How can you ensure the reliability and effectiveness? This is a clinical-target project so there might be a lot of risk.

A1: We are well aware of the risk of psychological intervention/therapy. That is why before put into use. We want to use experimental design to test the effectiveness of PsycGPT. We aim to find a faculty member in clinical psychology to supervise us in applying for the department's approval of clinical research. **Moreover,** we try to target mild to moderate patients only. We will specifically notify users to seek immediate help if they are in a very poor mental state.

Q2: How will you deal with the technical questions that may be encountered? LLM development seems to be very advanced for undergraduate students studying psychology.

A2: Professor Kwong-cheong Wong from data science and policy study is currently our supervisor. He is an expert in natural language processing and can assist if needed. Furthermore, LLM development is not as difficult as it may sound like. With reference to **key references**, it is easy to build and deploy the model.

Q3: What is the uniqueness of this project? Can't ChatGPT just do everything? Why train another model?

A3: ChatGPT, as well as other chatbots that are trained on a wide range of corpus. Are task-general chatbot. To deploy it in task-specific content, we need task-specific knowledge. In other words, we need knowledge in psychology to build better chatbots. Moreover, compared to other chatbots, we use the latest **Self-RAG** which has gained a lot of attention from the industry recently. The use of this is expected to improve the performance of the chatbot significantly. So far, this project is the first of its kind that uses Self-RAG to provide psychological services.