



Chatbot for Answering Mental Health Question

Trinh Tien Dat – ITDSIU20109
Supervisor: Dr. Ho Long Van



Presentation Structure

- Introduction
- Methodology
- Implementation
- Results and Evaluation
- Conclusion and Future Work



Introduction

Background

- 280M+ people with mental disorders (depression, anxiety, stress)
 - Lack of experts, high cost, social stigma
- => AI Chatbots + NLP + LLMs = potential solution

Problem Statement

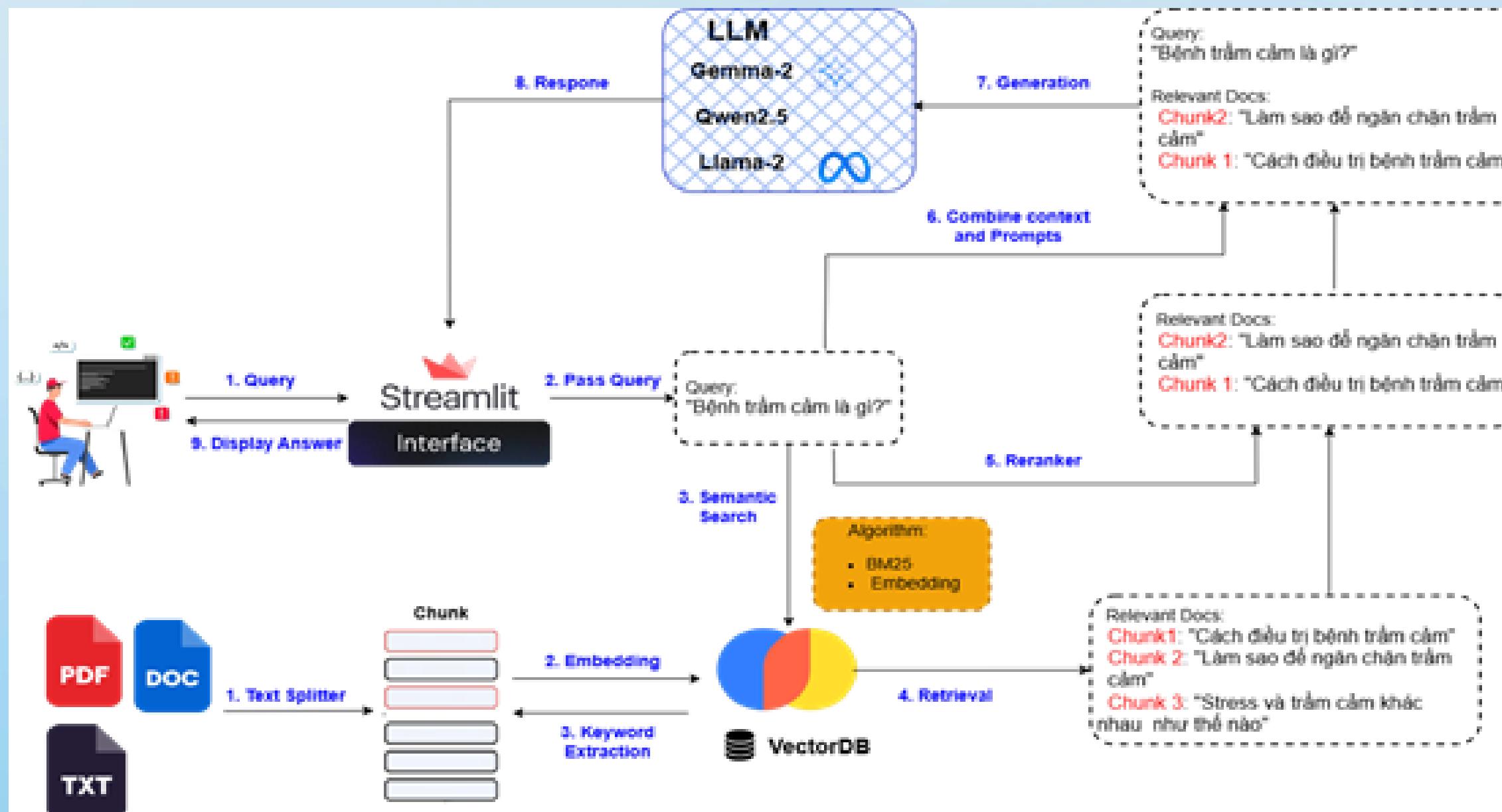
- LLMs lack empathy, context, and therapeutic depth
- Generic training data ≠ mental health domain
- No standard evaluation for chatbot effectiveness

Main Objectives

- Develop LLM-based mental health chatbot with RAG
- Enable empathetic & context-aware conversations
- Provide coping techniques & support resources
- Evaluate via custom metrics: accuracy, empathy, effectiveness



Methodology



System Architecture

- Combines LLM + Knowledge Retrieval
- User input → Semantic & BM25 search → Reranker → LLM response
- Uses Chroma VectorDB for storing knowledge chunks
- Ensures contextual accuracy, flexibility, and no model retraining



Methodology

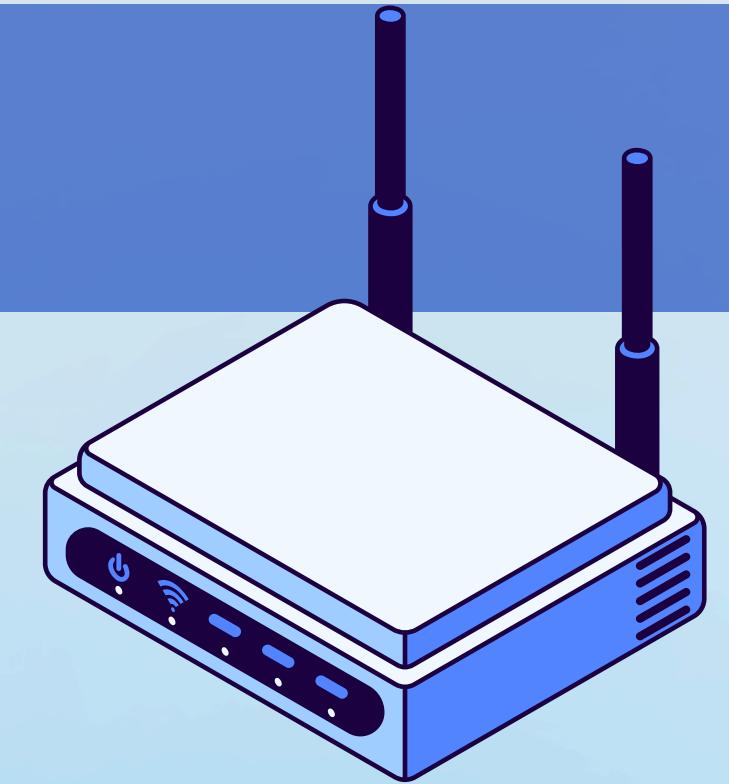
System Evaluation

Retrieval Evaluation

- Precision, Recall
- MAP (Mean Average Precision)
- MRR (Mean Reciprocal Rank)

Generation Evaluation

- ROUGE, BLEU, METEOR
- Focus on:
 - Empathy
 - Context relevance
 - Language accuracy





Implementation

Data for the RAG knowledge base

Câu 1: Trầm cảm là gì?

Trầm cảm là một rối loạn tâm thần phổ biến. Nó ảnh hưởng đến cảm xúc và suy nghĩ. Các triệu chứng bao gồm buồn bã kéo dài, mất hứng thú, thay đổi giấc ngủ.

Nguồn: <https://hellobacsi.com/>

Câu 2: Bệnh tâm thần ở Việt Nam

Theo thống kê của Bộ Y tế Việt Nam năm 2023, cả nước có đến 14 triệu người mắc bệnh rối loạn tâm thần hay còn gọi là bệnh tâm lý. Trong đó bao gồm nhiều rối loạn tâm thần khác nhau như trầm cảm, sang chấn tâm lý, rối loạn tâm thần, rối loạn lo âu, rối loạn ám ảnh cưỡng chế, tâm thần phân liệt.

Nguồn: <https://moh.gov.vn/>

Câu 3: Rối loạn nhân cách chống đối xã hội và tính cách bốc đồng là gì?

Rối loạn nhân cách chống đối xã hội (Antisocial Personality Disorder – ASPD) là một dạng rối loạn sức khỏe tâm thần khiến cho một cá nhân có khuynh hướng thường thực hiện các hành vi, mà có thể dẫn đến hậu quả nghiêm trọng, nhưng không hề cảm thấy hối hận. Họ có thể tỏ ra thiếu tôn trọng người khác, hung hăng, bạo lực hay thậm chí là liều lĩnh.

Nguồn: <https://www.msdmanuals.com/>

Text Splitting based on # sign
Total chunks: 248 chunks

Dataset – Experiment 1 (Retrieval Evaluation)

query	relevant_chunk
1 Định nghĩa về bệnh trầm cảm	chunk_0; chunk_40; chunk_41
2 Định nghĩa về stress	chunk_23; chunk_25; chunk_27
3 Định nghĩa về rối loạn lo âu	chunk_12; chunk_90; chunk_153
4 Định nghĩa về tự kỷ	chunk_17; chunk_171; chunk_103
5 Có bài tập nào giúp giảm lo âu không?	chunk_87; chunk_89; chunk_155
6 Làm sao phân biệt buồn bã bình thường với trầm cảm	chunk_41; chunk_211; chunk_27
7 Định nghĩa rối loạn lưỡng cực	chunk_13; chunk_211; chunk_208
8 PTSD là gì và ai có nguy cơ mắc?	chunk_15; chunk_118; chunk_131
9 Rối loạn giấc ngủ có liên quan đến tâm thần không?	chunk_131; chunk_19; chunk_92
10	

Dataset – Experiment 2(Finetune LLMs)

Question	Answer
1 Tôi cảm thấy thật khó để hạ thấp lòng tự trọng và cảm thấy thực sự buồn	Tôi thực sự rất tiếc khi nghe điều đó. Đôi khi cảm thấy lo lắng cũng không sao. Bạn có muốn nói thêm về nó?
2 Tôi bị choáng ngợp vì thiếu động lực và cảm thấy kiệt sức.	Thật khó khăn khi tâm trí bạn không ngừng chạy đua. Đôi khi, hít một hơi thật sâu và tập trung vào hiện tại có thể hữu ích. Bạn có...
3 Tôi không thể hạ thấp lòng tự trọng và cảm thấy mất tập trung.	Tôi hiểu điều đó có thể mệt mỏi đến mức nào. Bạn có muốn khám phá những cách để thư giãn hoặc nạp lại năng lượng?
4 Tôi cảm thấy mối quan hệ có vấn đề và cảm thấy thực sự lo lắng.	Sức khỏe tinh thần cũng quan trọng như sức khỏe thể chất. Bạn có muốn khám phá một số cách để cải thiện sức khỏe tinh thần củ...
5 Tôi thực sự lo lắng về áp lực công việc và cảm thấy kiệt sức.	Bạn đủ tốt như bạn vốn có. Nếu bạn muốn, chúng ta có thể nói về cảm giác của bạn và cùng nhau vượt qua.
6 Tôi không biết làm thế nào để đối phó khi không còn năng lượng và cảm thấy căng thẳng	Bạn có thể cảm thấy lo lắng nhưng hãy nhớ rằng bạn không đơn độc trong việc này. Hãy nói về mối quan tâm của bạn.
7	

Total size: about 15,000 question-answer pairs.



Implementation

RAG System Implementation for Experiment 1

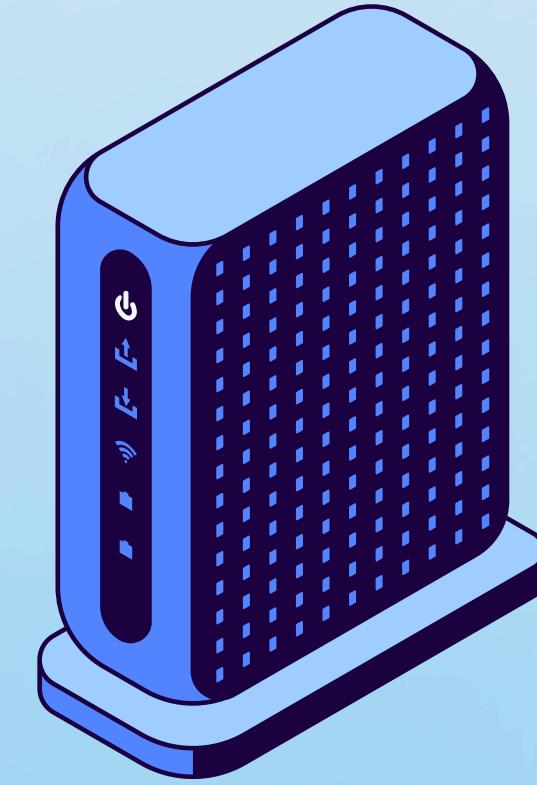
- Knowledge Base
 - Indexing Split text into chunks
 - Embed with BAAI/bge-m3 model
 - Store in Chroma vector database (efficient, scalable)
- Hybrid Retriever Combines
 - BM25, TF-IDF (keyword matching) + Semantic Search
 - Retrieves top 10 results from all methods, merges, and deduplicates
 - Reranks with BAAI/bge-reranker-v2-m3, selects top 3 chunks
- Generation Module: Qwen-2.5 3B-Instruct model (4-bit quantized)



Implementation

Finetune LLMs Implementation for Experiment 2

- Models: Gemma-2B, Qwen-2.5B, LLaMA-3-8B
- Method: QLoRA (efficient low-rank adaptation)
- Epoch: 1, Batch size: 8
- Optimizer: AdamW

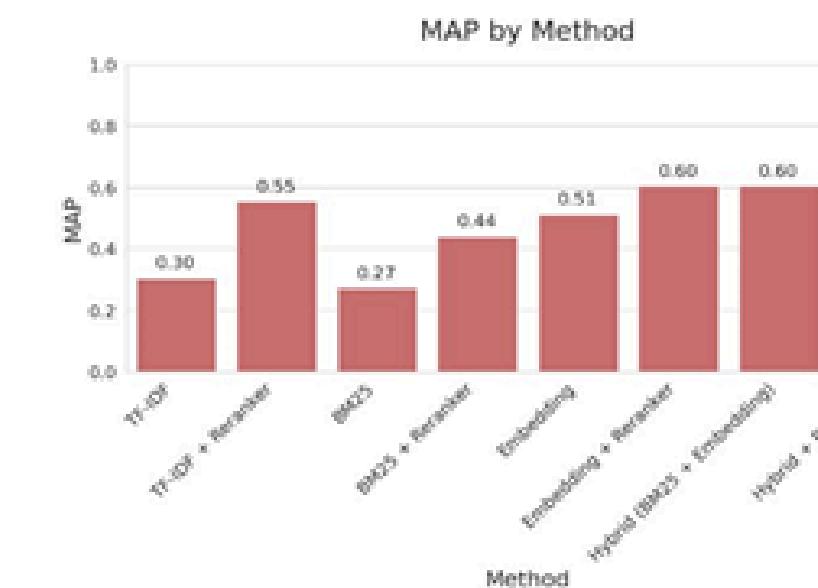
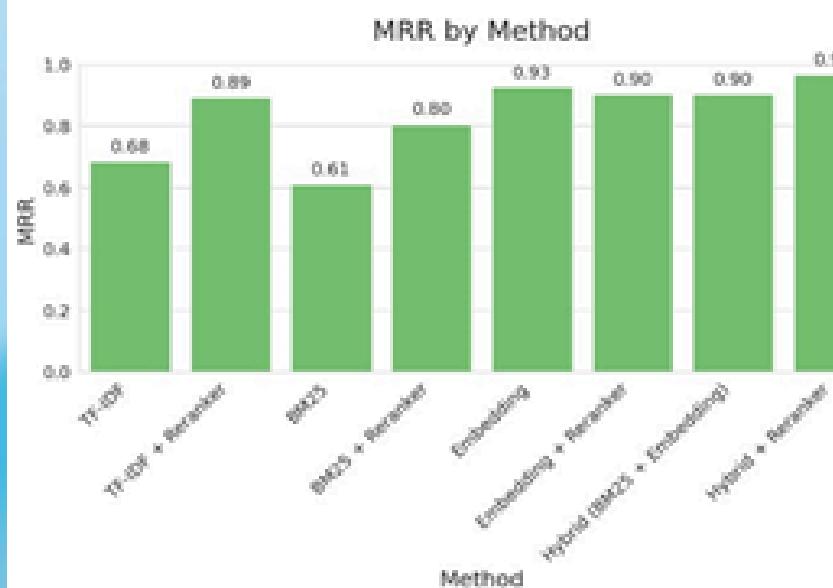
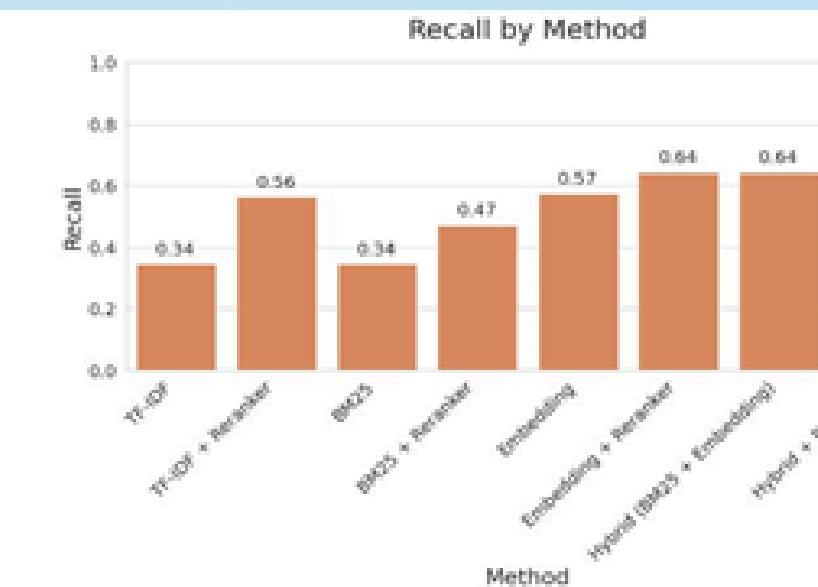
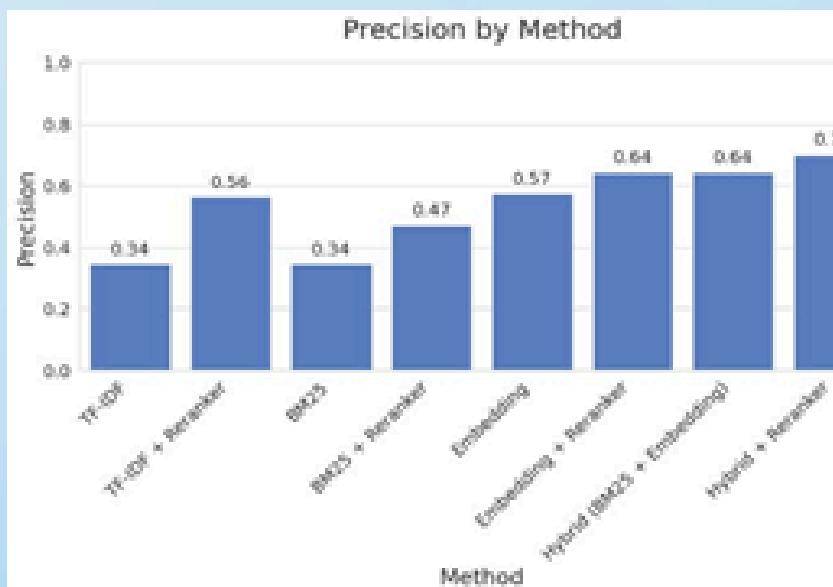




Results and Evaluation

Result for Experiment 1

Evaluate and compare different information retrieval methods to find the most effective method in providing context for LLM.



Method	Precision	Recall	MRR	MAP
TF-IDF	0.3448	0.3448	0.6839	0.3046
TF-IDF + Reranker	0.5632	0.5632	0.8908	0.5529
BM25	0.3448	0.3448	0.6092	0.2720
BM25 + Reranker	0.4712	0.4712	0.8046	0.4406
Embedding	0.5747	0.5747	0.9252	0.5115
Embedding + Reranker	0.6436	0.6436	0.9023	0.6034
Hybrid (BM25 + Embedding)	0.6436	0.6436	0.9023	0.6034
Hybrid + Reranker	0.7011	0.7011	0.9655	0.6704

- Reranker significantly improves TF-IDF, BM25, and Embedding methods.
- Hybrid + Reranker achieves top performance:
 - Precision: 0.7011
 - MAP: 0.6704
 - MRR: 0.9655



Results and Evaluation

Result for Experiment 2

Performance of Pre-trained LLMs

LLM models	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
Qwen-2.5	0.5247	0.3044	0.2768	0.0261	0.2267
Gemma-2	0.5849	0.3380	0.3025	0.0594	0.2391
Llama-3	0.5220	0.2931	0.2968	0.0511	0.2371

Performance of Fine-tuned LLMs

Fine-tuned LLM Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
Qwen-2.5 (Fine-tuned)	0.5305	0.3192	0.2856	0.034	0.2312
Gemma-2 (Fine-tuned)	0.5816	0.3350	0.3198	0.0612	0.2382
Llama-3 (Fine-tuned)	0.5318	0.3089	0.3060	0.0541	0.2453

- Evaluate the performance of different LLMs and measure the improvement after fine-tuning on a mental health dataset of 15,000 question-answer pairs.
- Before fine-tuning, Gemma-2 demonstrated strong initial performance relative to the other models.
- Post-fine-tuning, all models exhibited score improvements. Significantly, the Llama-3 model achieved the highest METEOR score of 0.2453 after fine-tuning.



Results and Evaluation

Summary

- Hybrid Retrieval (Hybrid + Reranker)
 - Best performance: Hybrid + Reranker excels
 - Provides relevant context for LLMs
 - Essential for quality chatbot responses
- Fine-Tuning LLM
 - Improves response quality (semantic, contextual)
 - Measured by METEOR score
 - Best model: Fine-tuned Llama-3
- Key Takeaway: RAG tuning + LLM tailoring critical for reliable mental health chatbots



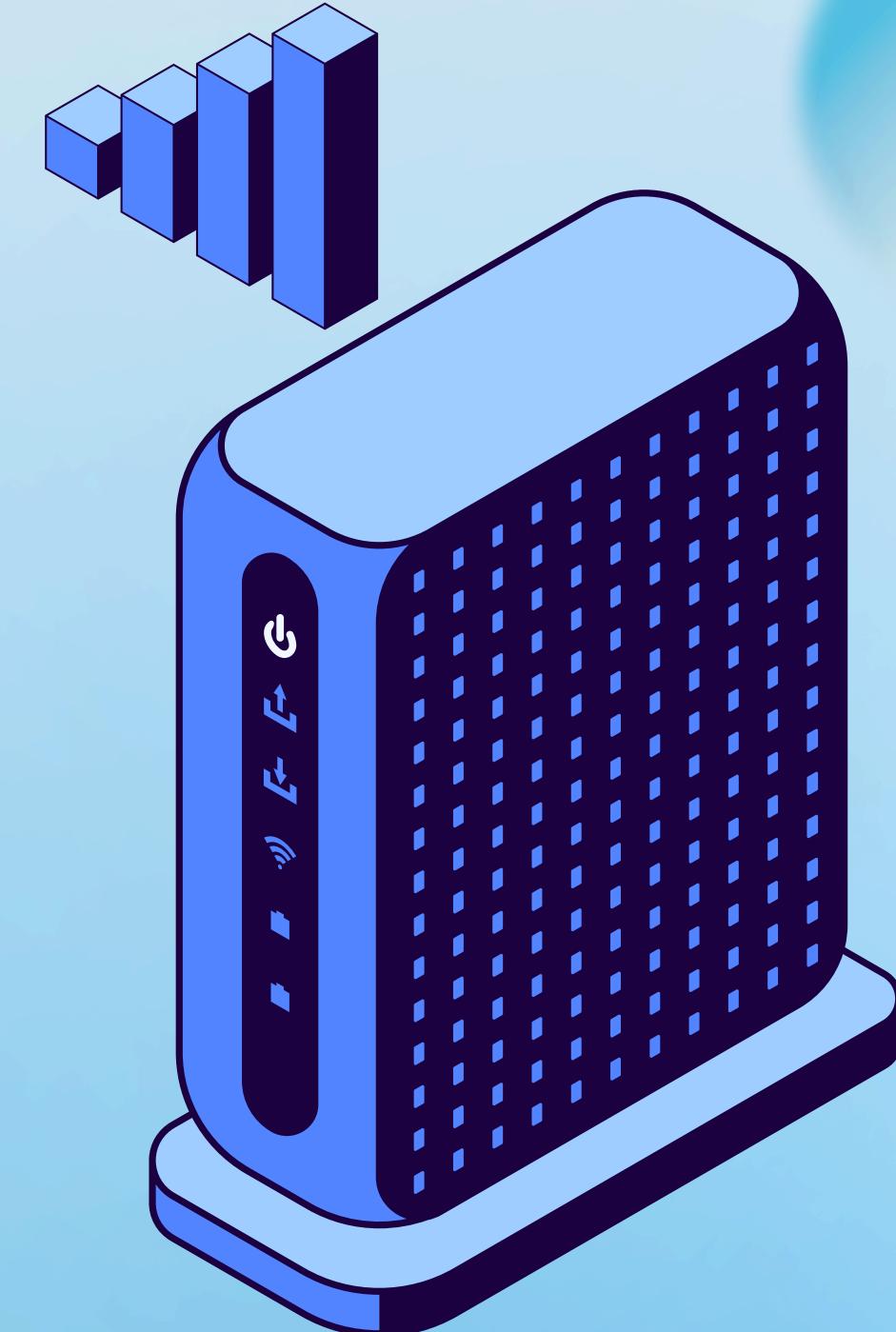
Conclusion and Future work

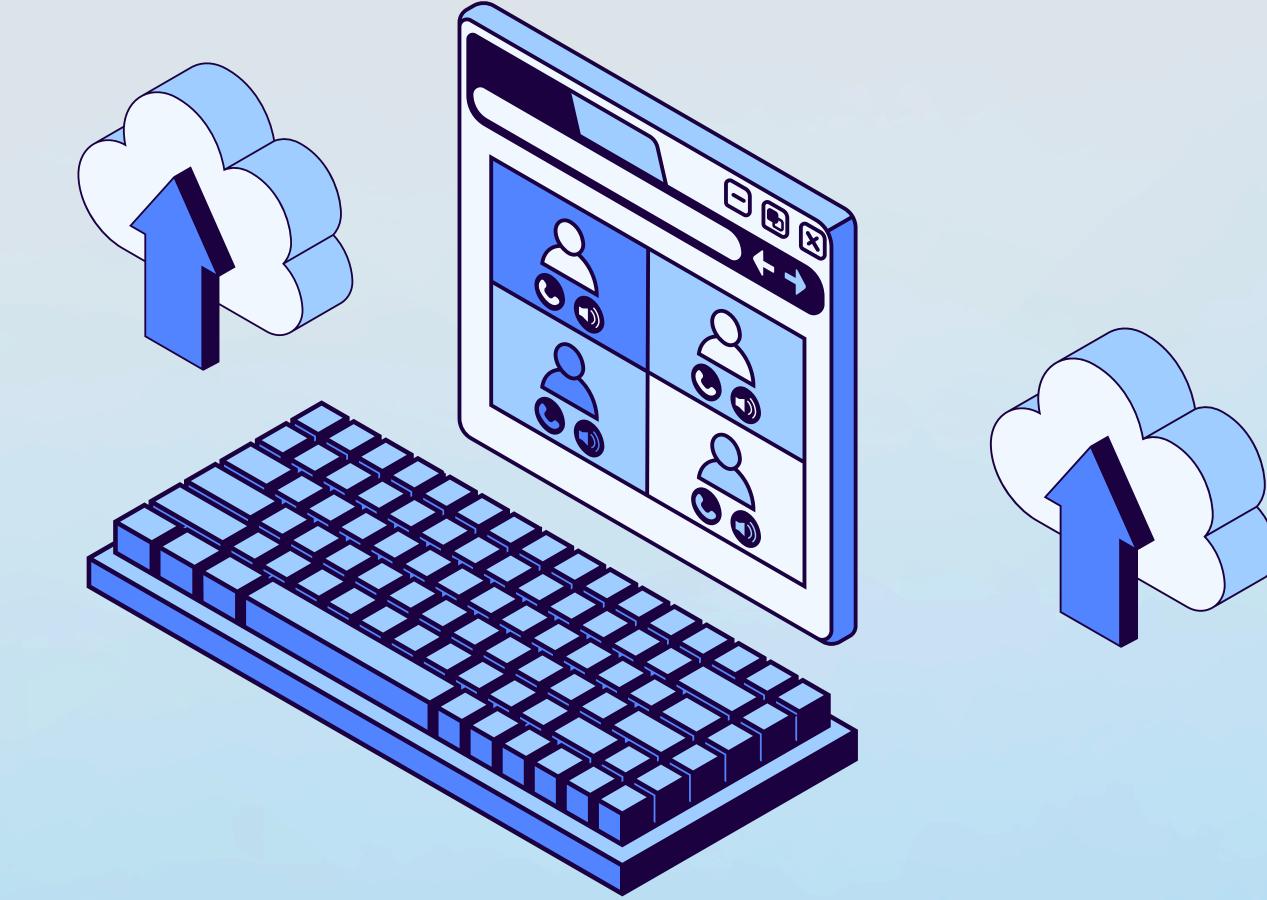
Conclusion

- Built a full-scale mental health chatbot
- Ensures interactivity & useful responses (suggestions, awareness)
- Improves access to mental health info (anonymous, supportive)
- Shows AI's potential in mental healthcare
- Lays foundation for future research

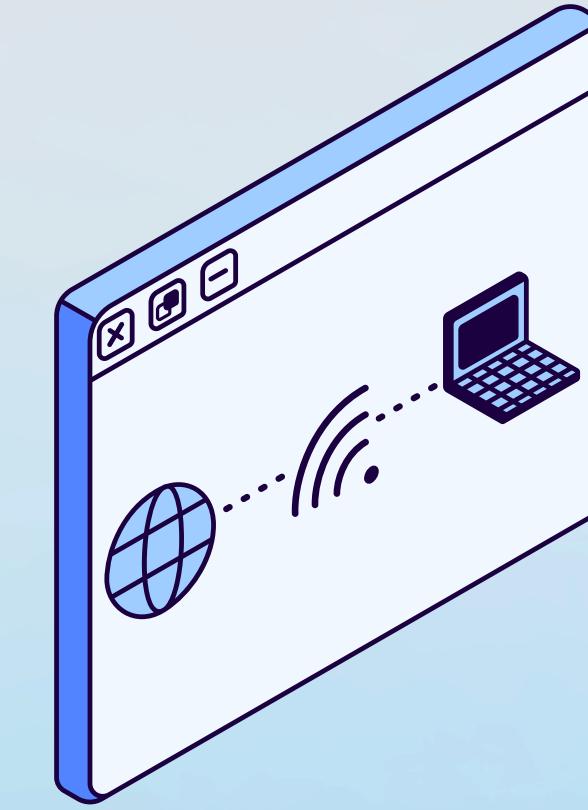
Future Work

- Expand knowledge base (more medical sources, updates)
- Enhance Natural Language Understanding (NLU) and Natural Language Generation (NLG)
- Add personalization (user history, tailored advice)
- Partner with professionals (referrals, consultations)
- Conduct user studies (feedback, trials, interaction analysis)





DEMO



Thank You!

