Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
5(2): 154–167.
DOI: 10.3934/aci.2025010
Received: 25 March 2025
Revised: 15 July 2025
Accepted: 16 July 2025
Published: 23 July 2025

Research article

Large language model enabled mental health app recommendations using structured datasets

Kris Prasad*, **Md Abdullah Al Hafiz Khan and Yong Pei**

Department of Computer Science, Kennesaw State University, Marietta, GA, 30060, USA
* **Correspondence:** Email: kprasad@students.kennesaw.edu.

Academic Editor: Jidong Yang

Abstract: The increasing use of large language models (LLMs) in mental health support necessitates a detailed evaluation of their recommendation capabilities. This study compared four modern LLMs—Gemma 2, GPT-3.5-Turbo, GPT-4o, and Claude 4 Sonnet—in recommending mental health applications. We constructed a structured dataset of 55 mental health apps using RoBERTa-based sentiment analysis and keyword similarity scoring, focusing on depression, anxiety, ADHD, and insomnia. Baseline LLMs demonstrated inconsistent total accuracy (ranging from 60% to 75%) and often relied on outdated or generic information. In contrast, our retrieval-augmented generation (RAG) pipeline enabled all models to achieve 100% accuracy, while maintaining good diversity and recommending apps with significantly better user ratings. These findings demonstrated that dataset-enhanced LLMs, regardless of being open-source or proprietary, can excel in domain-specific applications like mental health resource recommendations, potentially improving accessibility to quality mental health support tools.

Keywords: mental health applications; retrieval-augmented generation (RAG); large language models (LLMs); natural language processing (NLP); data analysis

1. Introduction

The COVID-19 pandemic and resulting lockdowns have led to a significant increase in mental health issues, mostly among teens and young adults. A brief from the World Health Organization shows that depression and anxiety rates have increased 25% worldwide, causing 90% of the countries surveyed to include mental health support in their COVID-19 response plans [1]. Despite the growing prevalence of mental health concerns, these age groups tend to remain silent about their struggles due to stigma and limited access to support systems [2].

Recently, large language models (LLMs), a type of artificial intelligence trained on vast amounts of text data to understand and generate human-like language, have shown promise in assisting with mental health support, commonly through applications such as conversational agents or resource recommendations. Prominent examples of these models include the generative pre-trained transformer (GPT) series. However, despite their potential, LLMs struggle to address more nuanced and complex mental health cases [3]. Typical LLMs' responses can lack context, depth, and accuracy, which raises concerns about their reliability in sensitive mental health applications. To improve their effectiveness, LLMs can be supplemented with domain-specific datasets, providing the necessary context and depth to enhance their recommendations.

One potential application of LLMs is improving access to high-quality mental health apps. Borghouts et al. [4] found that mental health apps can greatly improve symptoms, making them a valuable resource for individuals seeking support. However, the study also highlights several challenges that prevent users from easily finding the right apps. Factors like demographics, condition severity, and mental health stigma can make it difficult for individuals to identify and access quality resources. Additionally, the volume of available apps and the lack of quality assessments can make the selection process overwhelming, especially when each app requires a large time investment to see any payoff.

To address these challenges, we explore the potential of contextually adapted LLMs, enhanced with mental-health-specific datasets, to provide more effective and personalized app recommendations. We constructed a custom dataset that includes app descriptions, user sentiment analysis, and labeled mental health conditions to achieve this. The dataset is designed to prioritize apps that are both highly rated and explicitly intended for specific conditions, ensuring more precise and beneficial recommendations. By incorporating user reviews and developer descriptions, we can assess not only the intended purpose of an app but also its real-world effectiveness based on user feedback. This dataset serves as a critical foundation for improving LLM recommendations by reducing the risk of irrelevant or misleading app suggestions, which are common challenges when relying solely on generic artificial intelligence (AI) models.

To ensure the reliability of our recommendations, we established two key evaluation criteria:

(1) Explicit condition targeting: Recommended apps must indicate their intended use for a specific mental health condition, such as depression, anxiety, ADHD, etc.

(2) Verified effectiveness: User reviews and ratings must reflect positive feedback, indicating that the app is helpful and reliable for its targeted condition.

Finally, we integrated a retrieval-augmented generation (RAG) pipeline with our dataset and tested LLMs. This enables the models to retrieve relevant, high-quality mental health applications while eliminating irrelevant and misleading recommendations. Through this approach, we aim to develop a cost-effective, efficient, and accessible tool that can help individuals accurately navigate the abundance of digital mental health resources. Ultimately, this research contributes to the ongoing effort to reduce barriers to mental health care and improve access to reliable support through LLM-based solutions.

Our research makes the following key contributions:

- Development of a structured mental health app dataset with sentiment analysis and condition-specific labeling.
- Implementation of an RAG pipeline that significantly improves recommendation accuracy and quality.
- Demonstration that enhanced open-source models can match or exceed the performance of proprietary LLMs at lower cost.

- Empirical evidence that contextually adapted LLMs provide higher-quality mental health app recommendations based on user ratings.

2. Related works

AI has been widely explored in mental health care, particularly in clinical applications and data processing. Le Glaz et al. [5] reviewed various AI applications in mental health, including machine learning for medical imaging and models based on natural language processing (NLP), a field of AI focused on enabling computers to understand and analyze human language, for processing electronic health records. AI has proven useful in navigating large datasets, extracting key medical terms, and offering new perspectives in mental health analysis. However, the study concludes that AI is not yet at a stage where it can generate new clinical knowledge. Instead, it should be viewed as a supportive tool for improving existing mental health frameworks.

To improve mental health support provided by LLMs, we also looked at RAG. Lewis et al. [6] explored RAG, a technique that retrieves relevant documents from large external knowledge sources to enhance a model's generated outputs. Their work demonstrates that RAG models improve the specificity and factuality of generated outputs compared to non-RAG models.

This approach has also been validated in the broader medical domain; for instance, Shi et al. [7] recently demonstrated that a specialized RAG pipeline called MKRAG (medical knowledge retrieval-augmented generation) significantly improved an LLM's accuracy on medical question-answering tasks by retrieving facts from a medical knowledge base. Building on these findings, our work focuses on constructing a curated dataset and integrating an RAG pipeline to enhance LLMs' recommendations of mental health apps. We aim to improve access to accurate and high-quality mental health resources in a nonclinical, more casual setting. By designing a dataset carrying sufficient detail—eliminating the need for explicit, detailed prompting—we aim to make our resource higher quality and more accessible to a general audience.

Leivada et al. [8] also examined LLMs' reasoning and accuracy in decoding character sequences with substituted characters. Their results show that while LLMs such as GPT-4o can provide accurate decodings, their reasoning abilities are poor. In nuanced domains like mental health—where there is no single correct answer—the reasoning step is crucial for determining acceptable outputs. Without robust reasoning, an LLM might recommend mental health apps that are unlikely to be helpful or well-received. Our curated dataset is designed to mitigate such risks by limiting errors and misinterpretations, ultimately leading to more accurate and high-quality recommendations.

3. Mental health app recommendation approach

3.1. Recommendation approach summary

To improve the accuracy and relevance of mental health app recommendations, we developed a structured methodology that integrates data-driven filtering, semantic similarity analysis, sentiment classification, and RAG.

Our approach consists of four key components: First, we constructed a dataset of 55 mental health apps, including app descriptions, user reviews, and sentiment analysis. Second, we employed models based on RoBERTa (robustly optimized BERT pretraining approach), a machine learning technique for understanding natural language text, to extract relevant keywords, classify app sentiments, and determine app suitability for specific mental health conditions. Third, we implemented a vector database using FAISS (Facebook AI similarity search) [9], a library designed for efficient similarity searching within large sets of vector embeddings, to store and retrieve app-related embeddings, enabling LLMs to generate more precise and contextually relevant mental health app recommendations. Finally, we assessed the performance of four LLMs—Gemma 2, GPT-3.5-Turbo, GPT-4o, Claude 4 Sonnet—in both their baseline and RAG-enhanced configurations, using metrics such as accuracy, variety, and quality of recommendations.

This pipeline allows LLMs to provide contextually relevant and high-quality app suggestions while reducing irrelevant or misleading recommendations. The approach also optimizes computational efficiency by filtering datasets before querying the LLMs, lowering token usage and improving response quality.

3.2. Data collection and filtering

We collected data from 55 mental health apps on the Google Play Store, focusing on four conditions: depression, anxiety, ADHD, and insomnia. Apps were initially selected using keyword searches such as “depression tracker” and “anxiety relief”. To ensure relevance, we included only apps that explicitly referenced at least one mental health condition in their description.

In addition to app descriptions, user reviews were scraped and filtered using two keyword banks: condition-specific keywords such as “depression” and “ADHD therapy”, and generalized mental health terms such as “wellbeing” and “stress management”.

3.3. Data analysis

3.3.1. Semantic similarity

To assess how well each app aligned with mental health conditions, we used the all-roberta-large-v1 model [10] to generate semantic embeddings for app descriptions and user reviews. We then computed cosine similarity scores between these embeddings and the two keyword banks mentioned earlier.

App descriptions underwent keyword filtering to identify condition-specific terminology, while app reviews were processed through keyword matching. Both these processes, as illustrated in Figure 1, contributed to our final app labeling system. This dual-source approach ensured that each app was labeled based on both developer intentions (from descriptions) and user experiences (from reviews).

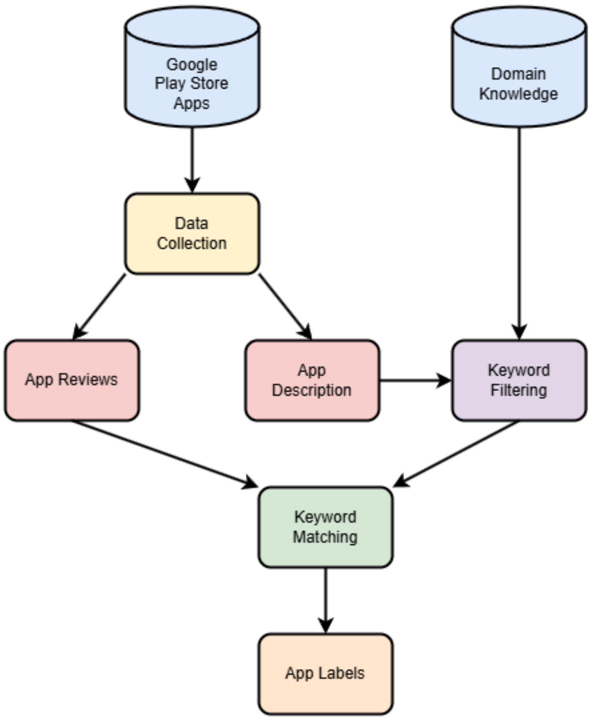

Figure 1. High-level representation of our app labeling system.

3.3.2. Sentiment analysis

We used twitter-roberta-base-sentiment from Cardiff NLP [11] to classify user reviews into positive, negative, or neutral categories. This model was chosen due to its fine-tuned performance on short-form text, making it well-suited for user-generated content.

The model's sentiment scores were processed through a softmax function to determine the most probable sentiment category. This process was applied to multiple reviews for each app, as represented by the iterative flow in Figure 2. Then, app-level sentiment scores were derived by aggregating the sentiment values across all of an app's reviews.

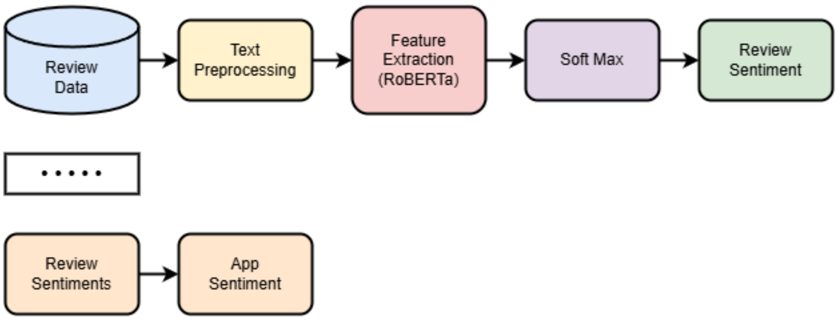

Figure 2. High-level representation of our sentiment analysis system.

3.3.3. Validation

To ensure accuracy and consistency in labeling, we cross-validated sentiment and condition labels using three locally hosted LLMs: Llama 3.2, Gemma 2, and Mistral, all accessed via Ollama [12].

We manually compared the models' predictions with our labeled dataset, identifying discrepancies such as mismatched condition tags or incorrect sentiment classifications. Any inconsistencies were reviewed manually, and the final labels required consensus from at least two annotators.

While errors from the similarity and sentiment models were minimal, this validation process further reduced inaccuracies, ensuring that any remaining errors had a negligible impact on overall results.

3.4. Dataset distribution

To ensure a rigorous and verifiable evaluation of LLM recommendations, our primary data curation protocol uses a strict filtering method. This approach establishes a ground truth for accuracy evaluation by requiring an app to meet two criteria: it must explicitly target a specific mental health condition in its official description, and it must have positive user reviews verifying its effectiveness. Applying this strict criterion resulted in our final dataset of 55 applications. As shown in Figure 3, the final counts for the conditions formally evaluated in this study were: anxiety (44 apps), depression (27 apps), ADHD (6 apps), and insomnia (5 apps).

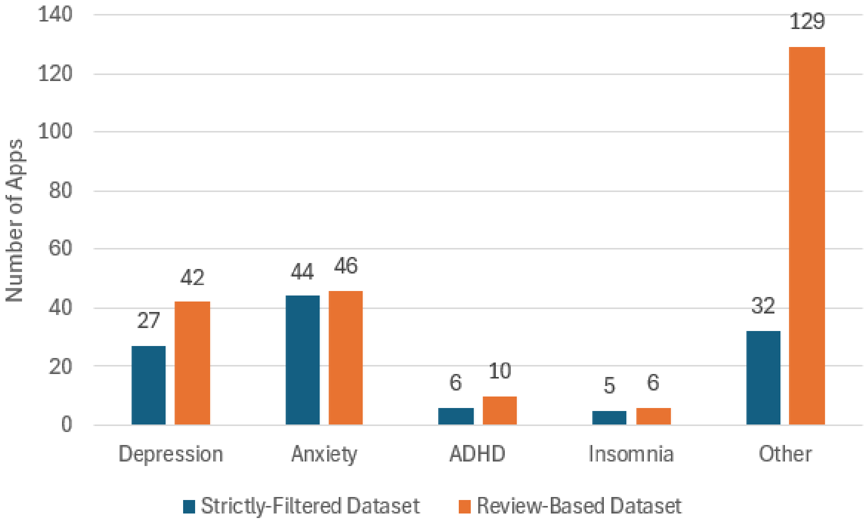

Figure 3. Distribution of evaluated labels in the final, strictly-filtered dataset and the review-based dataset with N = 5. The 'other' category includes all other labels, such as PTSD, mindfulness, mood tracker, etc. Most apps have multiple labels.

While this strict method is essential for our analysis, an alternative, more lenient filtering method can be used to surface a broader set of potentially helpful apps, particularly for underrepresented conditions. This alternative approach identifies apps based on having at least N positive user reviews that mention a specific condition. For example, using a threshold of N = 5, this review-based method identified a larger pool of applications that users anecdotally found helpful, with some of the most frequent labels being anxiety (46 apps), depression (42 apps), and PTSD (16 apps).

The impact of the chosen method is stark; for instance, while the lenient approach with N=5 identified 16 apps for PTSD, our strict protocol verified only one. This highlights a key trade-off: the lenient method expands discovery but presents a challenge for objective evaluation. A recommendation cannot be systematically verified as "correct" if an app's primary purpose differs from what is mentioned in reviews (e.g., a timer app being praised for “stress” management). Our strict process was therefore essential for ensuring that the LLM evaluation was based on recommending apps demonstrably and intentionally designed for the user’s specified condition.

Despite our formal accuracy evaluation focusing on the strictly-filtered dataset, our RAG application retains the flexibility to query the more lenient dataset. This allows users to find resources for a wider variety of conditions not explicitly evaluated here, and these recommendations still benefit from the high-quality data curation, including sentiment analysis and similarity metrics, that our RAG approach provides.

3.5. User query processing via RAG

3.5.1. Dataset structuring

Apps were stored with the following data elements and formatted in JSON (JavaScript object notation), a standard text-based format for representing structured data:

name:	[app name]
developer:	[app developer]
rating:	[N out of 5 stars]
labels:	[condition1, condition2, ...]
reviews:	[review1, review2, ...]
sentiment:	[pos, neg, neu]

3.5.2. Retrieval-augmented response generation

We implemented an RAG pipeline using LangChain [13] to improve LLM-based recommendations. The RAG-enhanced models used were Gemma 2, accessed via Ollama, and GPT-3.5-Turbo, GPT-4o, and Claude 4 Sonnet, accessed via their respective APIs (application programming interfaces) [14], which are protocols that allow different software systems to communicate with each other.

Figure 4 shows our RAG pipeline, which consisted of several integrated components. First, we filtered the dataset based on the mental health condition identified in the user query. For example, when a user asked about apps for depression, only apps labeled with depression were included in the subsequent processing steps. This initial filtering significantly reduced computational costs and improved recommendation relevance.

The filtered dataset was then converted into a single string and passed to the respective embedding functions for our models. These embeddings were then indexed in the FAISS vector database for efficient similarity-based searches.

The retrieved context was passed to an LLM via a conversational retrieval chain, which combined the retrieved context with the user’s query to generate recommendations.

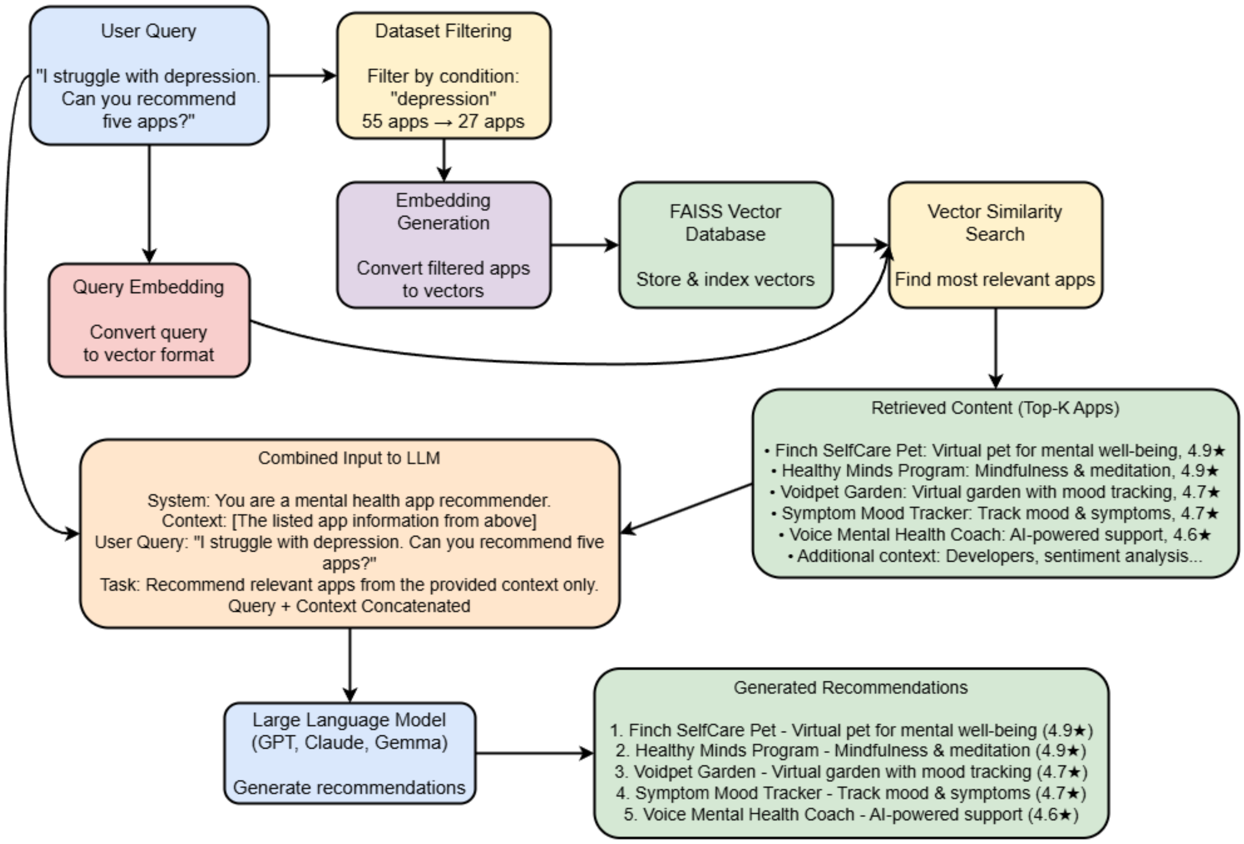
User Query
"I struggle with depression. Can you recommend five apps?"
Dataset Filtering
Filter by condition: "depression"
55 apps → 27 apps
Embedding Generation
Convert filtered apps to vectors
FAISS Vector Database
Store & index vectors
Vector Similarity Search
Find most relevant apps
Query Embedding
Convert query to vector format
Retrieved Content (Top-K Apps)
• Finch SelfCare Pet: Virtual pet for mental well-being, 4.9★
• Healthy Minds Program: Mindfulness & meditation, 4.9★
• Voidpet Garden: Virtual garden with mood tracking, 4.7★
• Symptom Mood Tracker: Track mood & symptoms, 4.7★
• Voice Mental Health Coach: AI-powered support, 4.6★
• Additional context: Developers, sentiment analysis...
Combined Input to LLM
System: You are a mental health app recommender.
Context: [The listed app information from above]
User Query: "I struggle with depression. Can you recommend five apps?"
Task: Recommend relevant apps from the provided context only.
Query + Context Concatenated
Large Language Model (GPT, Claude, Gemma)
Generate recommendations
Generated Recommendations
1. Finch SelfCare Pet - Virtual pet for mental well-being (4.9★)
2. Healthy Minds Program - Mindfulness & meditation (4.9★)
3. Voidpet Garden - Virtual garden with mood tracking (4.7★)
4. Symptom Mood Tracker - Track mood & symptoms (4.7★)
5. Voice Mental Health Coach - AI-powered support (4.6★)

Figure 4. High-level representation of our RAG system.

3.6. Recommendation quality evaluation

3.6.1. Evaluation Protocol

We tested four models, evaluating each in two configurations: a baseline version and a RAG-enhanced version. The models were Gemma 2, GPT-3.5-Turbo, GPT-4o, and Claude 4 Sonnet. Each model was evaluated using four standardized queries, such as: “I struggle with [condition]. Can you recommend five apps that might help?”

Recommendations were manually validated against two criteria: Explicit Targeting, where the app’s description must mention the condition, and Verified Effectiveness, where user reviews must confirm that the app is effective for that condition.

3.6.2. Evaluation metrics

To evaluate the performance of mental health app recommendations by different LLMs, we utilized three metrics tailored to the domain-specific requirements of mental health resources. Unlike conventional recommendation systems that prioritize preference matching or engagement, mental health app recommendations must balance clinical relevance, recommendation variety, and user satisfaction.

Accuracy measures whether the recommended app explicitly targets the requested condition and has verified effectiveness based on user reviews. This metric is crucial in mental health contexts where recommending inappropriate apps could be harmful.

Variety evaluates how diverse the recommendations are across queries for different conditions, which is important to provide users with meaningful alternatives that address different aspects of their condition or offer various therapeutic approaches.

We also use this metric to determine if a model recommends the same incorrect app for multiple different conditions. Quality assesses how well-received the recommendations are by actual users, recognizing that mental health apps with higher ratings (above 4.5 stars) are more likely to be effective, user-friendly, and supportive.

We define the following variables:

- N_c: Number of correct recommendations.
- N_u: Number of unique recommendations (apps recommended only once across all queries for different conditions).
- $N_{4.5+}$: Number of unique and correct recommendations with a user rating above 4.5 stars.

Accuracy is computed as the ratio of correctly recommended apps (N_c) out of 20 total recommendations.

$$\text{Accuracy} = \left(\frac{N_c}{20}\right) \times 100\%.$$

Variety is measured as the ratio of unique app recommendations (N_u) out of 20 total recommendations.

$$\text{Variety} = \left(\frac{N_u}{20}\right) \times 100\%.$$

Lastly, quality is represented by the ratio of unique and correct app recommendations with a rating above 4.5 stars ($N_{4.5+}$) out of the total correct recommendations (N_c):

$$\text{Quality} = \left(\frac{N_{4.5+}}{N_c}\right) \times 100\%.$$

4. Results

4.1. Recommendation accuracy

Figure 5 shows that the RAG pipeline universally corrects the accuracy deficiencies observed in baseline models. In their baseline state, the models showed varied performance, with Gemma 2 achieving 70%, GPT-3.5-Turbo at 65%, GPT-4o at 75% accuracy, and Claude 4 Sonnet at 60%. Notably, some of the baseline models tended to struggle with certain conditions, such as GPT-3.5-Turbo and Claude 4 Sonnet with ADHD recommendations and Claude 4 Sonnet with depression recommendations.

When the baseline LLMs provided incorrect recommendations, they displayed several distinct behaviors based on their reliance on generalized training data. These inaccuracies stemmed from a few key issues, including a reliance on generic or outdated information, a failure to target specific conditions—where recommendations were deemed incorrect if the app's description did not explicitly mention the condition or lacked verified effectiveness from user reviews—and the repetition of incorrect suggestions, which lowered the overall accuracy for some models. These shortcomings were particularly apparent in the ADHD recommendations from GPT-3.5-Turbo and the depression and ADHD recommendations from Claude 4 Sonnet. In stark contrast, the RAG pipeline corrected these behaviors entirely. The RAG-enhanced versions of all four models achieved 100% accuracy across all tested conditions. This perfect score demonstrates the power of the RAG pipeline to eliminate incorrect recommendations by grounding the models in a verified, domain-specific knowledge base.

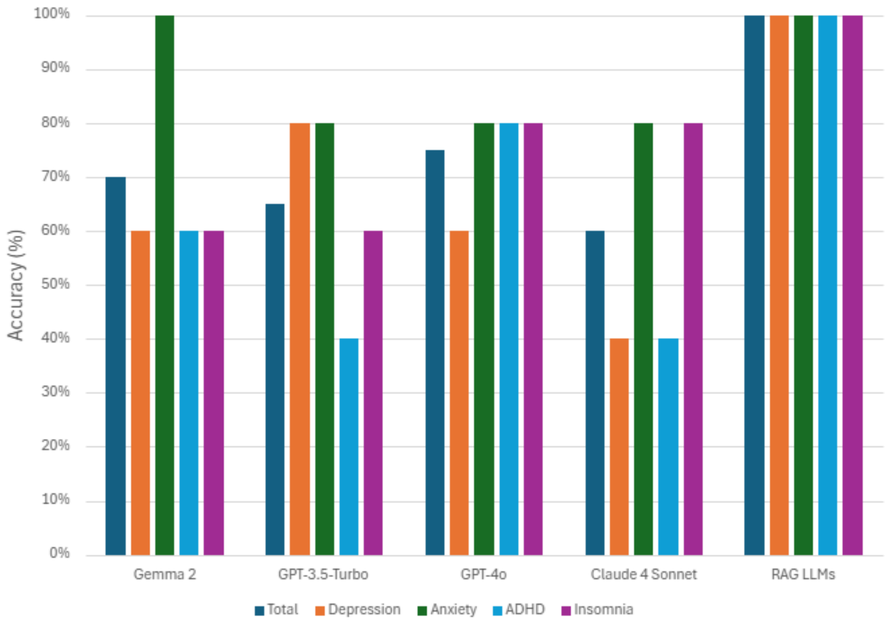

Figure 5. Percentage of accurate recommendations of varying conditions for each LLM.

4.2. Recommendation diversity

As shown in Figure 6, the RAG pipeline maintains strong recommendation diversity while ensuring accuracy. In the baseline comparison, GPT-4o provided the highest variety with 90% unique recommendations, while other models ranged from 60% to 80% . This indicates that some baseline models, like Gemma 2 and Claude 4 Sonnet, tended to repeat suggestions more often than other models.

Depending on the model, repeated suggestions may lower the overall accuracy if the repeated app is incorrect. Claude 4 Sonnet, in particular, had this issue, leading to lower accuracy with ADHD recommendations.

The aggregated "RAG LLMs" bar shows that the RAG-enhanced models collectively produced 80% unique recommendations. This is a crucial finding, as it demonstrates that the structured retrieval process does not force the models into recommending the same few apps. Instead, it strikes an effective balance between providing varied suggestions and guaranteeing their relevance, allowing the models to explore the breadth of the provided dataset to offer users a wide yet entirely accurate set of options.

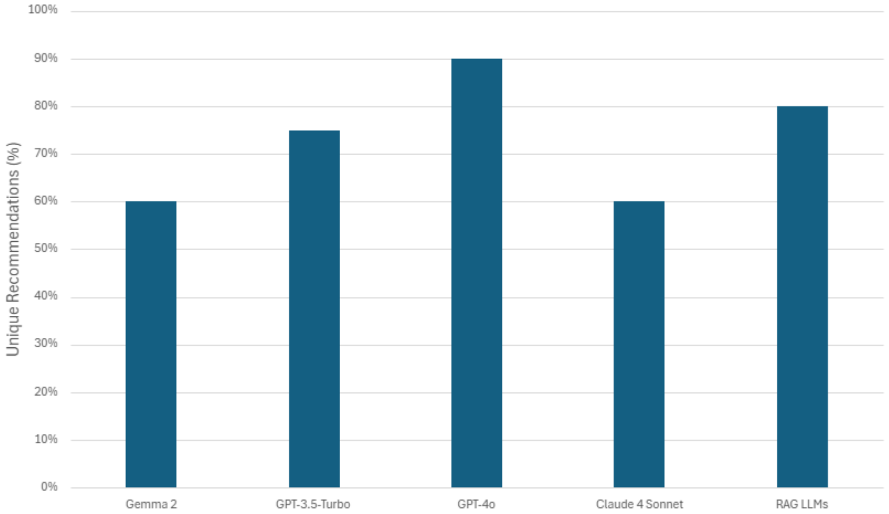

Figure 6. Percentage of unique recommendations for each LLM.

4.3. Recommendation quality

Figure 7 provides the most compelling evidence of the RAG pipeline's value, showing a dramatic improvement in the quality of recommended apps for all models. The chart directly compares the percentage of recommended apps with a user rating above 4.5 stars for each model in its baseline versus its RAG-enhanced state. While baseline models demonstrated low-to-moderate performance in recommending high-quality apps, every model saw a substantial increase with RAG enhancement, achieving comparable, high-quality scores. These findings indicate that the RAG pipeline does more than just improve factual accuracy; it guides the models toward recommending resources that are demonstrably more effective and better-liked by users. By leveraging a dataset enriched with sentiment analysis, the RAG models are better equipped to identify and prioritize apps that provide a positive user experience.

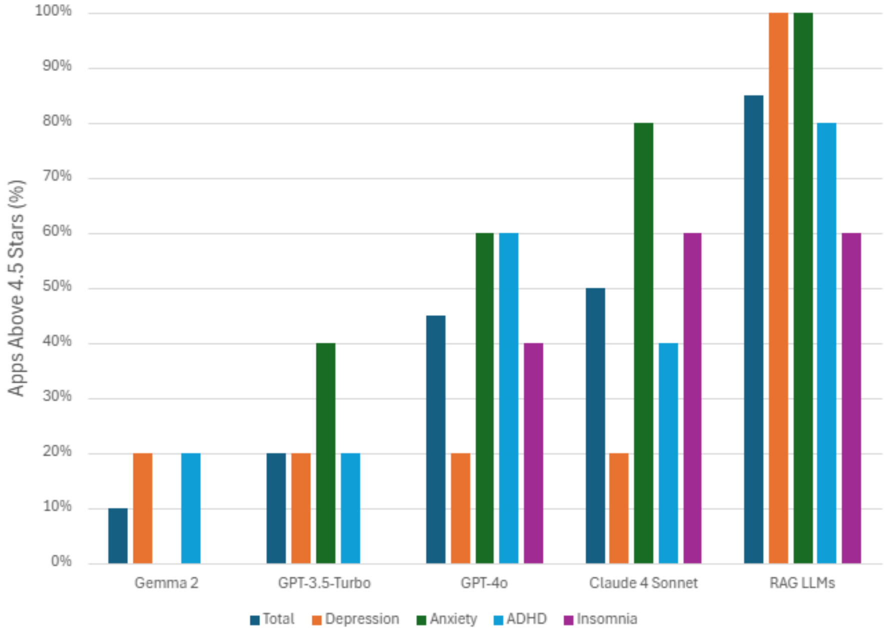

Figure 7. Percentage of recommendations rated above 4.5 stars for each LLM.

5. Limitations

Several factors may have influenced the accuracy of the LLMs' app recommendations.

First, despite utilizing an extensive keyword bank, some relevant terms may have been overlooked, potentially affecting classification accuracy. This limitation pertains to the labeling of both reviews and apps—missing keywords could lead to labels that are less specific or inclusive than intended.

Additionally, the pretrained models used to generate sentiment and similarity scores may carry biases from their original training data. While this could impact the accuracy of sentiment analysis and similarity measurements in mental health contexts, our validation process suggests that any such effects were negligible. Another challenge is the dynamic nature of app content. Developers frequently update app descriptions, and user reviews evolve. These changes could affect the reproducibility of this research, as some apps in our dataset may no longer be valid in future studies if they undergo substantial modifications.

Furthermore, the performance of the RAG pipeline is fundamentally dependent on the quality and comprehensiveness of the underlying dataset and vector database. The most significant limitation of this study was the availability of apps that met our criteria. While depression and anxiety apps were abundant, those for ADHD and insomnia were scarcer. We were unable to include PTSD in our evaluation, as only one app passed our filtering process, falling short of the five required for testing. This highlights a potential gap in the market for high-quality mental health apps targeting niche or underrepresented conditions.

6. Conclusions

This study systematically evaluated the impact of an RAG pipeline on the performance of four distinct large language models for mental health app recommendations. The results unequivocally demonstrate that enhancing LLMs with a structured, domain-specific dataset is a universally effective strategy for improving recommendation accuracy, quality, and diversity. Our head-to-head comparison revealed that while baseline models exhibit inconsistent accuracy, the RAG-enhanced versions of every model achieved 100% accuracy. This performance equivalence is particularly noteworthy given the significant cost differences between the models tested—Gemma 2, GPT-3.5-Turbo, GPT-4o, and Claude 4 Sonnet, in increasing order of cost—and demonstrates that RAG can enable lower-cost models to perform on par with more expensive ones for this task. This finding underscores the critical importance of factual grounding in sensitive domains like mental health. Furthermore, the RAG pipeline dramatically increased the quality of recommendations across all models, guiding them to suggest apps with significantly higher user ratings. This enhancement was universal, elevating every model to a comparable, high-quality score and demonstrating that even the most capable models require factual grounding to perform reliably in specialized domains. The study's key takeaway is that the value of RAG extends beyond elevating lower-cost models; it is an essential component for ensuring reliability and consistency across a wide range of LLMs. By grounding models in a curated knowledge base, we mitigate the risk of outdated or generic responses and transform them into highly reliable, domain-specific experts. Future research should focus on several promising directions. First, expanding the dataset to include a wider range of mental health conditions, particularly underrepresented ones like PTSD, would improve the system's applicability. Second, implementing more sophisticated retrieval algorithms to handle nuanced or implicit requests would enhance the user experience. Finally, conducting user studies to evaluate the real-world impact of these recommendations would provide valuable insights into their practical effectiveness.

These findings emphasize the necessity of tailored, contextually aware LLMs for domain-specific tasks. RAG-enhanced models, whether open-source or proprietary, can serve as accessible, scalable, and highly reliable tools for improving mental health support, ultimately helping to bridge the gap between the need for and the accessibility of quality mental health resources.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Md Abdullah Al Hafiz Khan is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. *World Health Organization, COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*, World Health Organization news release, 2022. Available from: `https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide`.
2. J. Radez, T. Reardon, C. Creswell, F. Orchard, P. Waite, Adolescents' perceived barriers and facilitators to seeking and accessing professional help for anxiety and depressive disorders: a qualitative interview study, *Eur. Child Adolesc. Psychiatry*, **31** (2022), 891–907. https://doi.org/10.1007/s00787-020-01707-0
3. M. Omar, S. Soffer, A. Charney, I. Landi, G. Nadkarni, E. Klang, Applications of large language models in psychiatry: a systematic review, *Front. Psychiatry*, **15** (2024), 1422807. https://doi.org/10.3389/fpsyt.2024.1422807
4. J. Borghouts, E. Eikey, G. Mark, C. De Leon, S. Schueller, M. Schneider, et al., Barriers to and facilitators of user engagement with digital mental health interventions: systematic review, *J. Med. Internet Res.*, **23** (2021), e24387. https://doi.org/10.2196/24387
5. A. Le Glaz, Y. Haralambous, D. Kim-Dufor, P. Lenca, R. Billot, T. Ryan, et al., Machine learning and natural language processing in mental health: systematic review, *J. Med. Internet Res.*, **23** (2021), e15708. https://doi.org/10.2196/15708
6. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Proceedings of the 34th International Conference on Neural Information Processing System*, 2020, 9459–9474.
7. Y. Shi, S. Xu, T. Yang, Z. Liu, T. Liu, Q. Li, et al., Mkrag: medical knowledge retrieval augmented generation for medical question answering, *AMIA Annu Symp Proc.*, **2024** (2025), 1011–1020.

8. E. Leivada, G. Marcus, F. Günther, E. Murphy, A sentence is worth a thousand pictures: can large language models understand human language and the world behind words? arXiv: 2308.00109. https://doi.org/10.48550/arXiv.2308.00109

9. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P. Mazaré, et al., The faiss library, arXiv: 2401.08281. https://doi.org/10.48550/arXiv.2401.08281

10. *Hugging Face, All-roberta-large-v1*, Sentence-Transformers, 2025. Available from: `https://huggingface.co/sentence-transformers/all-roberta-large-v1`.

11. *Hugging Face, Twitter-roberta-base for sentiment analysis*, Cardiff NLP, 2020. Available from: `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment`.

12. *GitHub, Get up and running with large language models*, Ollama, 2025. Available from: `https://ollama.com/`.

13. *GitHub, Langchain*, LangChain, 2025. Available from: `https://github.com/hwchase17/langchain`.

14. *OpenAI platform, OpenAI API*, OpenAI, 2025. Available from: `https://platform.openai.com/docs/api-reference`.

AIMS Press