RAG-Based Privacy Policy Analysis For Mental Health Apps

Muhan Ding, Erchen Qu, Jie Xu, Ziyu Kang, Advisor: Dr. Yanxia Jia
Department of Computer Science and Mathematics, Arcadia University

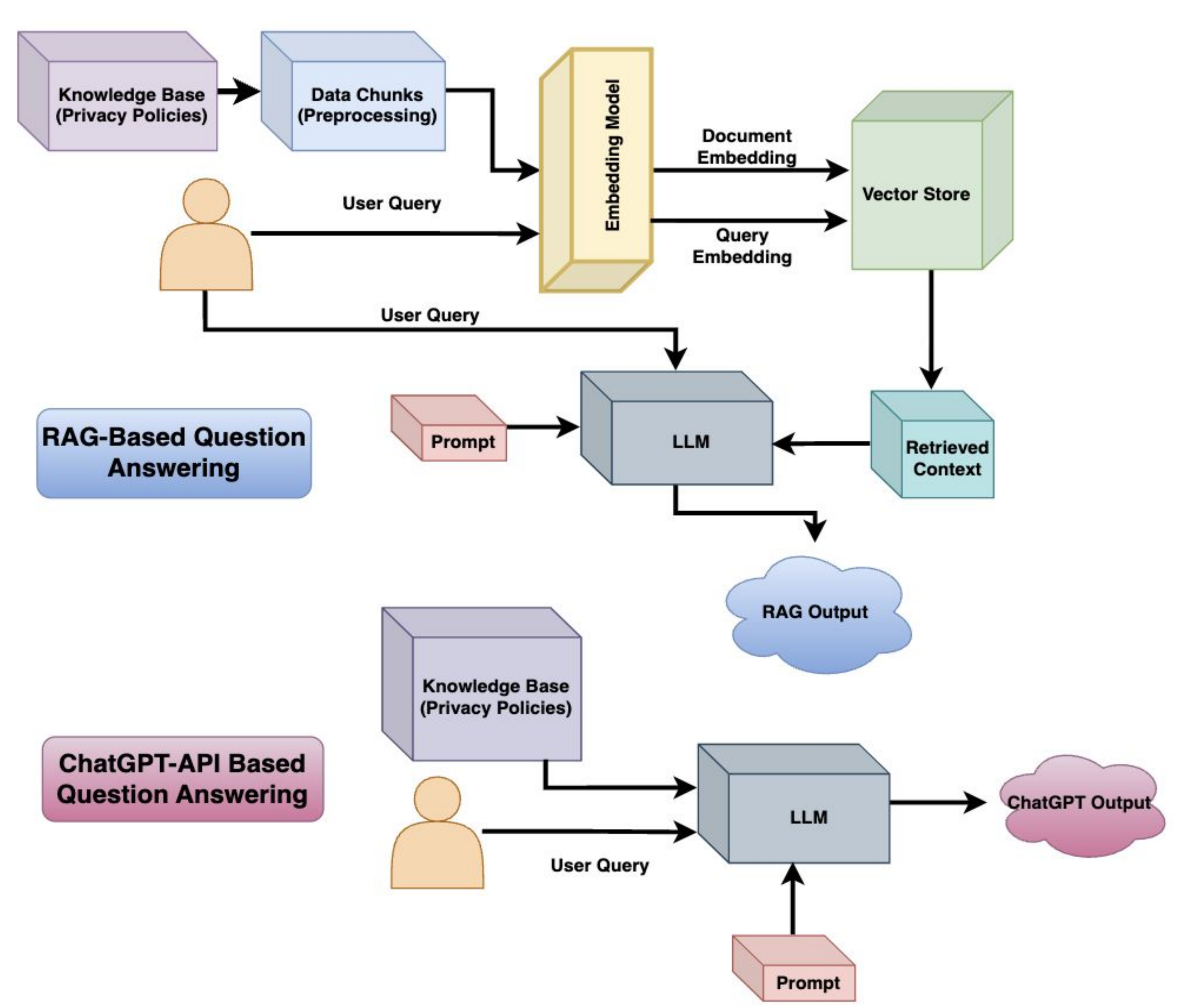
Introduction

Privacy policies explain how apps handle user data, which is crucial for mental health apps due to sensitive information. These policies are often long and complex, making manual analysis difficult. We developed a Retrieval-Augmented Generation (RAG)-based [1] Question Answering system using Haystack [2] to automate privacy policy analyze. Our framework evaluates context relevance, faithfulness, and semantic answer similarity [2].

Research Questions

1. How does the Retrieval-Augmented Generation (RAG) system perform compared to a standalone LLM (e.g., ChatGPT) in analyzing mental health app privacy policies?
2. What patterns and trends can be observed in mobile apps' transparency regarding data collection and sharing practices, and in the extent to which users are able to opt out or delete their data?

Methodology



For RQ1, we compare the performance of our RAG-based method and the standalone GPT-based method by evaluating precision, recall and F1 scores, as well as Context Relevance, Faithfulness and Semantic Answer Similarity (SAS).
The Privacy-related questions are developed based on the the app evaluation model [3] proposed by the American Psychiatric Association (APA). We use the dataset from [4] and manually annotated 233 documents as ground-truth references.

Privacy-related questions (User Query) ;

1. Does the app declare the collection of data? (Y/N)
2. If the app collect user data, what type of data does it collect? (Open-Ended)
3. Does the app declare the purpose of data collection and use? (Y/N)
4. Can you opt out of data collection or delete data? (Y/N)
5. Does the app share data with third parties? (Y/N)
6. If the app shares data with third parties, what third parties does the app share data with? (Open-Ended)

Performance metrics

Table 1: Per-Question Classification Metrix (Y/N questions)

Question	Confusion Matrix	Precision	Recall	F1	Accuracy
Q1(RAG)	[[1,1],[6,196]]	0.995	0.97	0.983	0.966
Q1(GPT)	[[1,1],[5,205]]	0.995	0.976	0.986	0.972
Q3(RAG)	[[1,1],[4,199]]	0.995	0.98	0.988	0.976
Q3(GPT)	[[0,2],[10,200]]	0.99	0.952	0.971	0.943
Q4(RAG)	[[3,19],[6,177]]	0.903	0.967	0.934	0.878
Q4(GPT)	[[4,17],[9,182]]	0.915	0.953	0.933	0.877
Q5(RAG)	[[13,3],[28,161]]	0.982	0.852	0.912	0.849
Q5(GPT)	[[14,2],[46,150]]	0.987	0.765	0.862	0.774
Avg(RAG)	-	0.969	0.942	0.954	0.917
Avg(GPT)	-	0.972	0.911	0.938	0.891

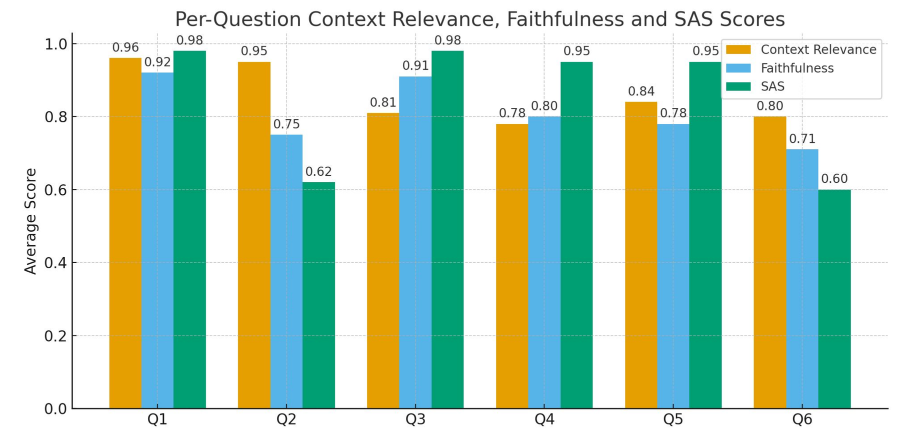


Figure 1: RAG Performance Measurement: Context Relevance, Faithfulness and SAS

GPT Answers(6 Questions)

-	Avg.
SAS	0.8339

RAG Answers(6 Questions)

-	Avg.
Context Relevance	0.86
Faithfulness	0.81
SAS	0.85

Figure 2: SAS Comparison between ChatGPT Baseline and RAG

- Precision = TP / (TP + FP);
- F1 = harmonic mean of precision and recall.
- Recall = TP / (TP + FN);
- Accuracy = (TP + TN) / Total.
- Context Relevance: relevance of the retrieved context to the query
- Faithfulness: the faithfulness of the generated answer to retrieved context
- SAS : semantic alignment between generated answers and ground truth

Results

1. Classification Performance Comparison for Y/N questions

According to Table 1, RAG shows overall improvements compared to the ChatGPT baseline. Specifically, accuracy increased from .89 to .92, recall from 0.91 to 0.94, and F1 from 0.94 to 0.96, indicating better alignment with ground truth.
For more challenging, such as Q5, RAG achieved a notable gain, improving accuracy from .77(ChatGPT) to .85.

2. RAG Performance in Context Relevance, Faithfulness and SAS

As shown in Figure 1, the RAG model demonstrate promising performance results across context relevance, faithfulness and SAS. For Q5, the context relevance reaches .80, while the faithfulness and SAS are both below 80%, indicating room for improvement.
Figure 2 indicates that while ChatGPT baseline and RAG demonstrates comparable SAS performance, RAG shows a slight advantage.

3. App Privacy Statistics: Data Collection and Sharing

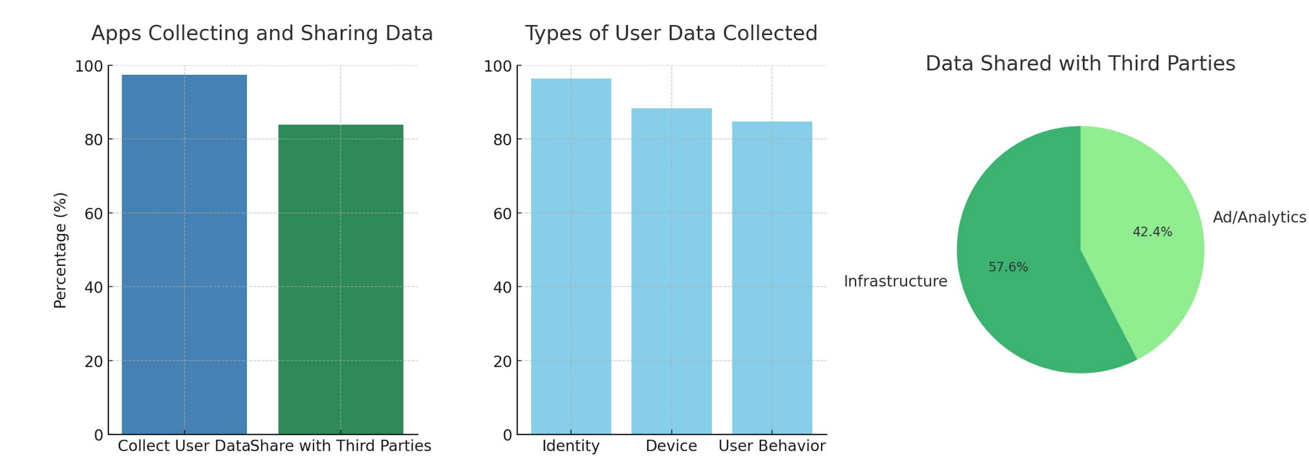


Figure 3: statistics from Ground Truth

Conclusion

RAG slightly outperforms ChatGPT on classification tasks, especially for more challenging questions. It demonstrates promising performance across context relevance, faithfulness and semantic answer similarity.

References

[1] Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.
[2] https://haystack.deepset.ai/
[3] You, X., Wang, W., Shen, Z., and Jia, Y. (2025, June). From Data Trends to Privacy Insights in Mental Health Apps: an LLM-Powered Approach. In 2025 ASEE Annual Conference & Exposition.
[4] Rodriguez, D., Yang, I., Sadeh, N., & Del Alamo, J. M. (2024, May). *Large language models: A new approach for privacy policy analysis at scale. arXiv*. https://arxiv.org/abs/2405.07437