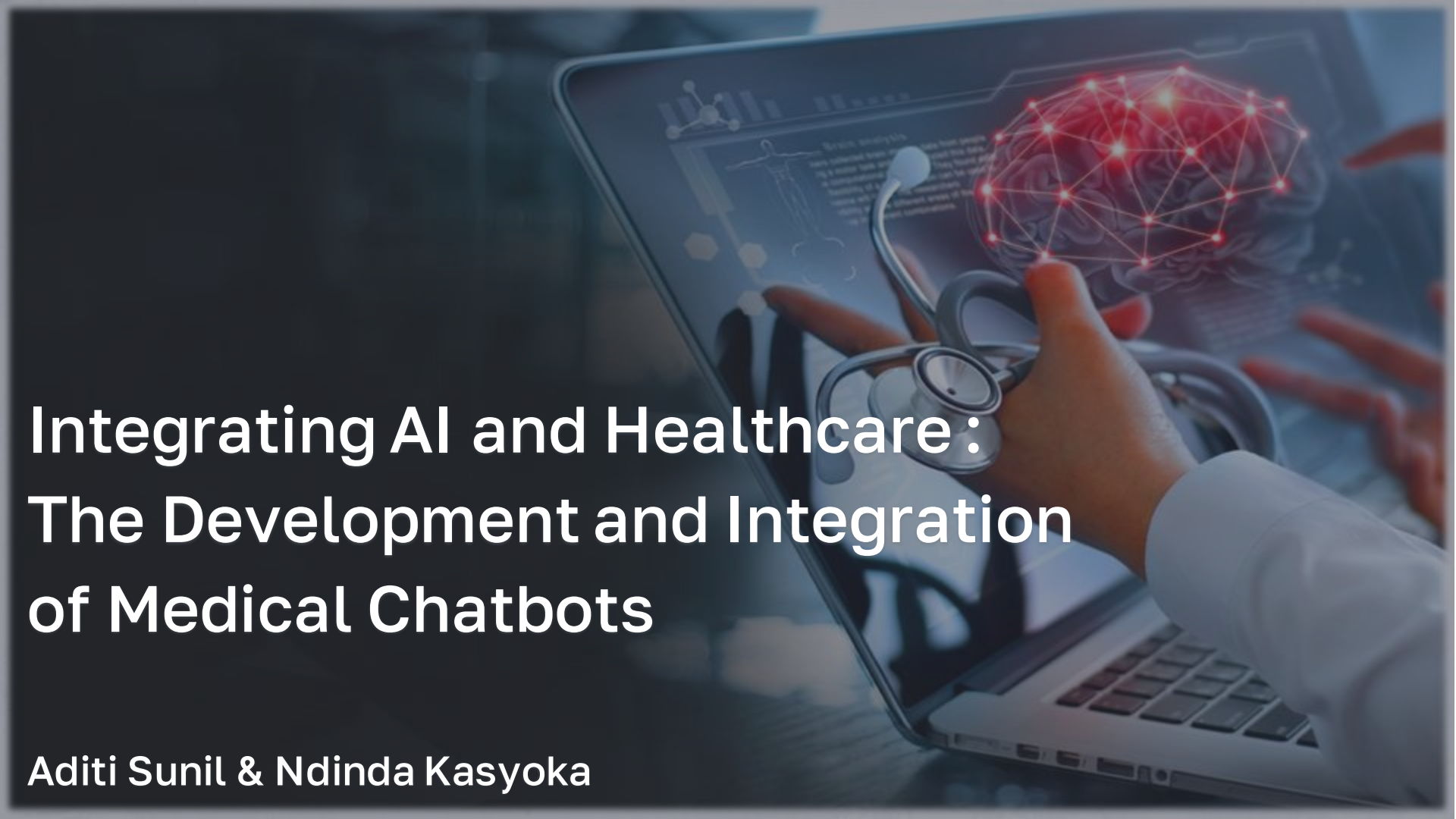


Integrating AI and Healthcare : The Development and Integration of Medical Chatbots

Aditi Sunil & Ndinda Kasyoka

Introduction

In the evolving landscape of healthcare, artificial intelligence stands at the forefront, promising to redefine patient care through innovative solutions.

Areas where AI has had an impact

Radiology
Image Processing used to accurately analyze a big chunk of x-rays

Public Health & Epidemiology
Used to analyze data from various sources to decrease reduce outbreak

Research and Development
Used to accurately review literature and identifying drug pattern by predicting molecular behavior

Effective Clinical Decisions
Used to analyze data from various sources to decrease reduce outbreak

Personalized Healthcare
By analyzing genetic information alongside medical records, LLMs can help tailor medical treatments to individual patients

Healthcare Chatbots
AI models power conversational agents that can provide basic health advice and information, thus enhancing patient engagement and service efficiency

Research Focus

Central Question :
How can the use of different Large Language Models (LLMs) enhance the accuracy and effectiveness of medical chatbots in providing information on specific healthcare topics?

Purpose Statement:
This presentation explores the development and integration of medical chatbots, focusing on different Large Language Models to enhance their accuracy and effectiveness.

Significance
This research can lead to integration of more precise and helpful medical chatbots which can be used to provide precise and accessible healthcare information.

Project Background

Understanding AI Large Language Models (LLMs):
Investigating and comprehending various AI LLMs

Integration for Improved Patient Understanding:
To create medical chatbots to aid patients with understanding their medical conditions, treatment options, and diagnostic procedures.

Method Pt.1

Why did we choose Kidney Cancer?

- Significant amount of kidney cancer research has been conducted, resulting in a substantial amount of data being available for analysis
- There exists high-quality kidney cancer patient data from various institutions which we have access to

Method Pt.2

LLM's Used:

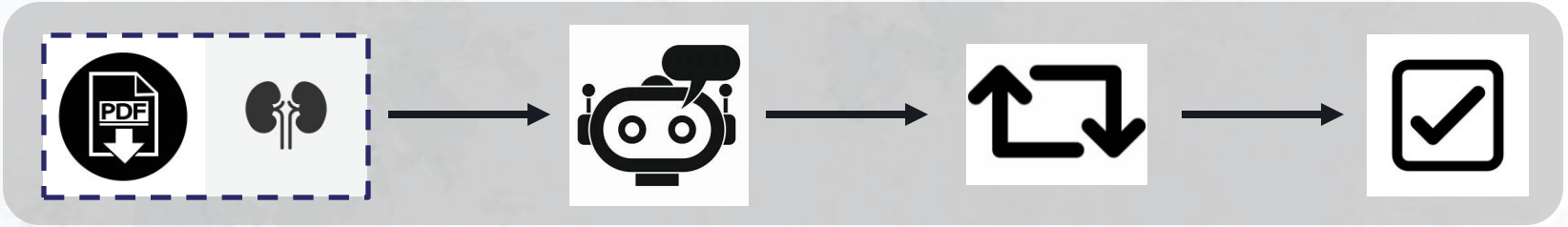
GPT 3.5 turbo

Chat GPT 4.0

Method Pt.3

Retrieval Augmented Generation (RAG) - the practice of enhancing the output of a large language model by incorporating information from an authoritative external knowledge base, beyond the data it was trained on, prior to generating a response

The process



Data Collection

Feeding data into our normal chatbot

Iteration

Checking correctness

Generating Questions & Answers

Manual vs AI Generated:

- Manually created 50 questions related to kidney cancer(procedures, definitions, surgeries, etc.) and read through the pre-selected PDFs to create our own answers
- Used GPT-4 to create 50 questions related to kidney cancer and answers to those questions
- Created a JSON file containing all 100 questions and answers (ground truth) to test the chatbot's responses using the BERT scoring system

Understanding Bert Score

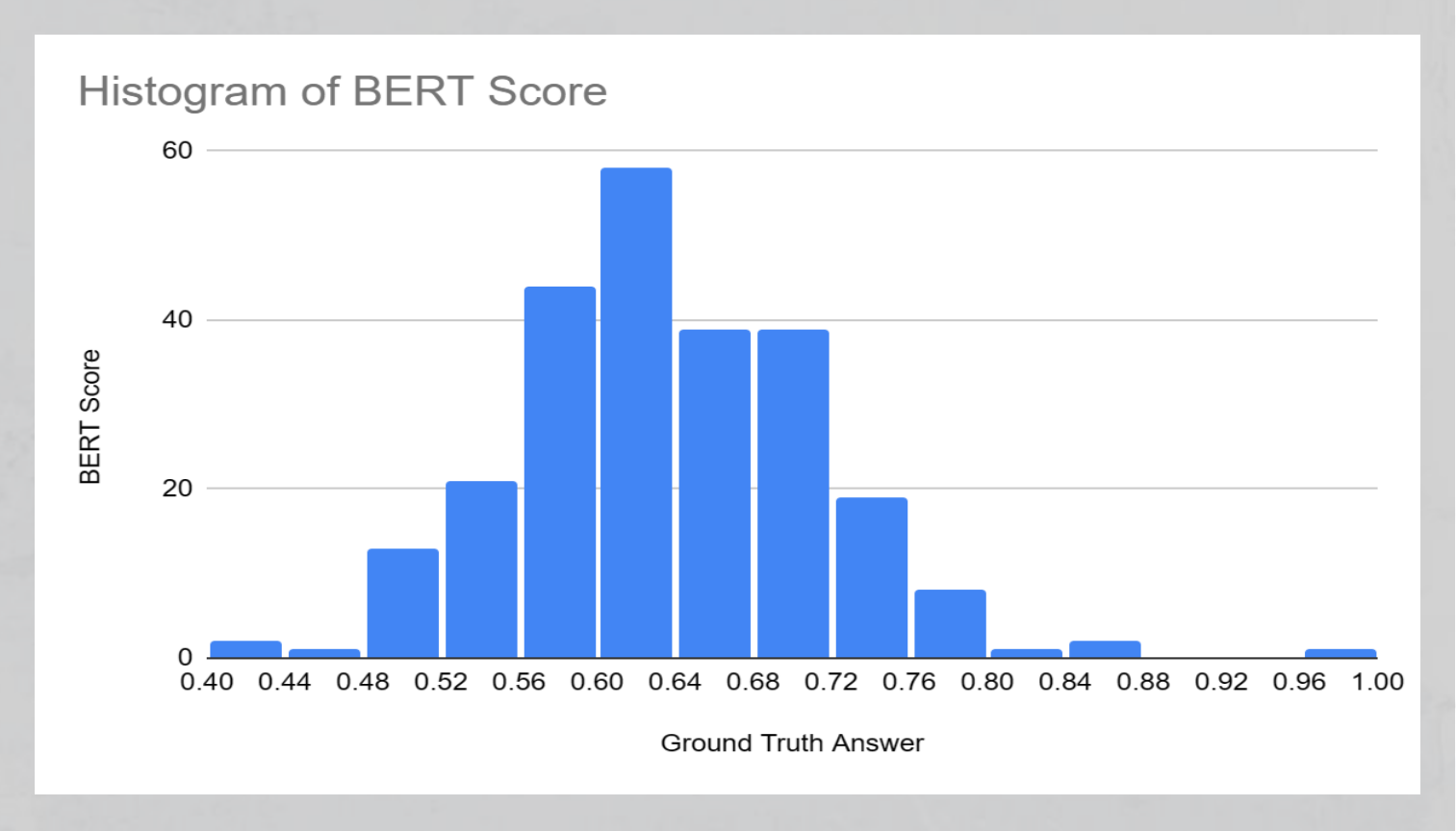
Bert Score is used to accurately check the similarity of the answer the chatbot gives to the correct answer it is expected to give (ground truth answer). The BERT SCORE consists of the **precision, recall and the f1 score.**

Precision – how precise the similarity between the chatbot response **(prediction)** and the ground truth answer is **(reference)**
Recall - measures how well the chatbot response avoids omitting relevant content.
F1 Score - a combination of both Precision and Recall to measure how well the candidate texts capture and retain relevant information from the reference texts.

```
from evaluate import load
bertscore = load("bertscore")
predictions = ["hello world", "general kenobi"]
references = ["goodnight moon", "the sun is shining"]
results = bertscore.compute(predictions=predictions, references=references, model_type="distilbert-base-uncased")
print(results)
{'precision': [0.7380737066268921, 0.5584042072296143], 'recall': [0.7380737066268921, 0.5889028906822205],
```

```
from evaluate import load
bertscore = load("bertscore")
predictions = ["hello world", "general kenobi"]
references = ["hello world", "general kenobi"]
results = bertscore.compute(predictions=predictions, references=references, model_type="distilbert-base-uncased")
print(results)
{'precision': [1.0, 1.0], 'recall': [1.0, 1.0]
```

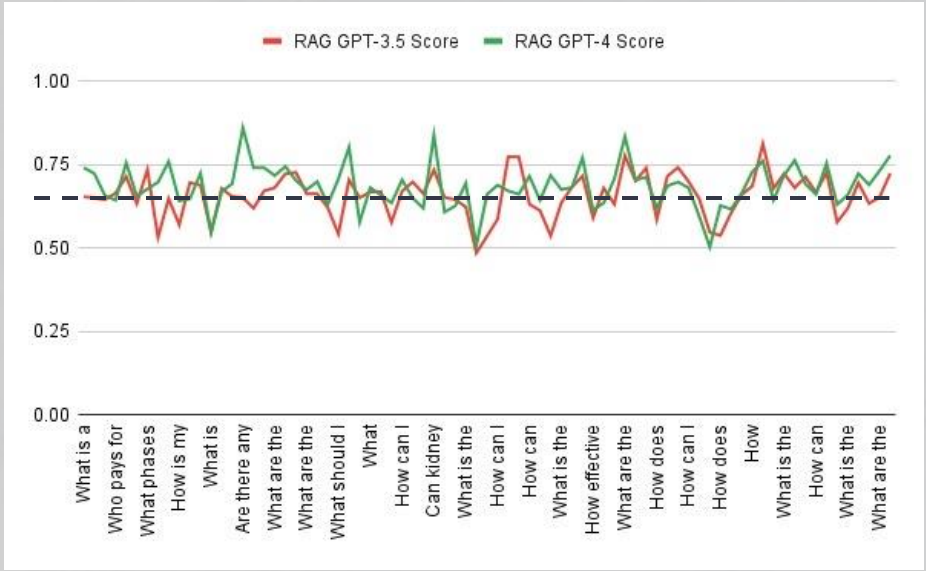
Distribution of BERT Scores for Chatbot Responses Compared to Ground Truth



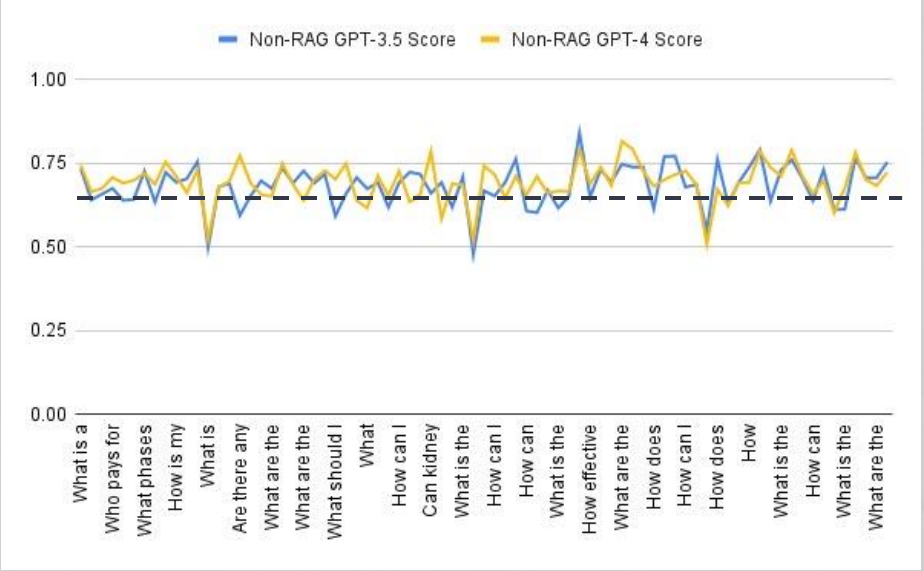
Comparative Analysis of RAG and Non-RAG BERT Scores Across GPT-3.5 and GPT-4 Responses

Comparative Performance of RAG and Non-RAG Responses in GPT Models Based on BERT Scores

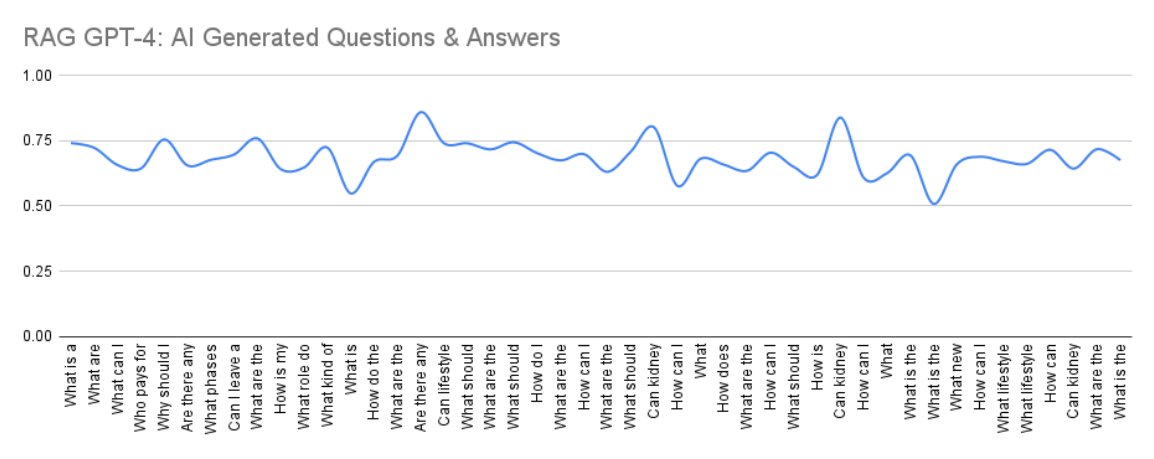
RAG

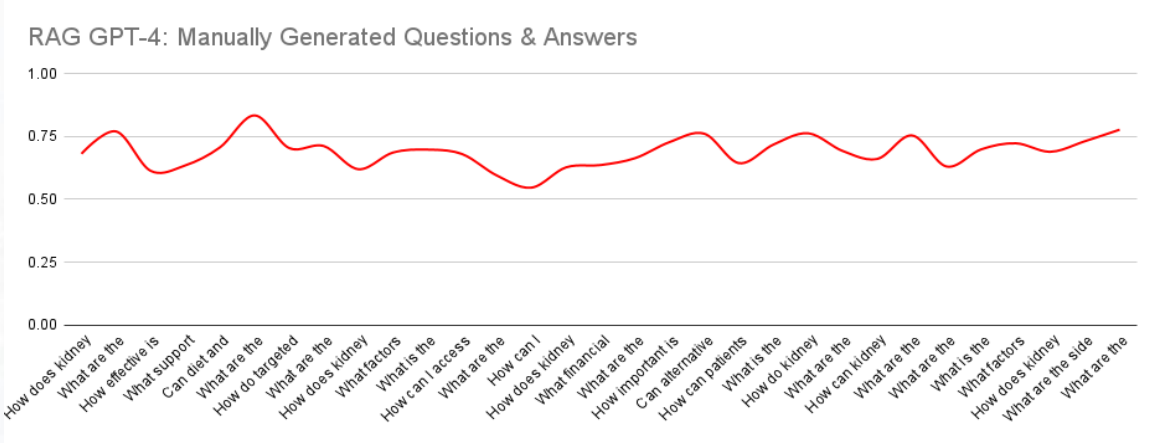


Non-RAG



Accuracy Measures: AI vs Manually Generated Questions & Answers





- Not a large difference between the scores for AI generated questions and ground truth answers vs the manual ones
- By manually checking the responses for accuracy, around 80% gave very accurate responses

Conclusion

Results:

- GPT-4 models enhanced by RAG achieved the greatest similarity with our ground truth answers
- Both RAG and non-RAG of GPT-4 exhibited higher BERT scores than GPT-3.5

Learning Outcomes:

- The difference in BERT scores between RAG and non-RAG models was minimal, suggesting that RAG had little impact on the overall effectiveness of the chatbot responses

Next Steps:

- Further research into optimizing the use of RAG
- Testing with other LLMs
- Expanding the scope of diseases covered by the chatbot

Acknowledgements

- Dr. Alan B McMillan

- URS Fellows:
    - Maria & Mason
    - Ian & Ash