# 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 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 work together towards a common goal.

# Project Background

**Understanding AI LLMs**: Investigating and comprehending various AI LLMs such as ChatGPT, Ollama, Langchain, and Autogen

**Integration for Improved Patient Understanding:** To create a medical chatbot to improve patient understanding of their healthcare

# 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.

- Access to high-quality kidney cancer patient information from various institutions dedicated to researching this disease

Cleveland Clinic
MAYO CLINIC
M ROGEL CANCER CENTER
World Cancer Research Fund
NIH NATIONAL CANCER INSTITUTE
National Kidney Foundation™
Kidney Cancer Association
American Cancer Society®

# Method Pt.2

LLM's Used:

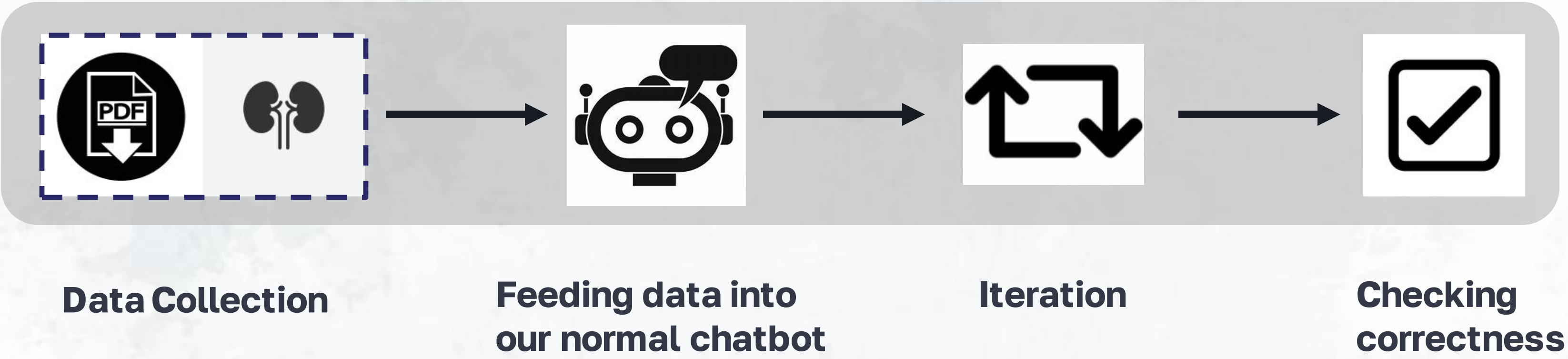
**GPT 3.5 turbo**

**Chat GPT 4.0**

# Method Pt.3

**Retrieval Augmented Generation (RAG) -** the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response.

## The process

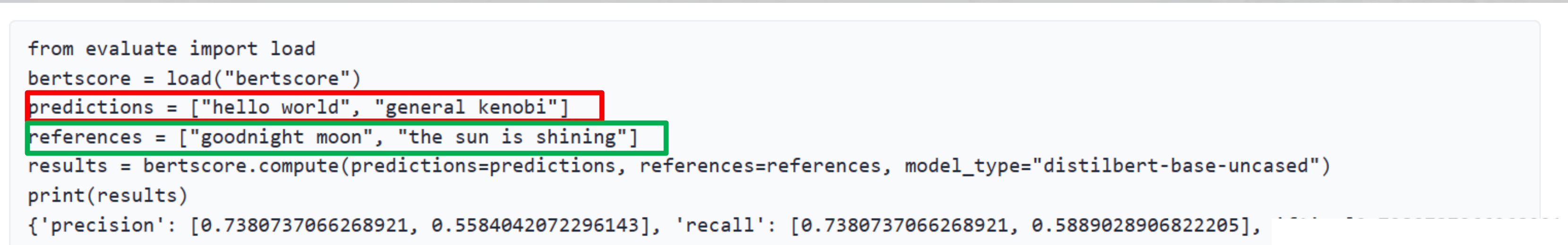


**Data Collection** → **Feeding data into our normal chatbot** → **Iteration** → **Checking correctness**

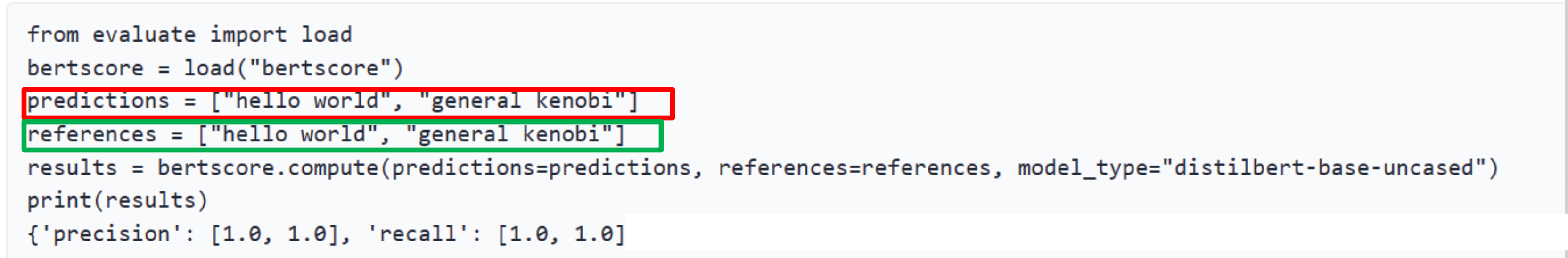
# Understanding Bert Score

**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],
```

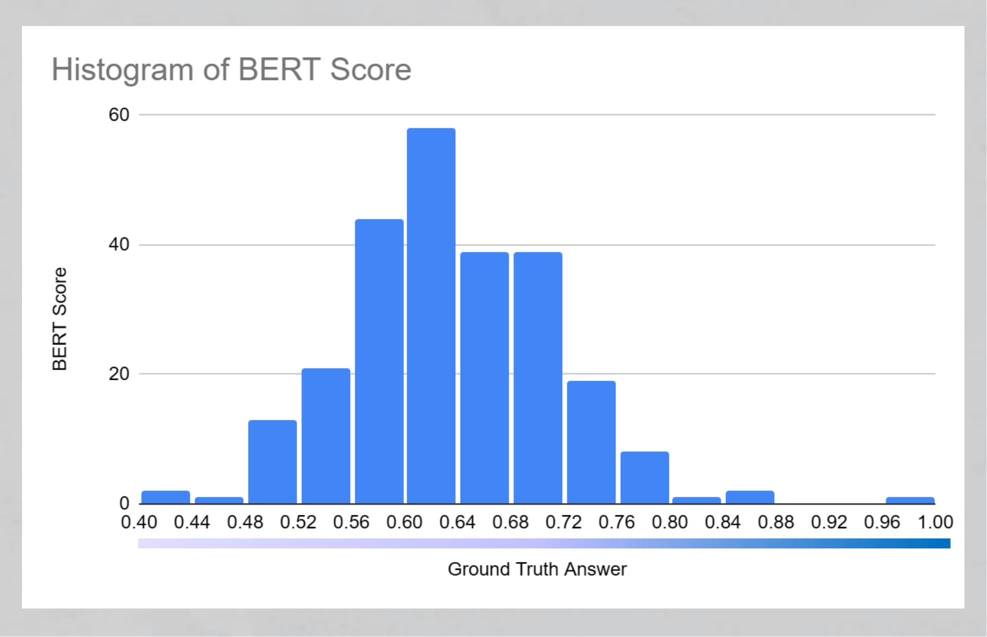


```
from evaluate import load
bertscore = load("bertscore")
predictions = ["hello world", "general kenobi"]
references = ["hello world", "general kenobi"]
results = bertscore.compute(predictions=predictions, references=references, model_type="distilbert-base-uncased")
print(results)
{'precision': [1.0,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

# Analysis

## Getting the Threshold

Below is a graph that shows a distribution of the scores given by the ground truth answers



## The Bert score Distribution with respect to ground truth answers ( Thresh hold ~ 0.62 )

**Very open - ended Questions – (Scores ~ 0.40-0.48)**
Question types were less medical about kidney cancer patients asked to confuse the chatbot eg; How do I cope with the fear of my kidney cancer returning after treatment?

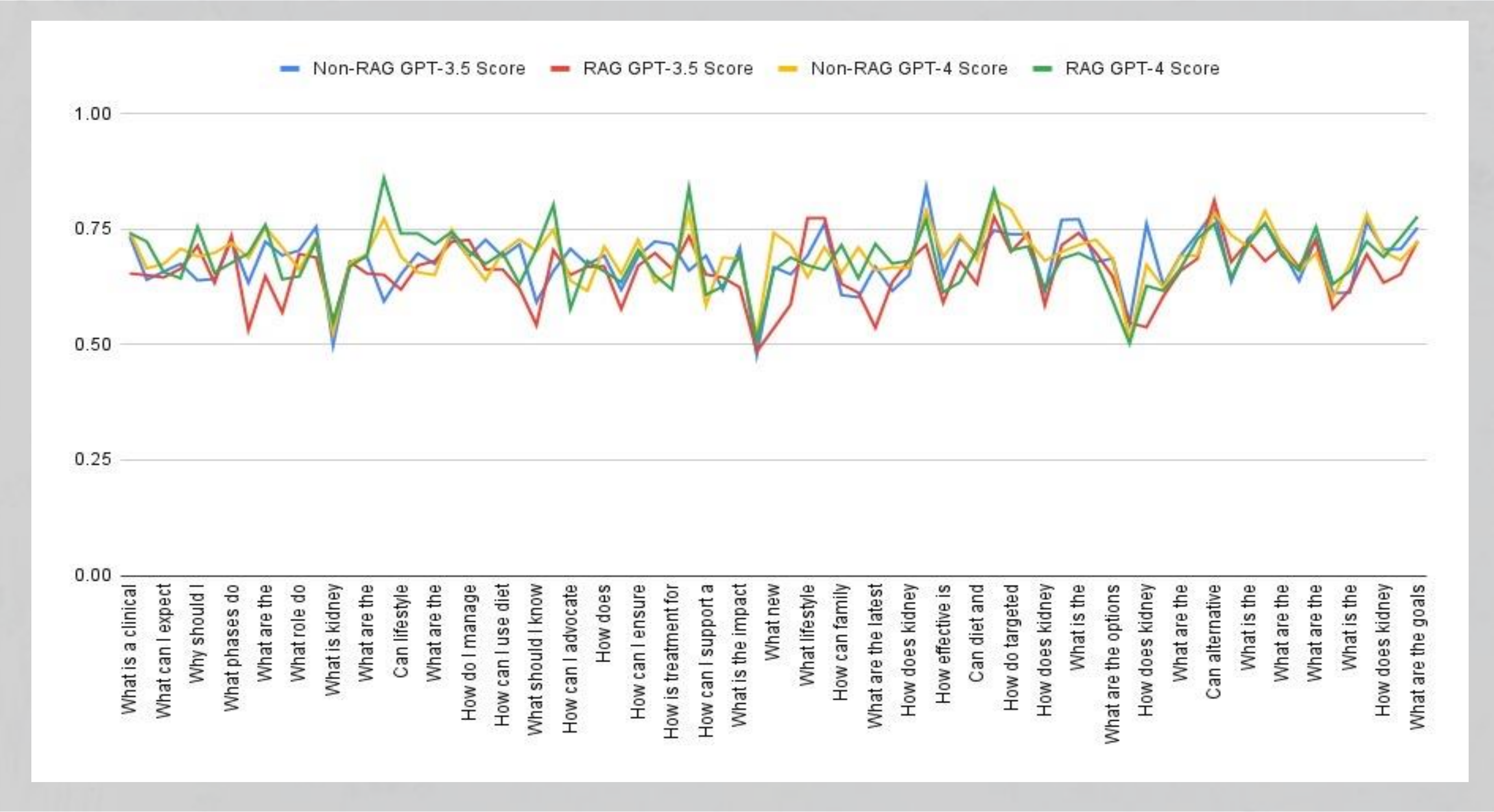
**Hard Question (Scores ~ 0.52- 0.60)**
Question types were more open ended hence predictions and references were less similar. eg; How is kidney function monitored after treatment for kidney cancer?

**Medium Questions (Scores ~ 0.60-0.64)**
Questions types were a balance of open-ended and medical eg; How should I prepare for potential changes in my fertility due to chemotherapy?

**Easy – Medium Questions (Scores ~ 0.68-0.70)**
Question types were procedural/listing questions eg; If the initial chemotherapy regimen doesn't work, what are the next steps?

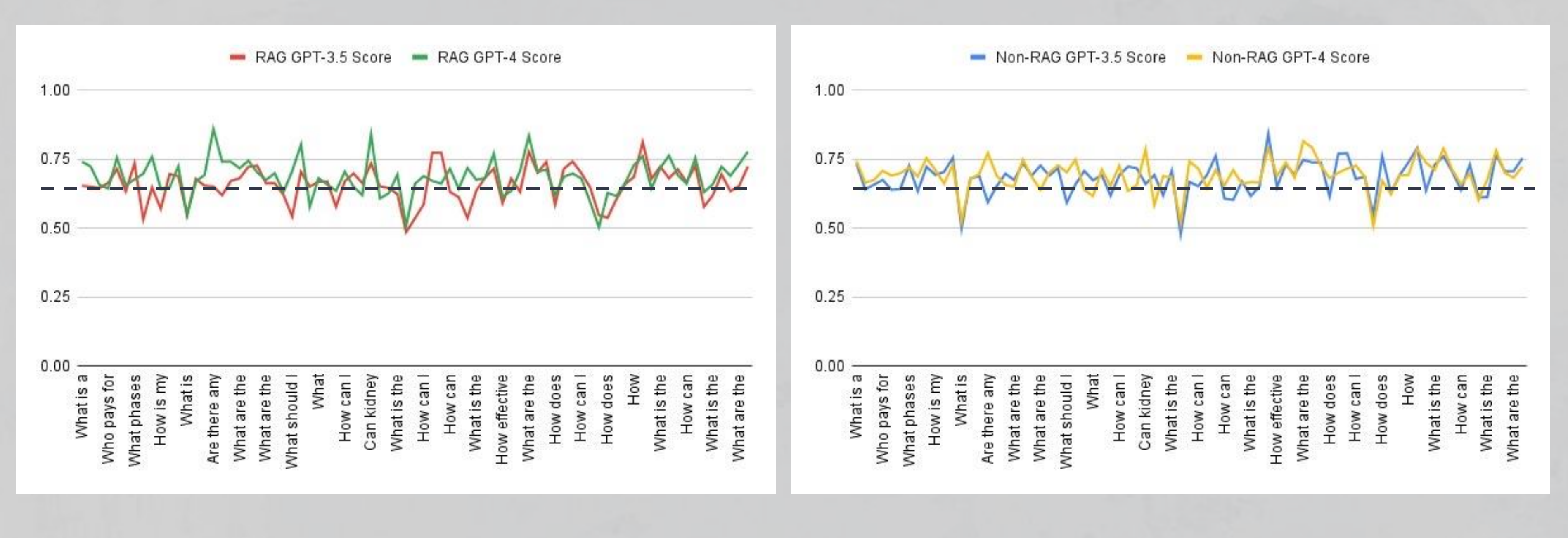
**Easy Questions (definition types) (Scores ~ > 0.70)**
Question types were mostly definition questions.

# Different Large Language Models Used (LLMS)

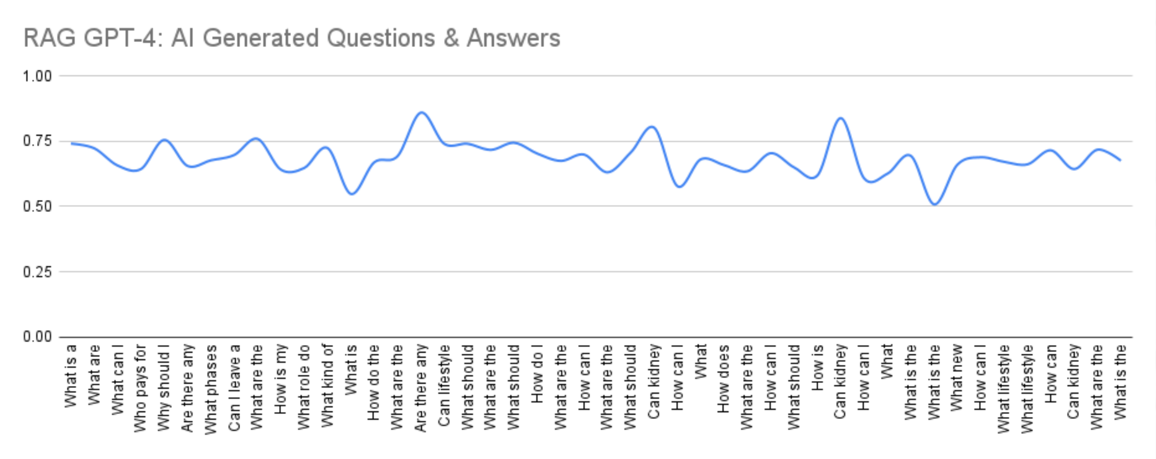


# Results

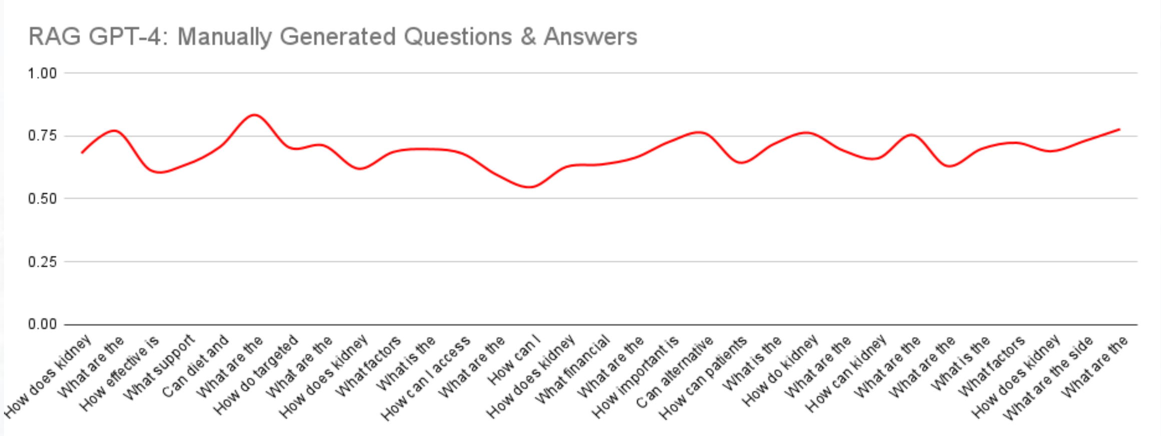
RAG VS NON RAG



# Results



RAG GPT-4: AI Generated Questions & Answers



RAG GPT-4: Manually Generated Questions & Answers

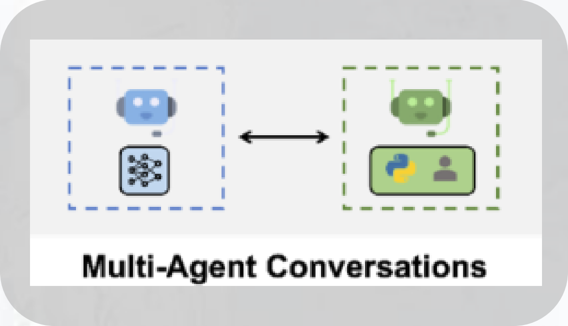
- Not a large difference between the scores for AI generated questions and ground truth answers vs the manual ones

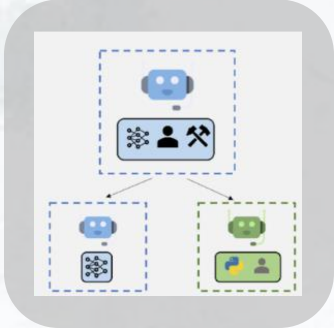
# Chatbot Accuracy & Effectiveness

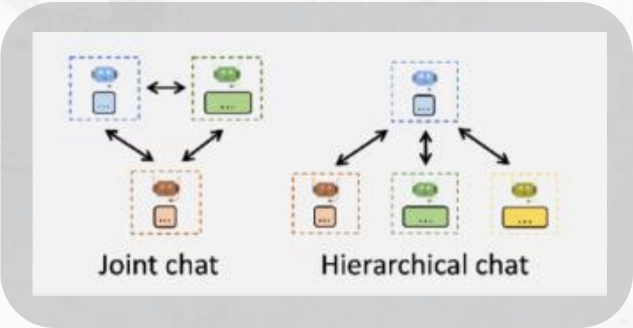
- Overall RAG GPT-4 gave the most similar responses to the ground truth
- GPT-4 (both RAG and non-RAG) had higher scores
- By manually checking the responses for accuracy, we notice that around 80% give very accurate responses

# Future of research

The integration of Large Language Models on Medical Chatbots is to promote specialization of tasks and have chatbots with specific functions work towards a common goal.







# Acknowledgements

- Dr. Alan B McMillan

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