“DonateFoodGoWhere”
A chatbot for donating specific food items in Singapore

By Mei Qi,
DSI 39

In 2022, Singapore generated

813 million kg

of food waste, accounting for

12%

of total waste generated

Annually, each household throws away **$258** worth of food, equivalent of

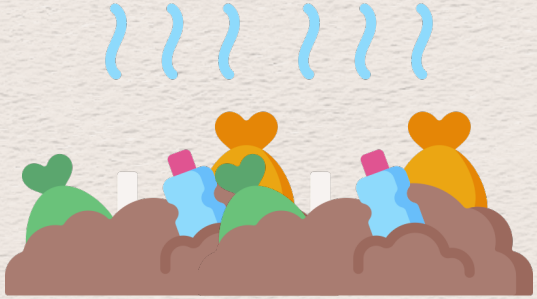
x 52 plates

Assuming each plate of nasi lemak costs $5

Amount of food waste has grown by **30%**
over past 10 years and is expected to rise further

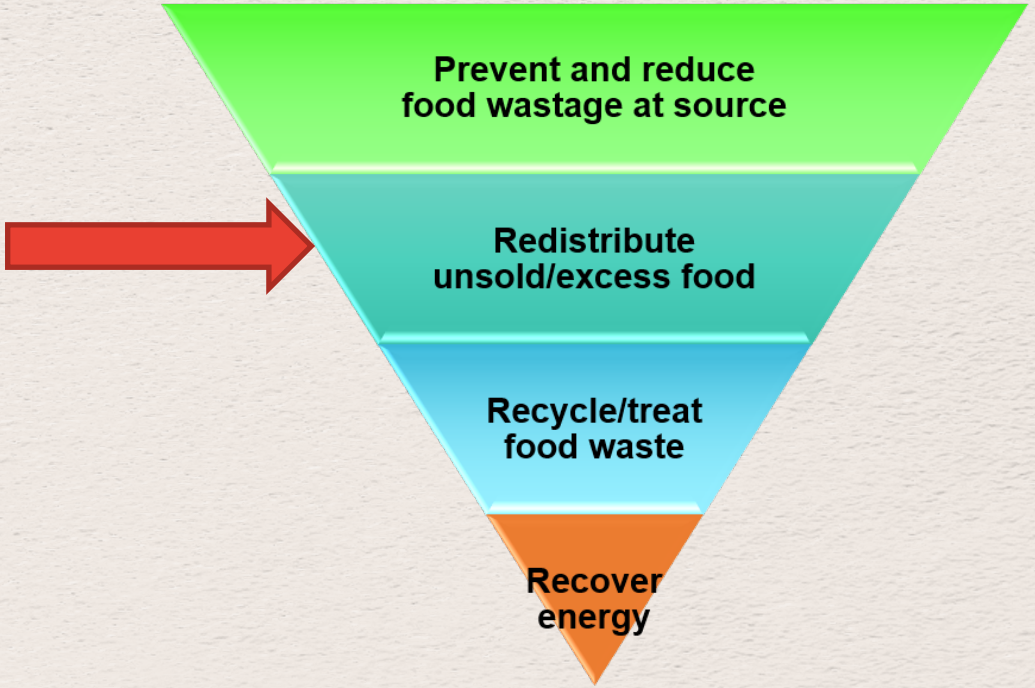
At current waste disposal rates, Singapore will need..

a new incineration plant
every **7-10 years**



a new landfill
every **30-35 years**

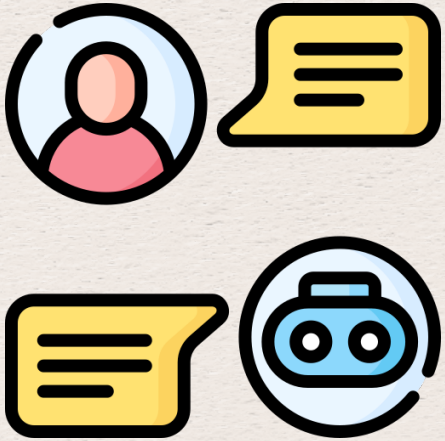
Households contribute around half of the food waste generated



Singapore food waste management strategies,
as part of Singapore' Zero Waste Masterplan
(by NEA)

Organisations have specific wish list of food items and donation requirements, and
it is time-consuming for individuals to find the right organisation and information.

How can we help link individuals up with organisations?



To develop a chatbot for individuals to enquire about donating **<u>specific food items</u>** – where and how to donate with relevant donation instructions

Agenda

Basic concepts of chatbot

Understanding how ChatGPT and Large Language Model (LLM) works

RAG and Fine-tuning

Leveraging the reasoning and generative capabilities of LLM for custom data

Evaluation

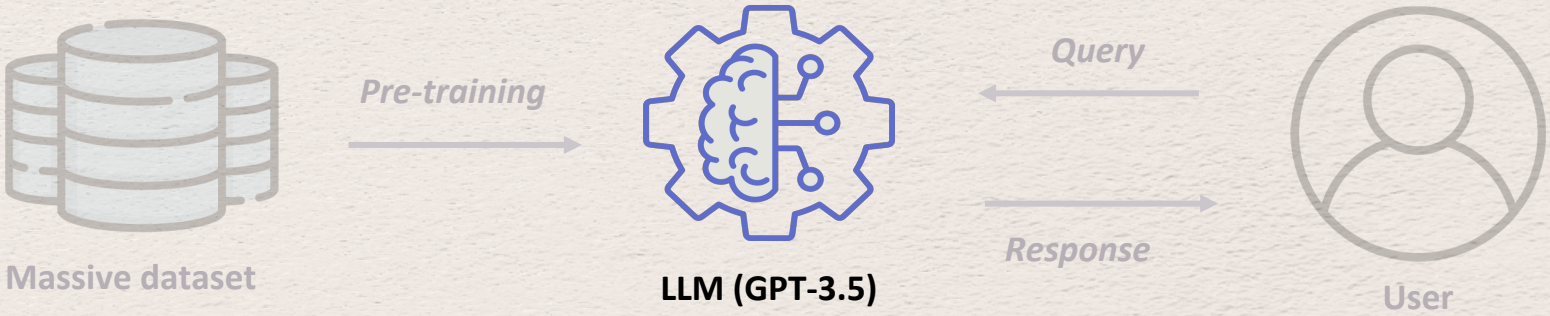
Evaluating the relevancy and accuracy of custom chatbot

Chatbot Design

Designing and deploying the chatbot

Basic concept of a chatbot – OpenAI's ChatGPT

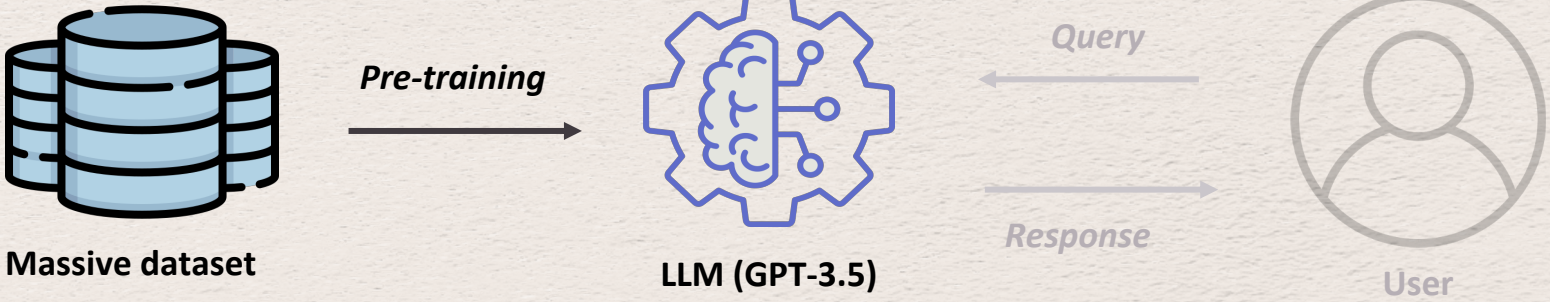
A large language model, LLM, is a deep learning model designed to understand, and generate human language.



Basic concept of a chatbot – OpenAI's ChatGPT

trained on massive amounts of text data from the internet, such as websites, books, Wikipedia, up to September 2021

A large language model, LLM, is a deep learning model designed to understand, and generate human language.

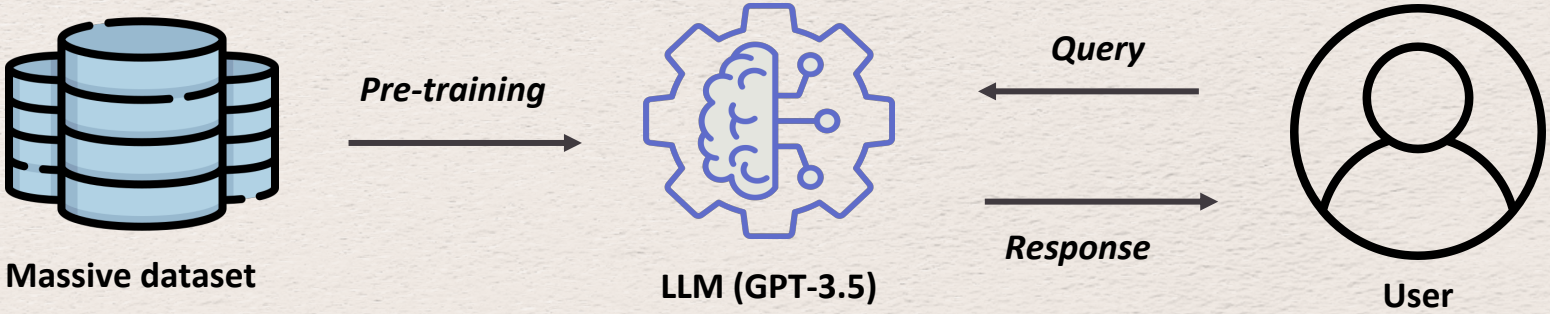


model learns statistical patterns and associations between words and phrases in the text.

Basic concept of a chatbot – OpenAI's ChatGPT

trained on massive amounts of text data from the internet, such as websites, books, Wikipedia, up to September 2021

A large language model, LLM, is a deep learning model designed to understand, and generate human language.



model learns statistical patterns and associations between words and phrases in the text.

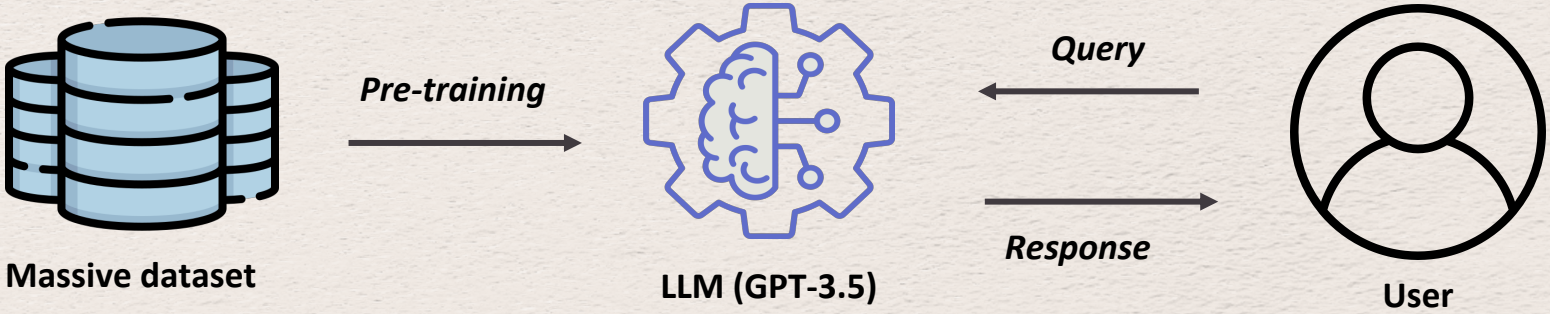
model estimates probabilities of possible words that can follow the words and context of the query
-> samples from distribution to select the word to form a coherent and contextually relevant response.

Basic concept of a chatbot – OpenAI's ChatGPT

Knowledge not up to date, and may not perform well in answering domain-specific queries

trained on massive amounts of text data from the internet, such as websites, books, Wikipedia, up to September 2021

A large language model, LLM, is a deep learning model designed to understand, and generate human language.

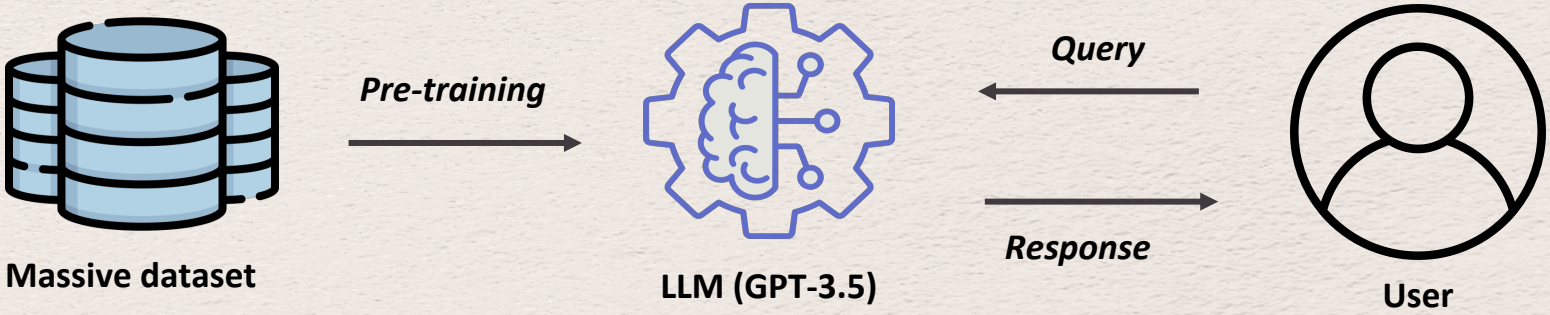


model learns statistical patterns and associations between words and phrases in the text.

model estimates probabilities of possible words that can follow the words and context of the query
-> samples from distribution to select the word to form a coherent and contextually relevant response.

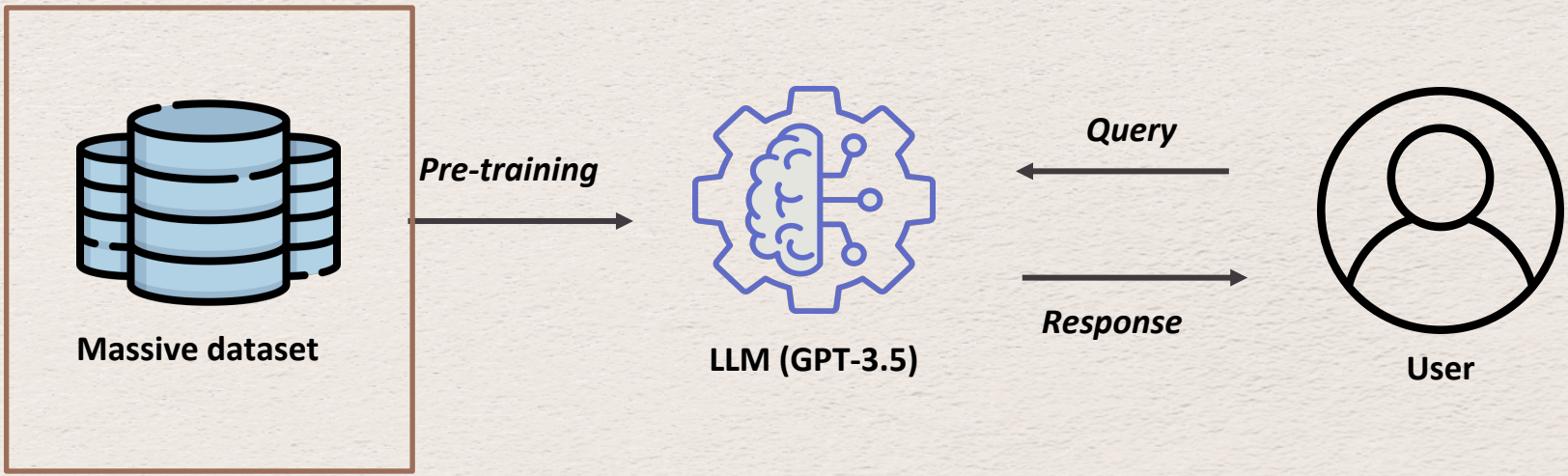
Two techniques to create custom chatbot with own data –

1) Retrieval Augmented Generation (RAG)
2) Fine-tuning



Two techniques to create custom chatbot with own data –

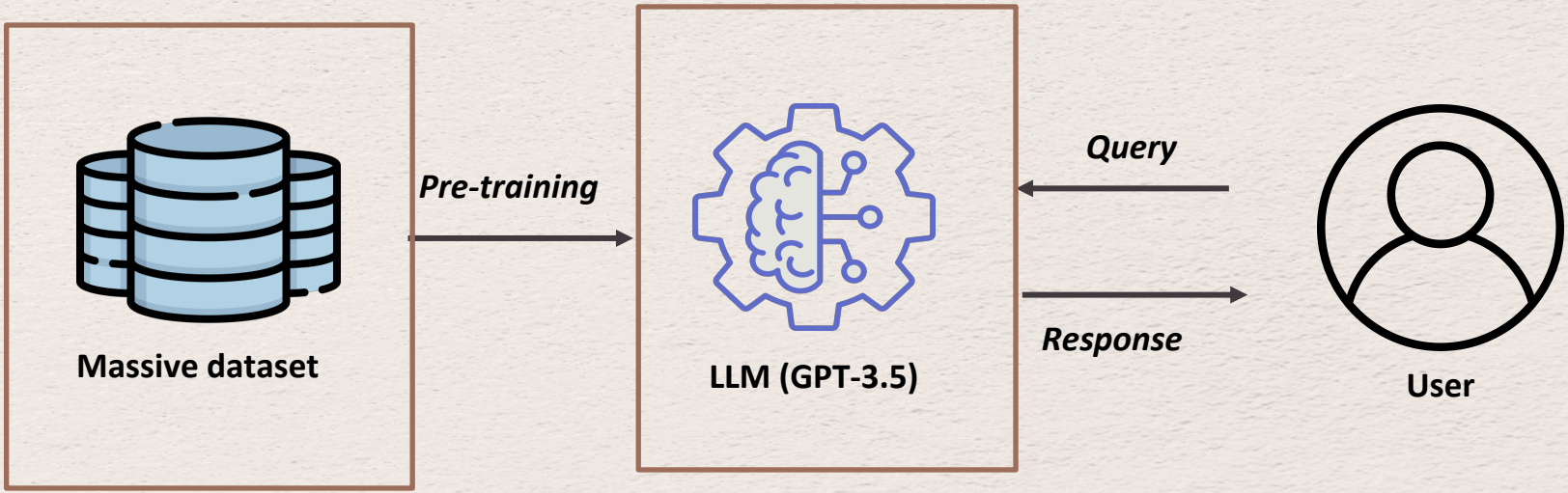
1) Retrieval Augmented Generation (RAG)
2) Fine-tuning



RAG adjusts knowledge the LLM has access to through external knowledge retrieval

Two techniques to create custom chatbot with own data –

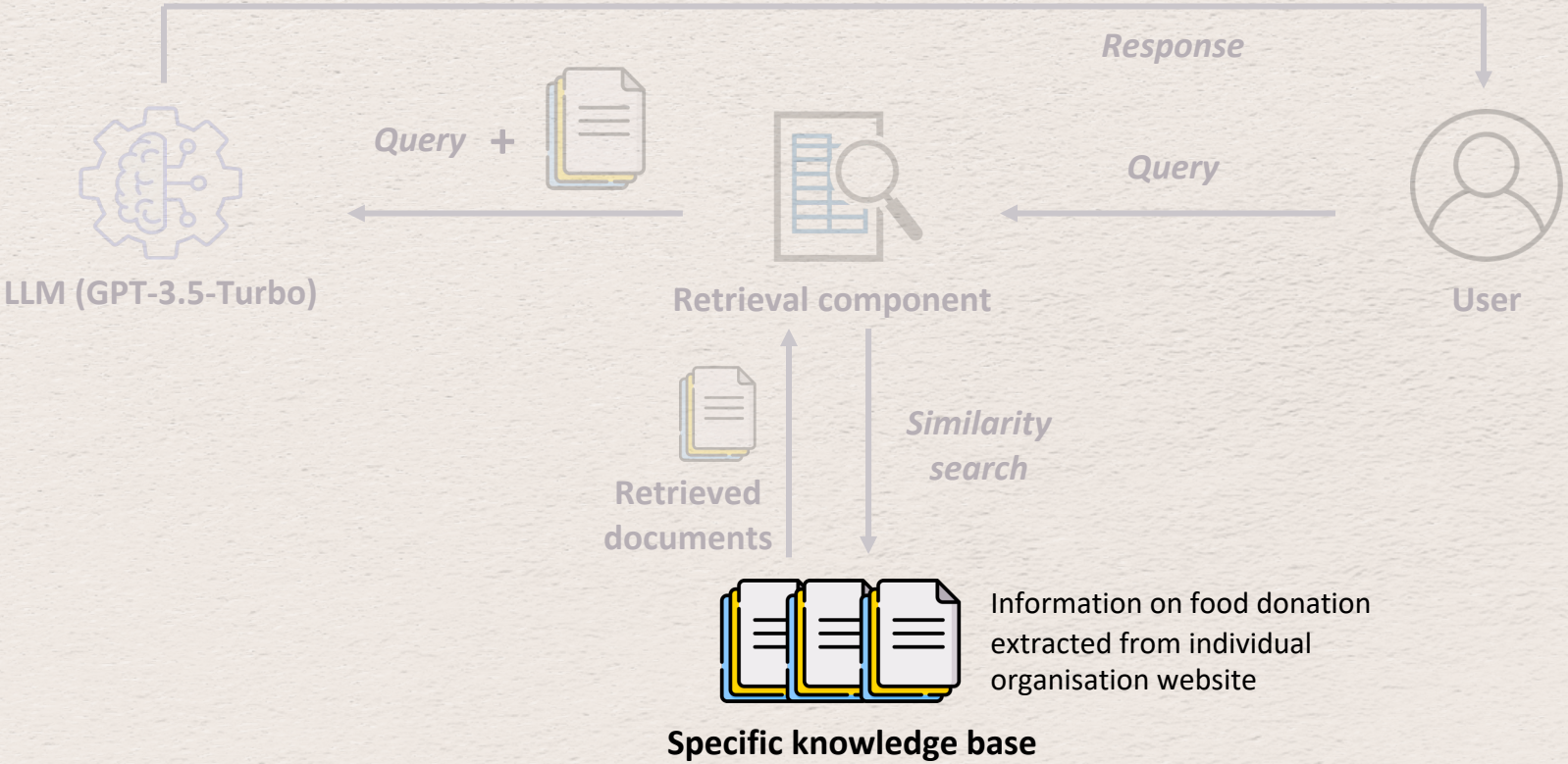
1) Retrieval Augmented Generation (RAG)
2) Fine-tuning



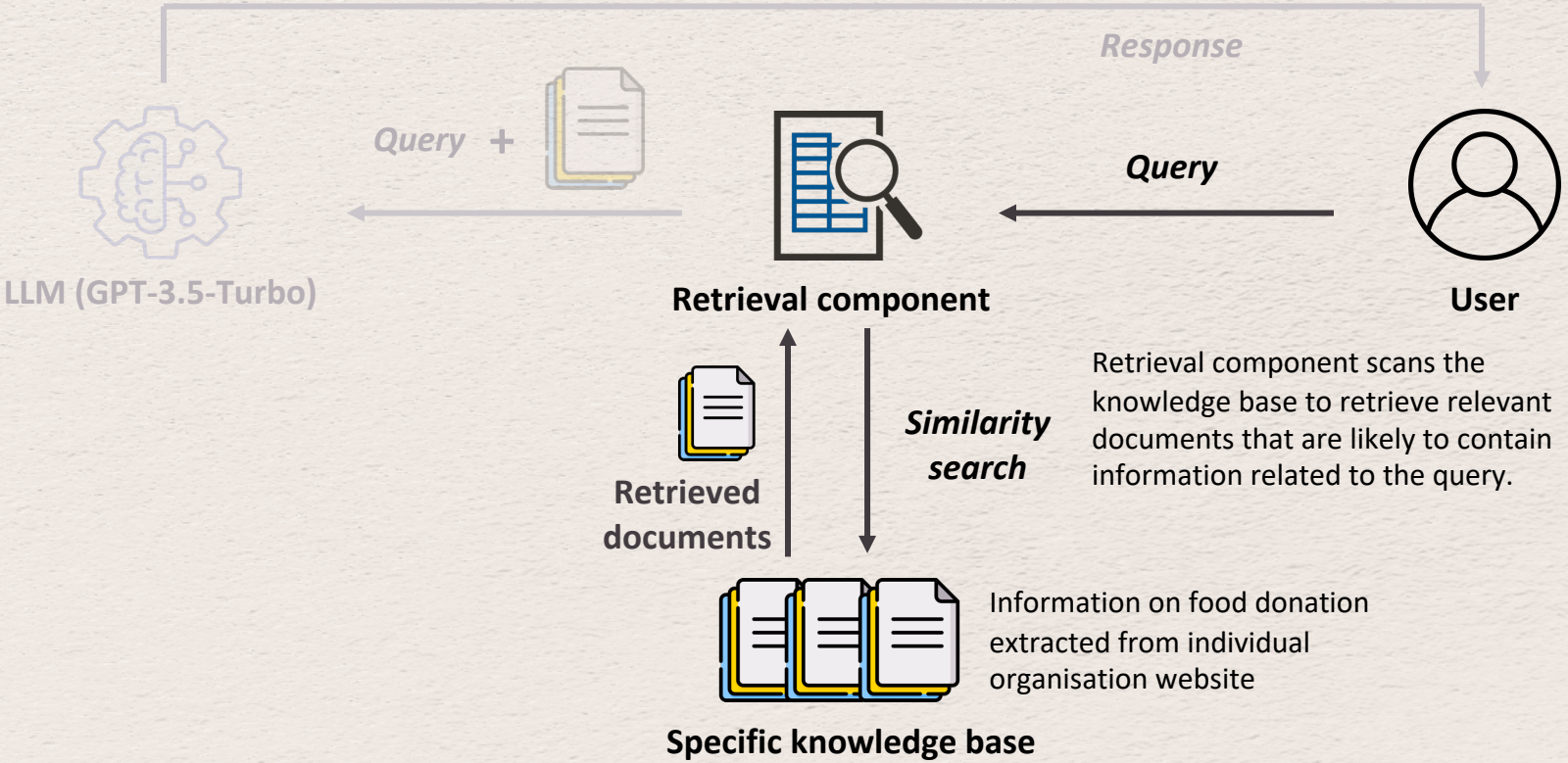
RAG adjusts knowledge the LLM has access to through external knowledge retrieval

Fine-tuning adjusts the behaviour of the LLM for specific tasks or domains by training it on a specific dataset

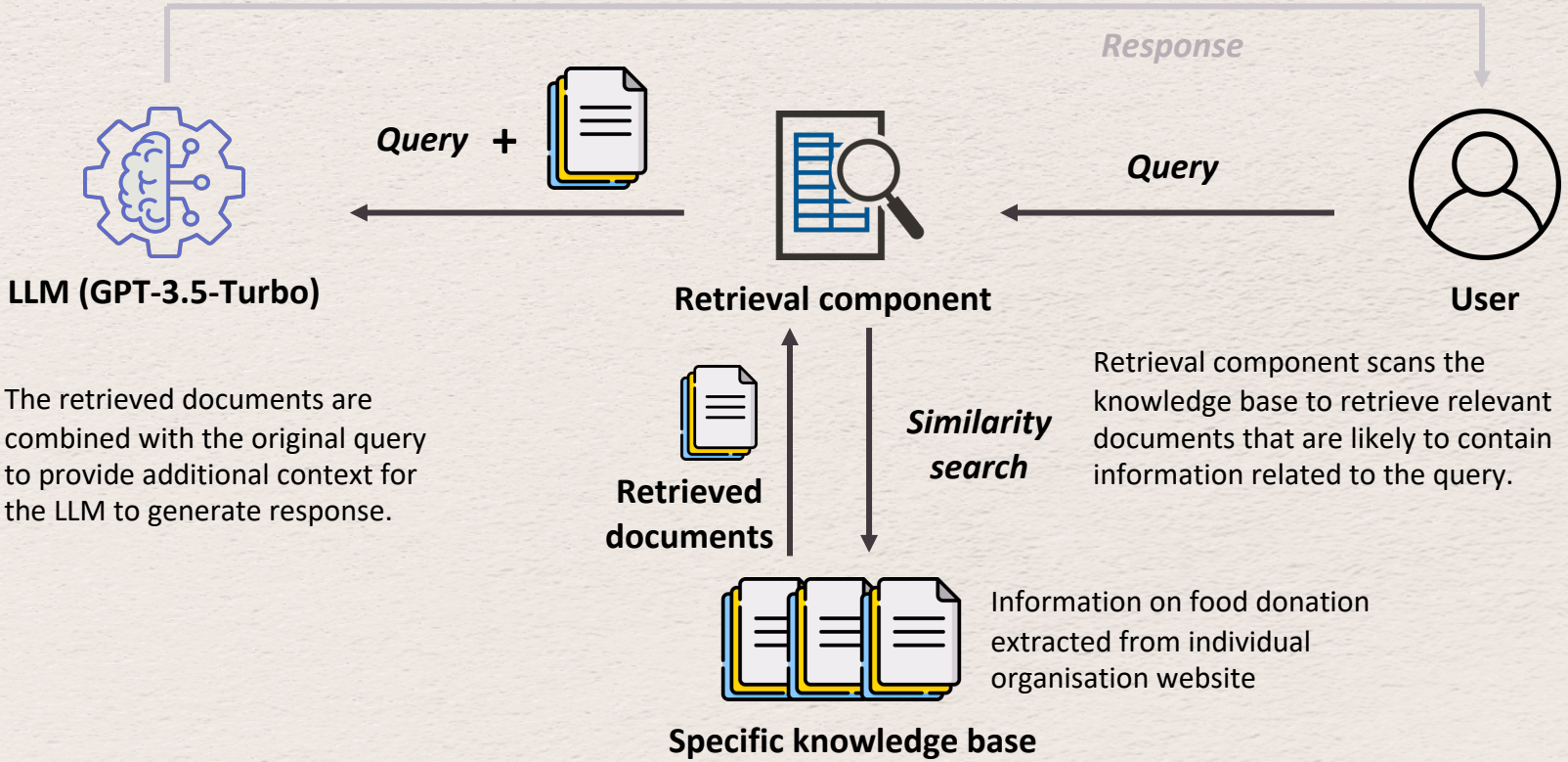
RAG adjusts knowledge the LLM has access to



RAG adjusts knowledge the LLM has access to



RAG adjusts knowledge the LLM has access to



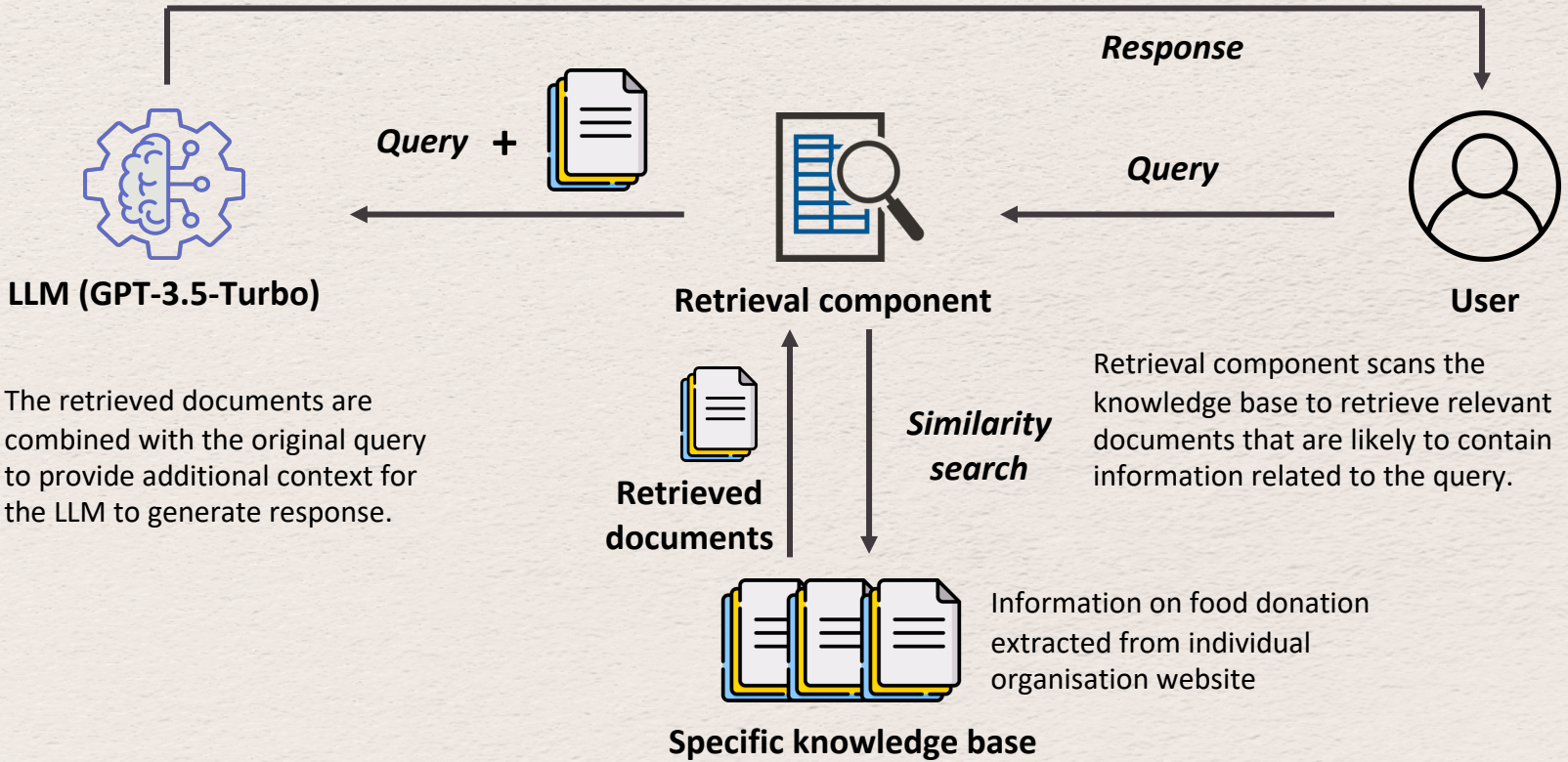
The retrieved documents are combined with the original query to provide additional context for the LLM to generate response.

Retrieval component scans the knowledge base to retrieve relevant documents that are likely to contain information related to the query.

Information on food donation extracted from individual organisation website

RAG adjusts knowledge the LLM has access to

RAG refines the probability distribution as words and phrases present in the retrieved context are assigned higher probabilities.

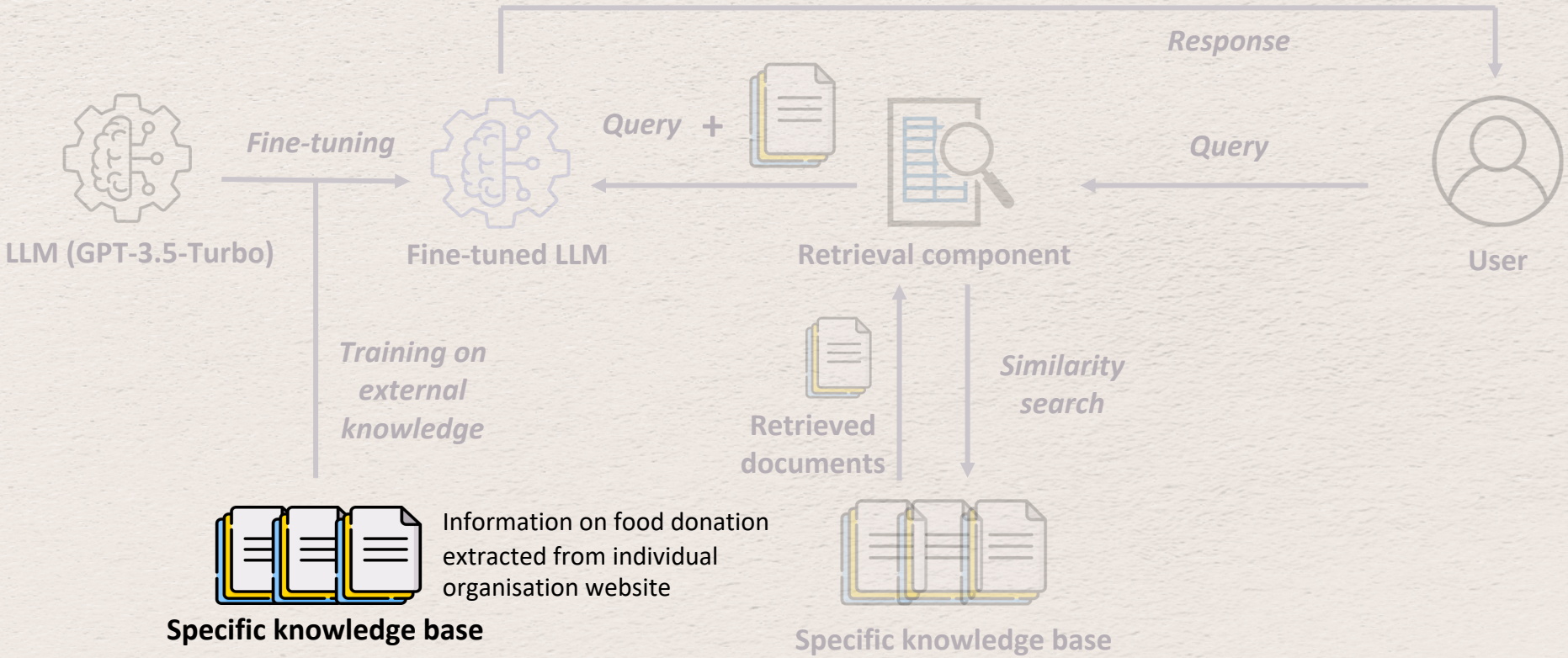


The retrieved documents are combined with the original query to provide additional context for the LLM to generate response.

Retrieval component scans the knowledge base to retrieve relevant documents that are likely to contain information related to the query.

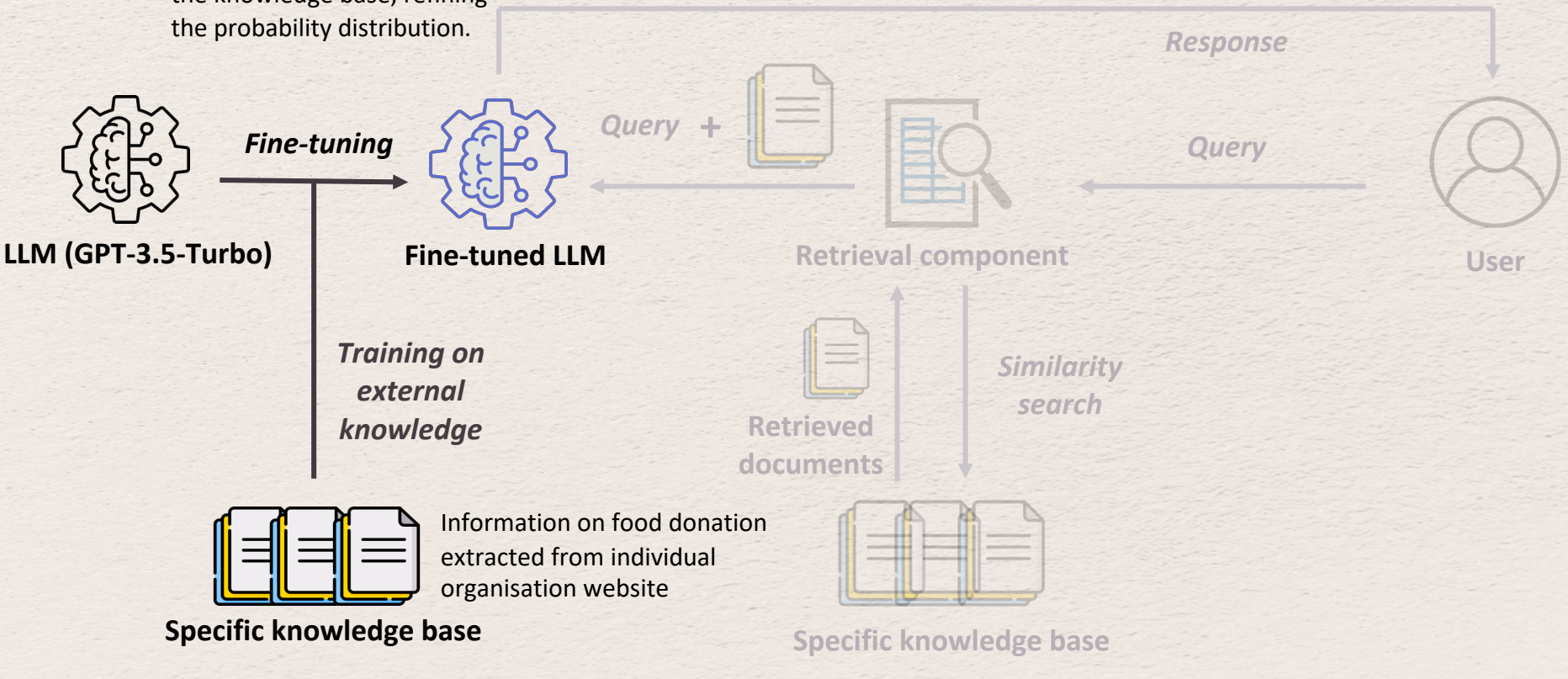
Information on food donation extracted from individual organisation website

Combine RAG and Fine-tuning to further improve the performance



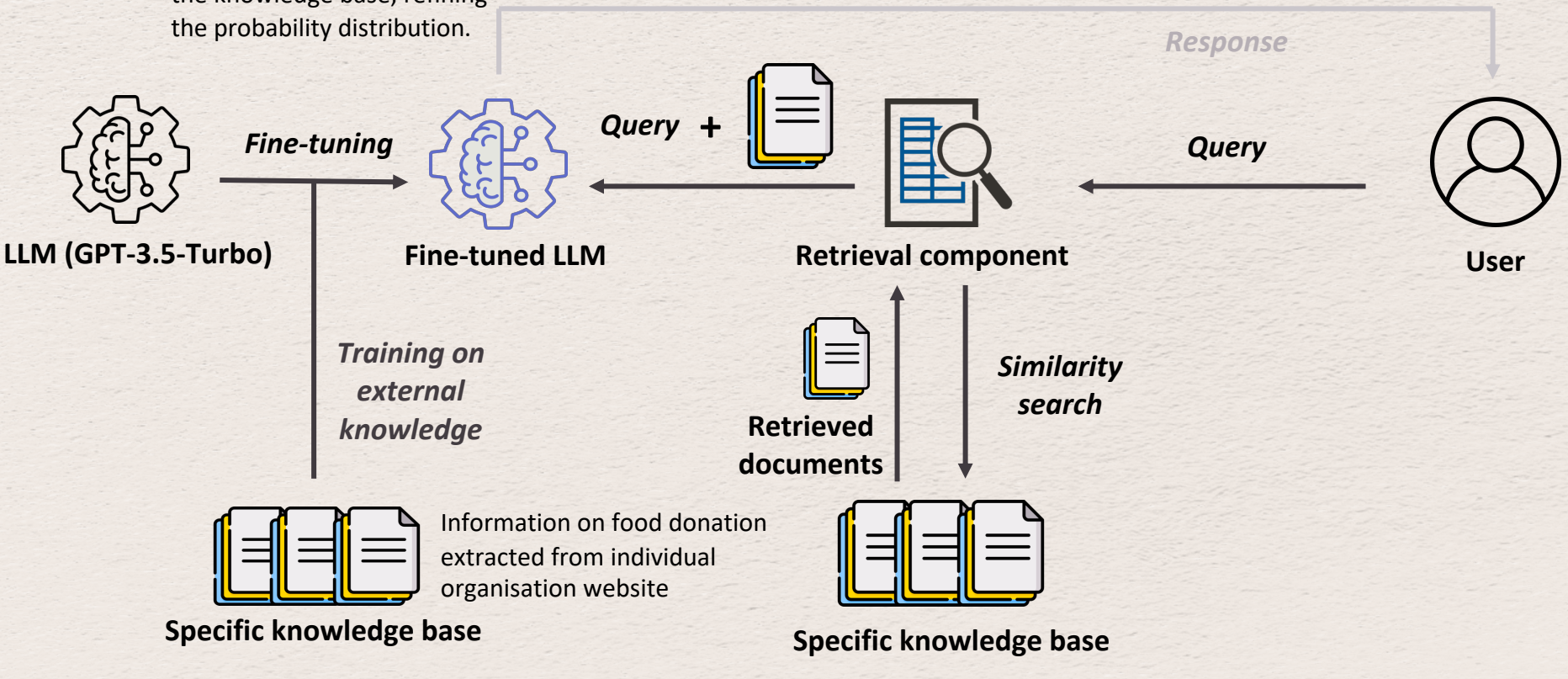
Combine RAG and Fine-tuning to further improve the performance

Fine-tuned model learns to assign different probabilities to words or phrases present in the knowledge base, refining the probability distribution.



Combine RAG and Fine-tuning to further improve the performance

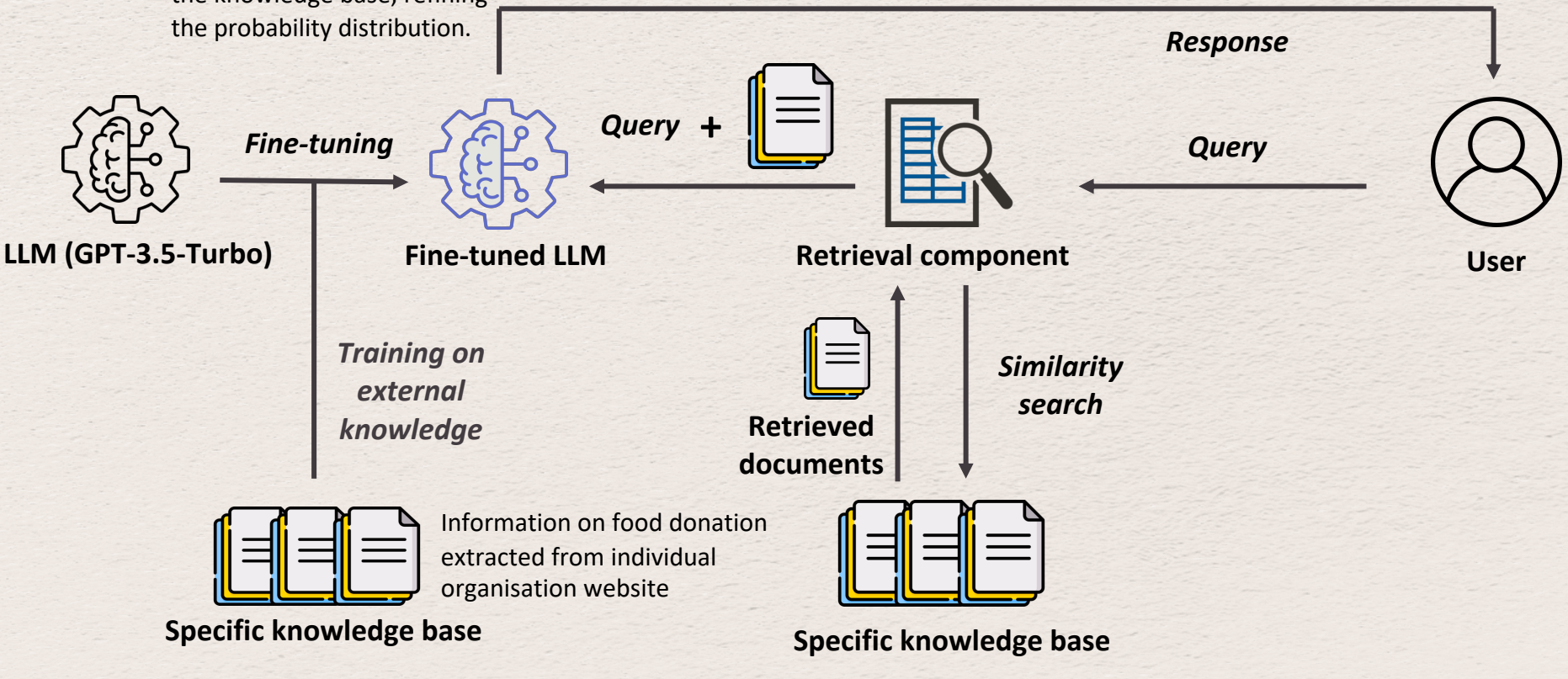
Fine-tuned model learns to assign different probabilities to words or phrases present in the knowledge base, refining the probability distribution.



Combine RAG and Fine-tuning to further improve the performance

Fine-tuned model learns to assign different probabilities to words or phrases present in the knowledge base, refining the probability distribution.

In combination with RAG, the likelihood of words and phrases present in the external dataset increases further, improving response relevancy and accuracy.



Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Answer relevancy : Measures if the generated answer can directly and appropriately address the question, i.e. answers that are complete and do not include unnecessary or duplicated information.

Score ranges from 0 to 1, higher score indicates concise and informative answers.

Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Answer relevancy : Measures if the generated answer can directly and appropriately address the question, i.e. answers that are complete and do not include unnecessary or duplicated information.

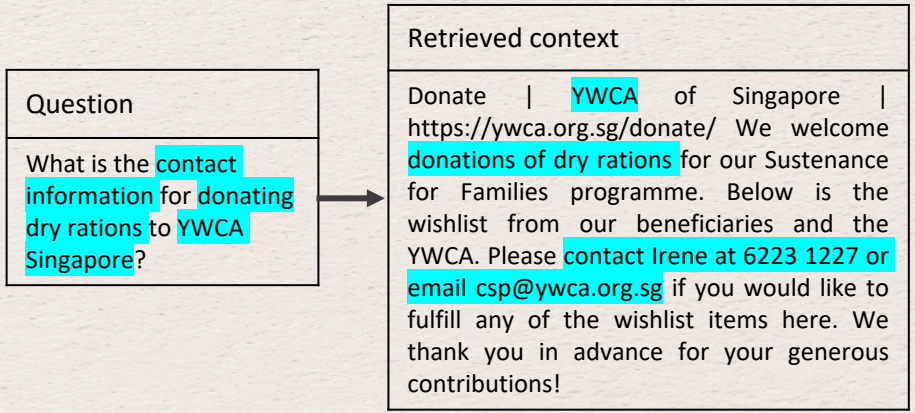
Score ranges from 0 to 1, higher score indicates concise and informative answers.

Question
What is the contact information for donating dry rations to YWCA Singapore?

Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Answer relevancy : Measures if the generated answer can directly and appropriately address the question, i.e. answers that are complete and do not include unnecessary or duplicated information.

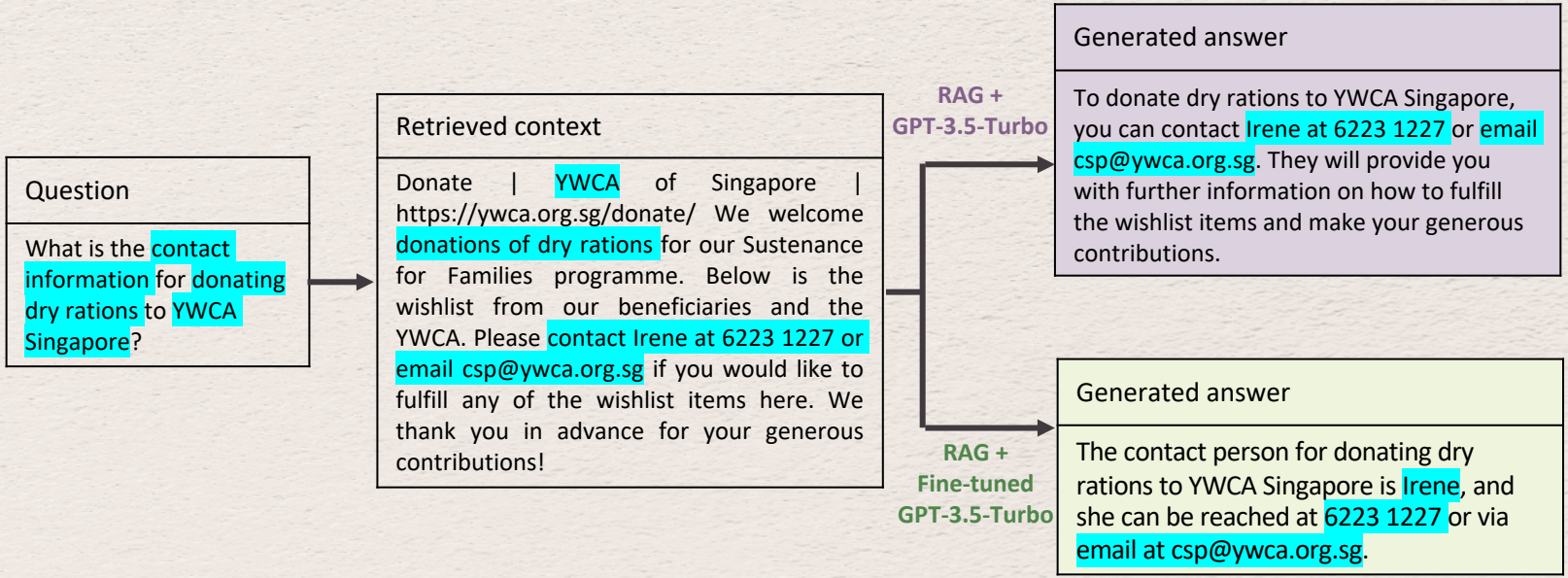
Score ranges from 0 to 1, higher score indicates concise and informative answers.



Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Answer relevancy : Measures if the generated answer can directly and appropriately address the question, i.e. answers that are complete and do not include unnecessary or duplicated information.

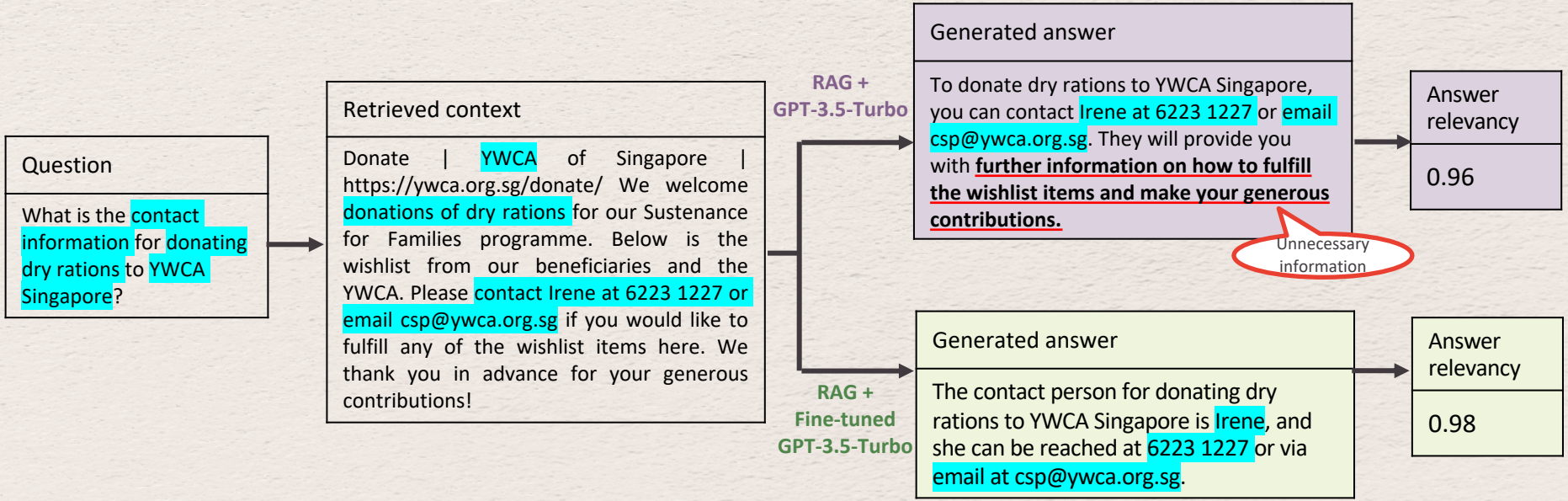
Score ranges from 0 to 1, higher score indicates concise and informative answers.



Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Answer relevancy : Measures if the generated answer can directly and appropriately address the question, i.e. answers that are complete and do not include unnecessary or duplicated information.

Score ranges from 0 to 1, higher score indicates concise and informative answers.



Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Faithfulness: Measures how factually accurate the generated answer is, i.e. answers that are derived from retrieved contexts and not hallucinated.

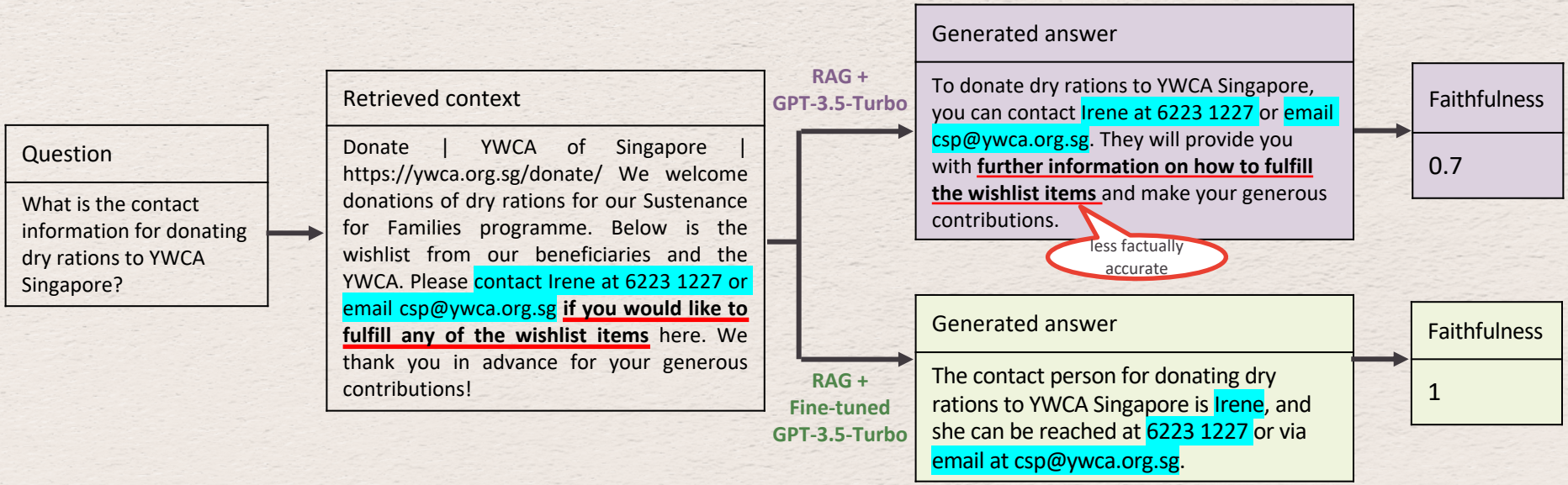
Score ranges from 0 to 1, higher score indicates reliability and truthfulness of the model to provide contextually accurate information.

Hallucinations: instances where the language model produces information or claims that are not accurate or supported by the input context.

Performance is evaluated using 2 metrics – answer relevancy and faithfulness

Faithfulness: Measures how factually accurate the generated answer is, i.e. answers that are derived from retrieved contexts and not hallucinated.

Score ranges from 0 to 1, higher score indicates reliability and truthfulness of the model to provide contextually accurate information.



Hallucinations: instances where the language model produces information or claims that are not accurate or supported by the input context.

RAG + GPT-3.5-Turbo is chosen for building the chatbot due to high performance and ease of scalability

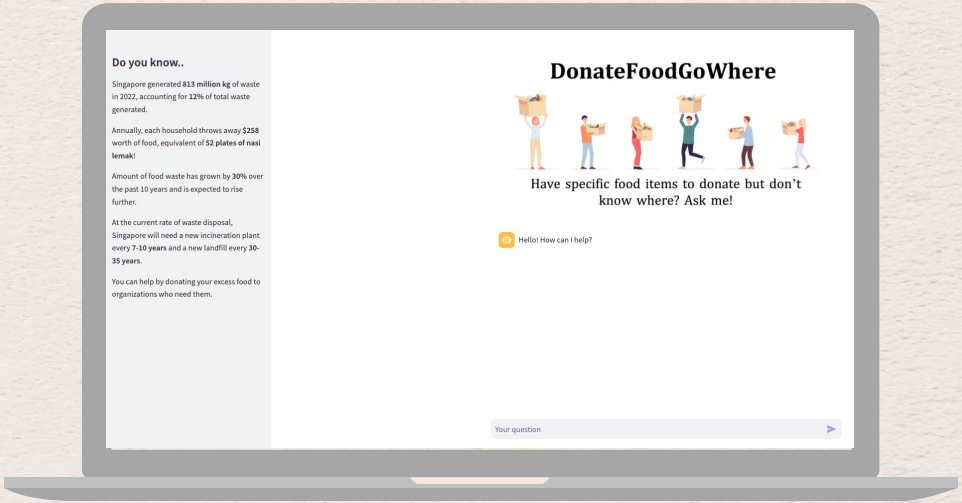
Model pipeline	Answer relevancy	Faithfulness
RAG + GPT-3.5-Turbo	0.97	0.87
RAG + Fine-tuned GPT-3.5-Turbo	0.98	0.88

High performance – Instead of fine-tuning the entire model on the specific dataset, we get comparable results using a base model, which involves far less computing resources.

Easily of scalability – Food items on organisation's wishlist change frequently, RAG alone can easily and quickly adapt to new data.

Chatbot demo

https://donatefoodgowhere.streamlit.app/



SCAN ME

Summary

As demonstrated by the chatbot, with just 1 or 2 query, individuals can easily and quickly find out where and how to donate specific food items with the relevant instructions or information.

Moving forward

Phase 1

Conduct road shows

- Expand list of organisations

Phase 2

First release of chatbot

- Integrate with NEA website – food waste management

Phase 3

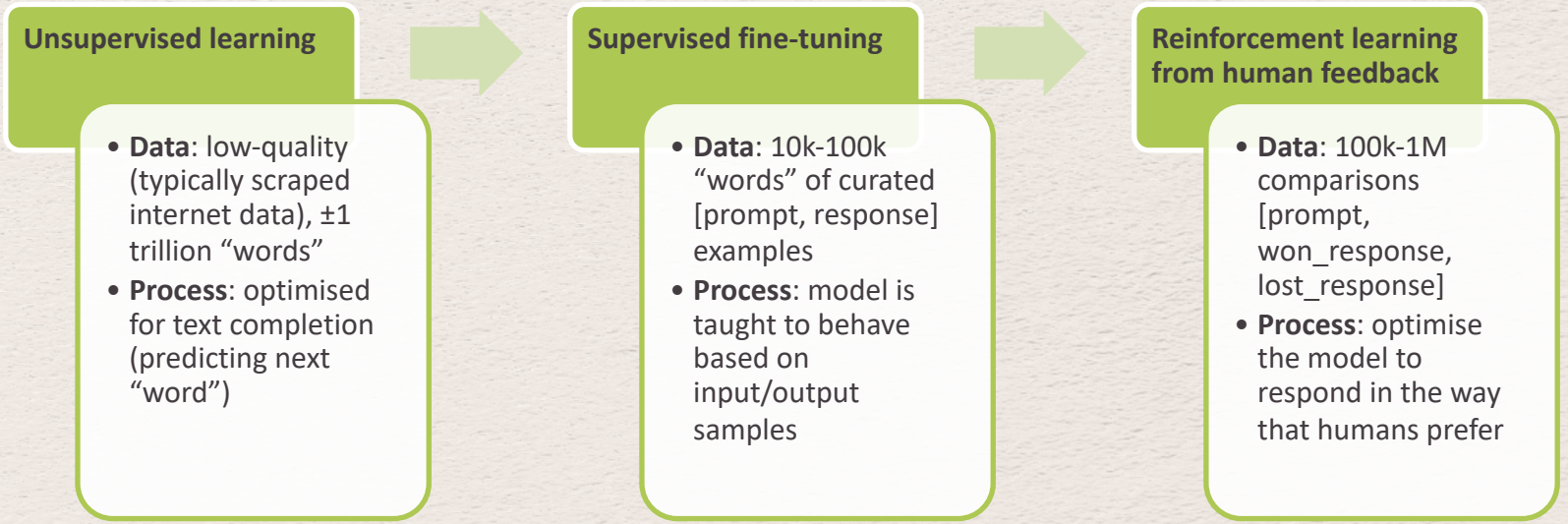
Second release of chatbot

- Based on feedback, apply prompt engineering to tune the response behaviour of the chatbot

Thank You!

Appendix

Large language model is trained in 3 general steps



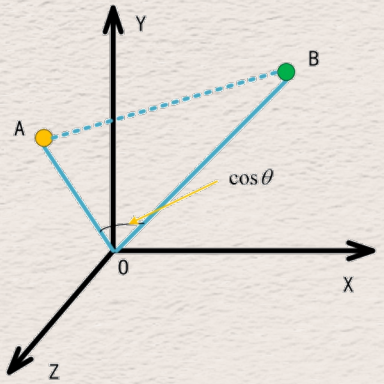
Source: https://blog.ml6.eu/to-fine-tune-or-not-to-fine-tune-837996713913

How does retriever component of RAG works

External knowledge documents → Convert into vector embeddings → Stored in vector database

Similarity search using cosine similarity → Return context vector with the shortest similarity distance

Query → Convert into vector embeddings



Vector embeddings: numerical representation of the text that captures semantic relationships between words.
Cosine similarity: a metric used to measure how similar two items are. It measures the cosine of the angle between two vectors projected in a multi-dimensional space. The output value ranges from 0–1 where the higher the value, the more similar both items are.

Computation of answer relevancy and faithfulness metric

Answer relevancy: Calculated using answer and question. For a generated answer, the LLM is prompted to generate an appropriate question for the multiple times, and the mean cosine similarity between these generated questions and the original question is measured. The underlying idea is that if the generated answer accurately addresses the initial question, the LLM should be able to generate questions from the answer that align with the original question, i.e. high mean cosine similarity, translating to high score.

Faithfulness: Calculated using answer and retrieved context. The LLM identifies statements within the generated answer and verifies if each statement is supported by the retrieved context.

$$\text{Faithfulness} = \frac{\text{number of statements in the answer that can be logically inferred from the context}}{\text{total number of statements in the answer}}$$

Cosine similarity: a metric used to measure how similar two items are. It measures the cosine of the angle between two vectors projected in a multi-dimensional space. The output value ranges from 0–1 where the higher the value, the more similar both items are.