



كلية معتمدة من الهيئة القومية  
لضمان جودة التعليم والإعتماد



# *MENTORIA*

## Team Members

Mohamed Hassan Fathy

## Under Supervision

Dr. Ahmed Hosny

Dr. Mohamed Yousef

# TABLE OF CONTENT

1. Overview of *MENTORIA*
2. Motivation
3. What is Gen AI?
4. What is the LLM?
5. Problems with Vanilla LLM
6. Existing Solutions & Their Limitations
7. The RAG: Description and Key Features
8. Technical Implementation: Tech Stack
9. Future Work
10. Final Thoughts



**Multi-functional Educational Navigator Toolkit for Outstanding Results and Interactive Assistance**

**A chatbot that enables you to chat with your data from different sources: documents or URLs.**

## Data Sources

URL

<https://medium.com/ai-advances/unbelievable-run-70b-llm-inferen>

Upload your docs here

Drag and drop files here

Limit 200MB per file

Browse files



paper.docx

214.7KB



Octopus.pdf

0.6MB



Mentoria Definition.txt

370.0B



GO



**MENTORIA**



Welcome to Mentoria, how can I help you?



What is event-loop?



The event-loop is the main core of running and executing asynchronous code on Node.js. Node.js Event loop registers your event with their associated listeners and once the operation is completed it executes its callback listener with the specified data. The only part that runs on a single thread on the Node.js is the event loop itself.

Enter your prompt here . . .





## Data Sources

URL

Upload your docs here

Drag and drop files here

Limit 200MB per file

Browse files



SIGCSE\_TS\_2024\_submission\_1369 [Zeyad Ahmed]... X  
192.1KB

GO



# MENTORIA



Welcome to Mentoria, how can I help you?



What is Roro8?



Roro8 is a system that emphasizes CPU control over a bus for memory and I/O. Its main microprocessor is RX8, which is an 8-bit, little-endian processor. I do not have any further information about Roro8.

Enter your prompt here . . .



# Motivation

- Explore the Generative AI.
- Deep dive into transformers and GPT models.
- Understand the architecture and impact on NLP.
- Apply knowledge to real-world challenges.

# What is Generative AI?

Generative AI is a technology that enables users to quickly generate new content based on a variety of inputs.

Inputs and outputs to these models can include **text**, **images**, **sounds**, **animation**, or other types of data.

Utilizes deep learning algorithms to understand and replicate patterns in data, enabling the creation of highly realistic and complex content.

Applications span multiple industries, including creative arts, healthcare, gaming, and customer service, revolutionizing how we interact with technology and create digital experiences.

# What is the LLM?

Large Language Models (LLMs) are very large deep learning models trained on vast amounts of textual data to understand, generate, and manipulate human language.

They can perform a range of language-related tasks such as translation, summarization, question-answering, and creative writing.

The magic of the LLM is that it just tries to ***predict the next word!!***



Is this the best  
solution?

# Problems with Vanilla LLM

## Knowledge cutoff

What is the weather tomorrow?



I can't provide real-time weather updates because my training data only goes up until October 2023, and I don't have access to the internet to check current conditions. For the most accurate weather forecast, please refer to a reliable weather service or app.

A cartoon illustration of a man with bright orange, spiky hair and thin-rimmed glasses. He is shown from the chest up, in profile, looking towards the right. He has a neutral expression. The background consists of dark blue diagonal lines. The text "I'VE SEEN THIS SOMEWHERE BEFORE..." is overlaid at the bottom in a bold, white, sans-serif font with a black outline.

**I'VE SEEN THIS  
SOMEWHERE BEFORE...**

# Problems with Vanilla LLM

## Hallucination

What is the REST API?



A REST API is a system used to manage sleep schedules and monitor rest cycles for individuals. It helps in tracking how much rest a person gets and recommends optimal times for naps and sleep to improve overall health and productivity.

For a technical integration, it can connect with smart home devices to adjust the environment for better sleep.



# Solutions & Limitations

## Prompt Engineering

The practice of designing and refining prompts—questions or instructions—to elicit specific responses from AI models.

Think of it as the **interface** between human intent and machine output.

**Garbage in garbage out!!**

# Don't do this



Recipe with what's  
in my kitchen



Experience  
Seoul like a local



Thank my  
interviewer



Pick outfit to look  
good on camera



Tell me about the RESTful API|



# Do this



Recipe with what's  
in my kitchen



Experience  
Seoul like a local



Thank my  
interviewer



Pick outfit to look  
good on camera



Explain the RESTful API to me with clarifying the importance of it, the use cases and the other approaches that I can use instead of it. And finally, give me implementation of it in python.



# Solutions & Limitations

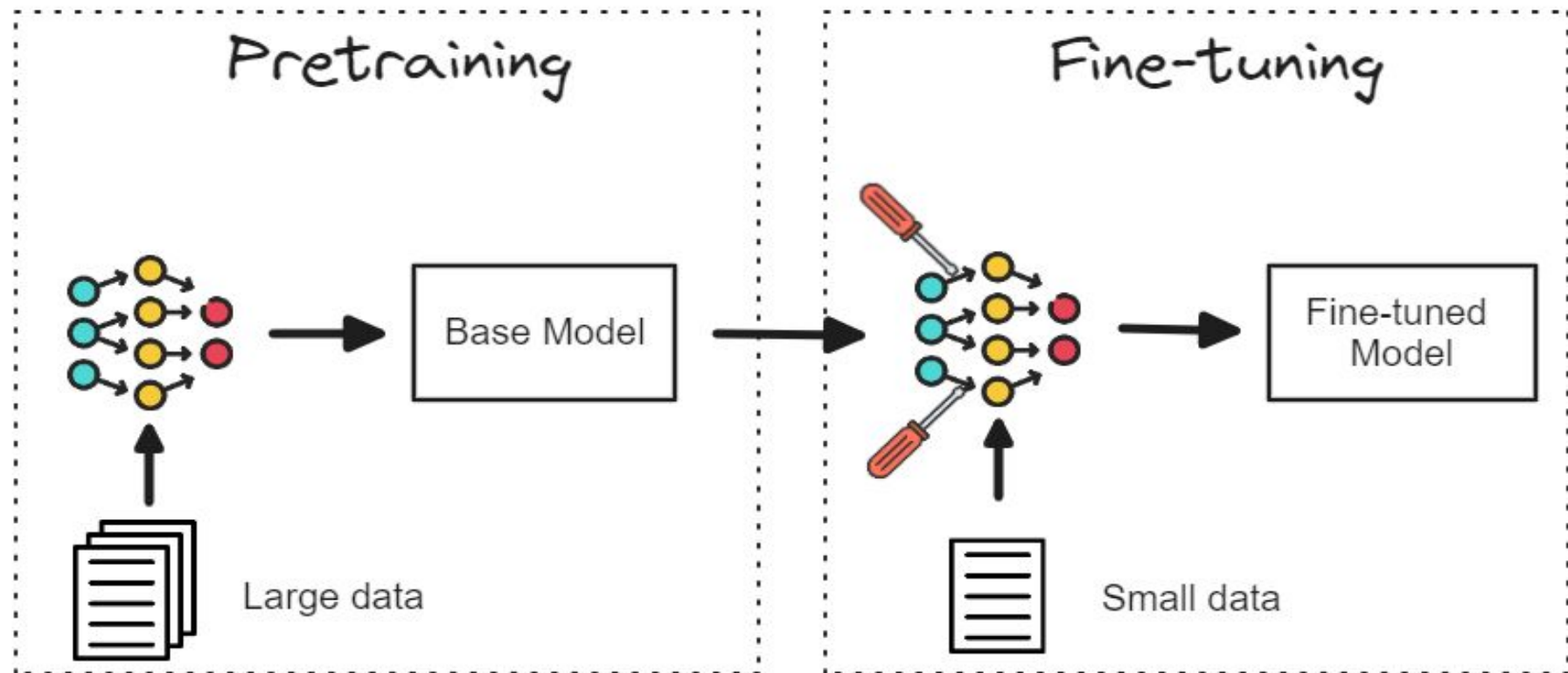
## Fine-tuning (PEFT)

Parameter-Efficient Fine-tuning, is an approach to **transfer learning** in which the parameters of a pre-trained model are trained on new data.

Fine-tuning can be done on the entire neural network, or on only a subset of its layers, in which case the layers that are not being fine-tuned are "frozen" (not updated during the backpropagation step).



# Large Language Model



# Solutions & Limitations

## Retrieval Augmented Generation (RAG)

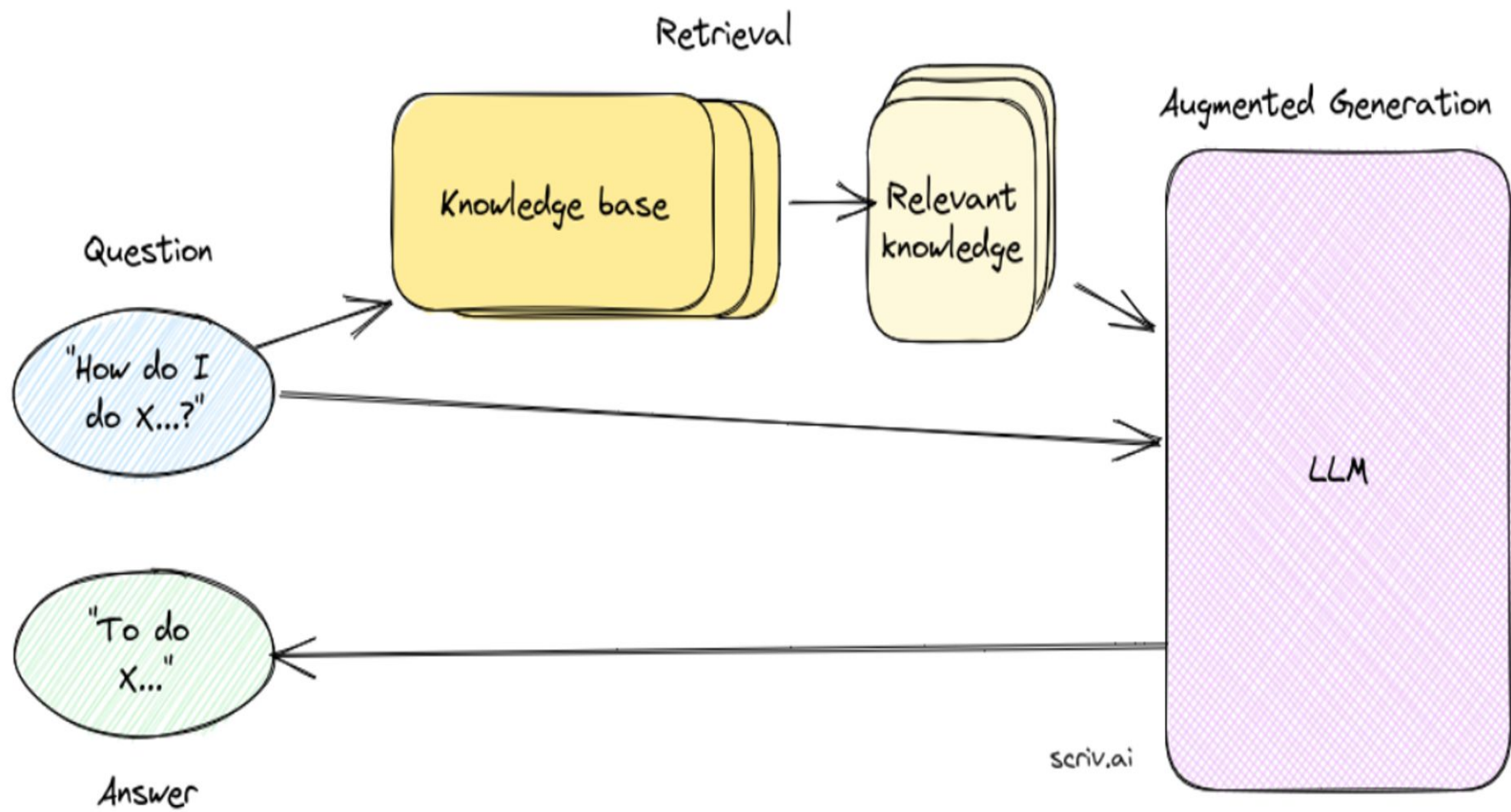
Imagine you're in a library. There are thousands of books and you have a question. But you don't know which book has the answer.

If you start searching, it may take you hours, or even days, to find an answer.

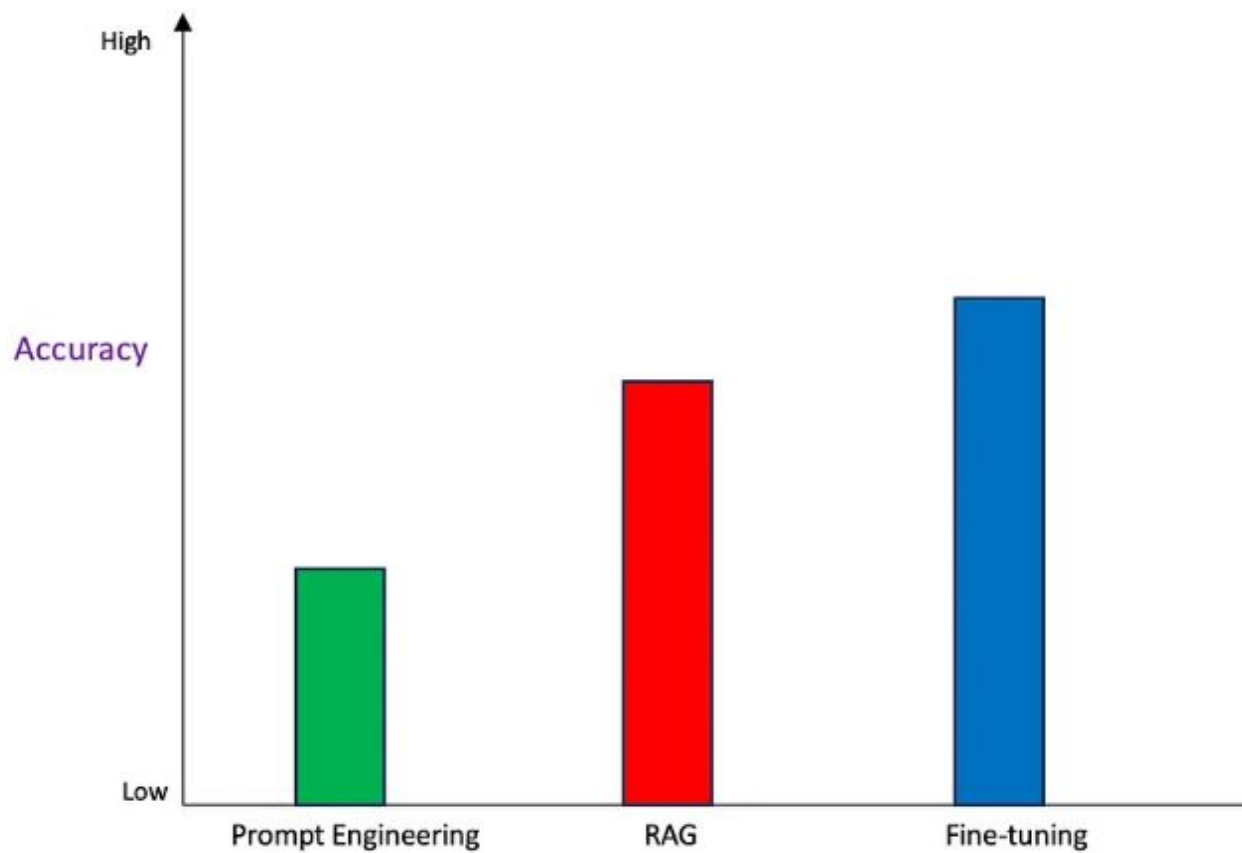
**What if, there was a digital agent in your possession to which when you ask the question, it scans all books quickly and fetches the best answer for you?**

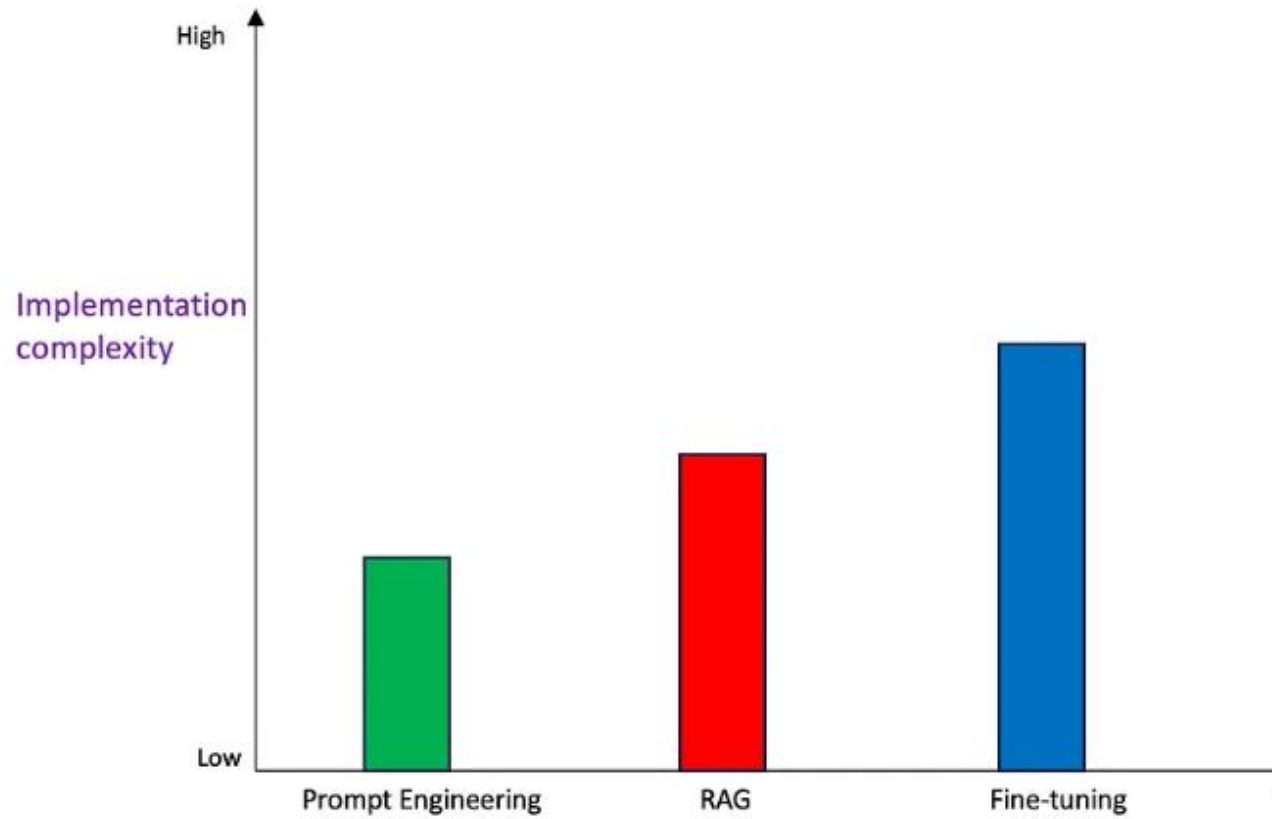


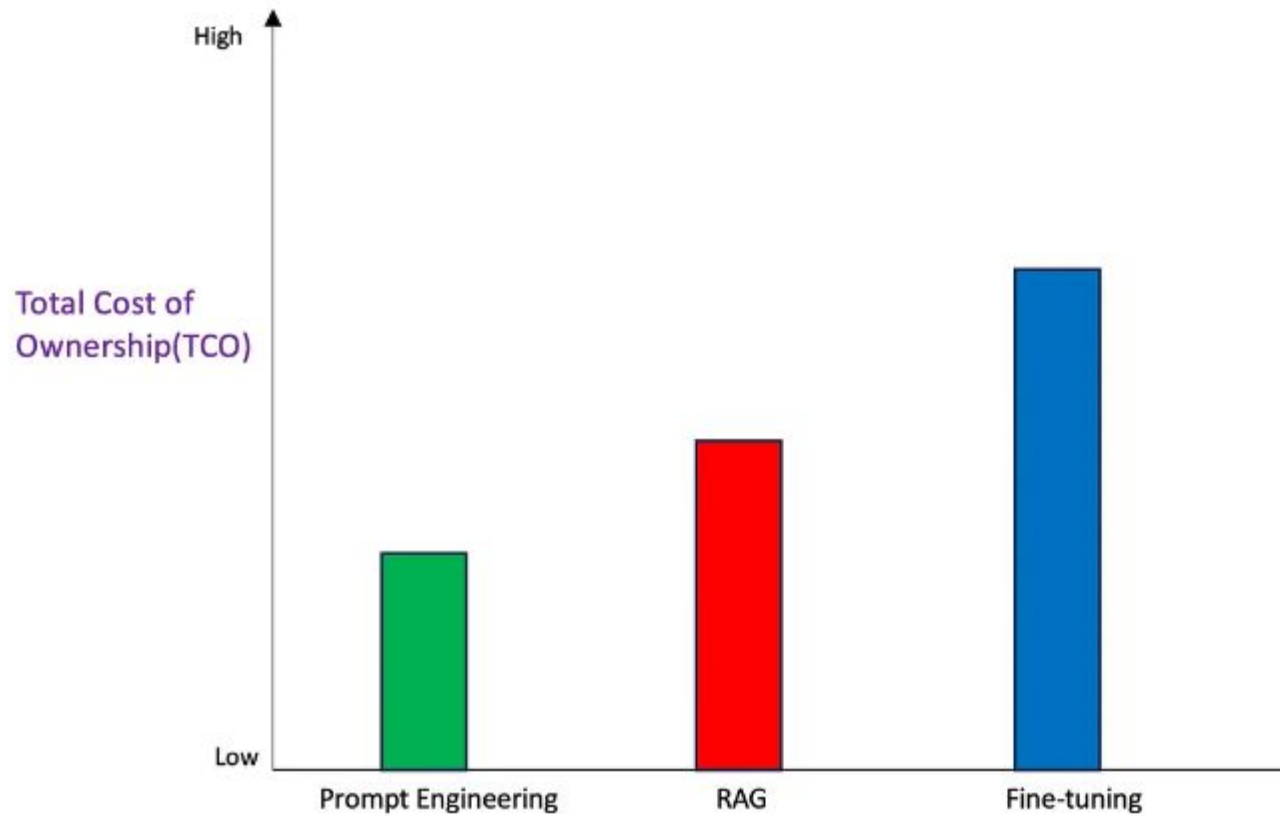




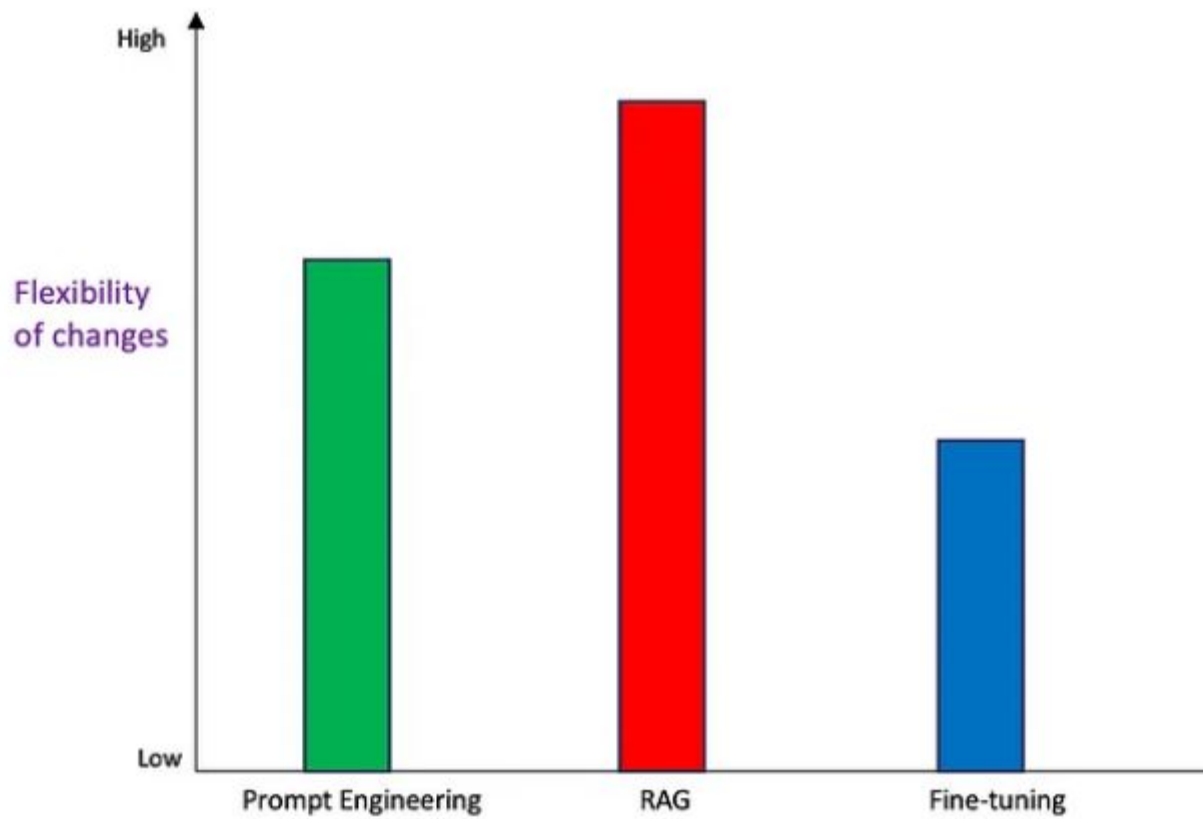
Let's  
Compare











# Vector Store!!

# Vector Store & Embeddings

Vector store stores the data in **numerical** form.

Embeddings means to map each token (word or part of it) to a vector of real numbers.

Word embeddings is capable of capturing the meaning of a word in a document, **semantic** and syntactic similarity, relation with other words.

**ChatGPT 3 maps each token in a vector length of 12288 dimensions!!**



Data



Traditional  
Database



Data

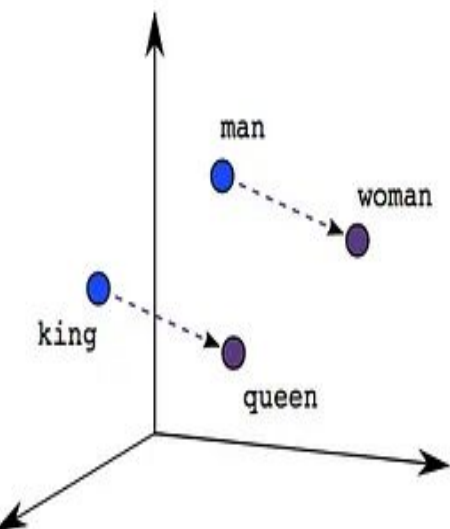


0101  
1001  
0101

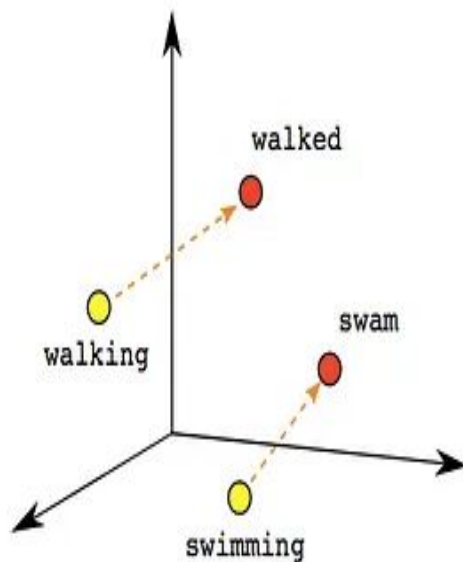
Vector



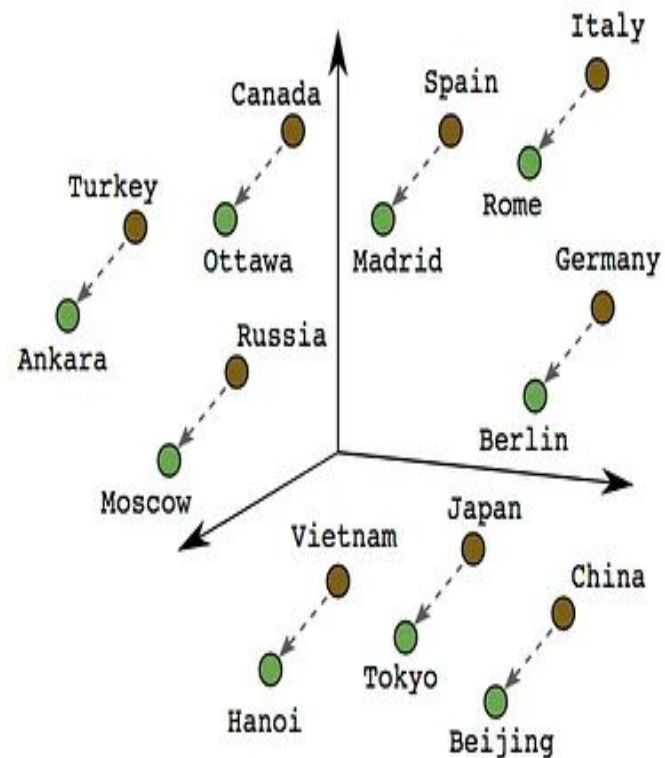
Vector Database



Male-Female



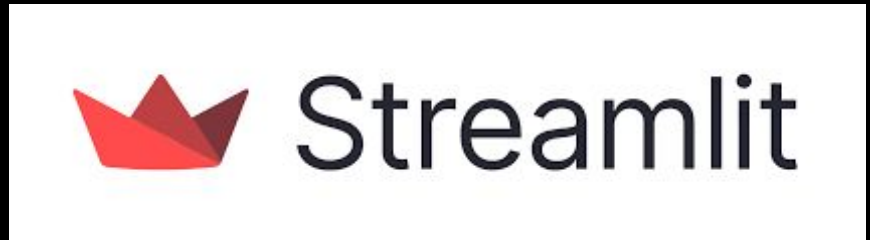
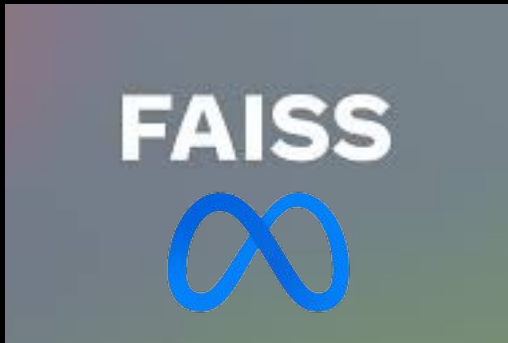
Verb Tense



Country-Capital



Gemini



# Future Work

- Incorporation of additional and bigger data sources.
- Enhanced natural language understanding.
- User interface and experience improvements.
- Give the ability to the user to choose a specific LLM, embedding model and specify the number of documents to retrieved from the vectorstore.
- Evaluation: automated testing and monitoring.

# Final Words



*crewai*



## MENTORIA

Graduation Project Documentation



Submitted by:  
**Mohamed Hassan Fathy**

Advisors:  
**Dr. Ahmed Hosny Ibrahim**  
**Dr. Mohamed Youssif Bassiouni**

6/30/2024 (V1.0)