DATA PROCESSING AND ARTIFICIAL INTELLIGENCE FOR ENHANCED INCIDENT MANAGEMENT

Members:
Tan Jun Xian Clement (Dunman High School)
Wong Qi Wen (Nanyang Girls' High School)

Mentors:
Dickson Tan Zhi Sheng
(Defence Science and Technology Agency)
Neo Jin
(Defence Science and Technology Agency)

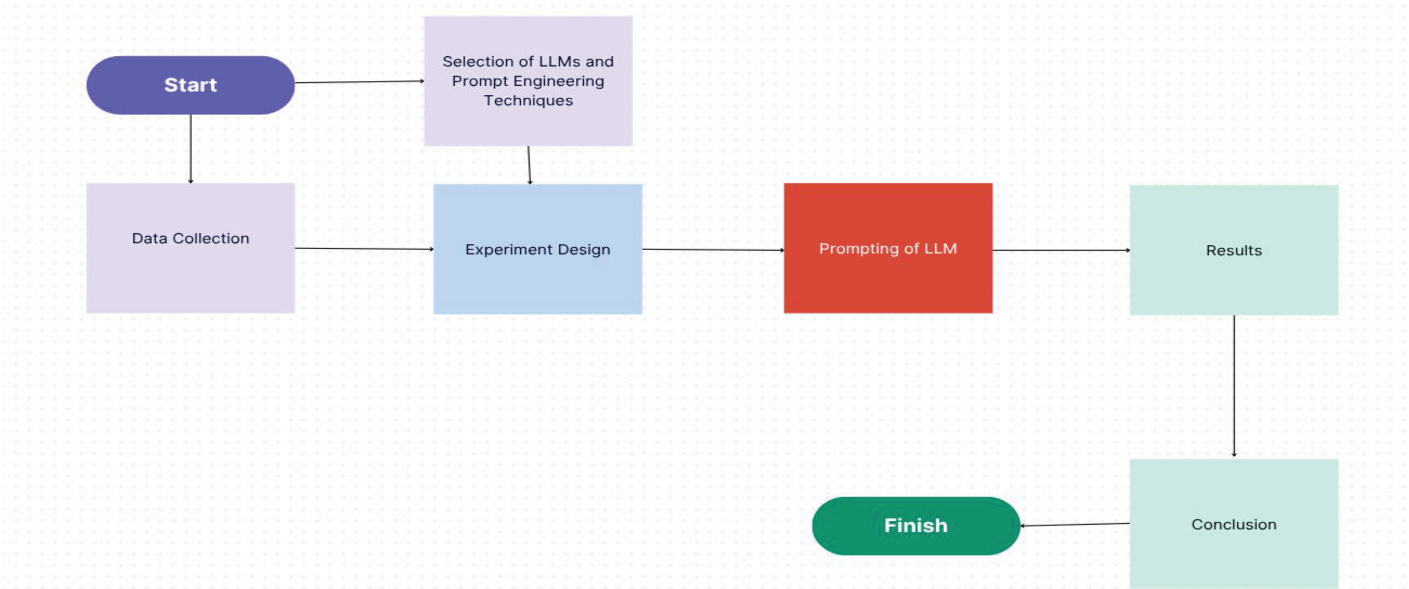
1. Introduction

Every day, there is a large volume of vehicles and people passing through Singapore's Land Checkpoints. With the emergence of Artificial Intelligence and Generative AI solutions, incident management at Singapore's Land Checkpoints can be optimised. This project aims to evaluate the use of LLMs and different prompt engineering techniques to find the most optimal way to retrieve information for ICA Land Checkpoint Operations.

2. Objectives

This project aims to evaluate the functionality of various existing LLMs, with a focus on their ability to retrieve information from structured and unstructured data. The goal is to enhance incident management effectiveness at Land Checkpoints. Potential operational scenarios include using LLMs to retrieve information from past incidents, which can aid in trend identification and decision making. The effectiveness of in-context in improving accuracy of incident retrieval will be explored, to identify optimal prompt engineering technique for efficient information retrieval, without manual searching or additional prompting.

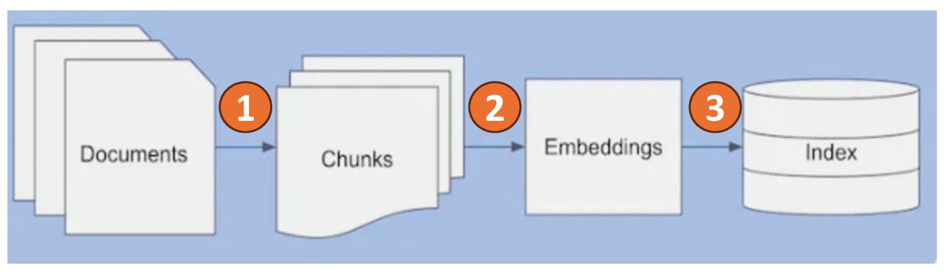
3. Methodology



1. **Data Collection**: ChatGPT Plus was used to generate 100 mock ICA incidents in csv file format.
2. **Selection of LLMs and Prompt Engineering Techniques**: 4 LLMs and 2 prompt engineering techniques, zero-shot and few-shot were selected.
3. **Experiment Design**: A RAG system was developed, using tools such as Python and LangChain, for us to send prompts to LLMs.
4. **Prompting of LLMs**: A total of 10 prompts for each prompt engineering technique was passed to 4 different LLMs.
5. **Results**: The responses generated by the LLMs were analysed for their accuracy and relevance.
6. **Conclusion**: The best LLM and prompt engineering for more accurate and relevant information retrieval to enhance land checkpoint operations.

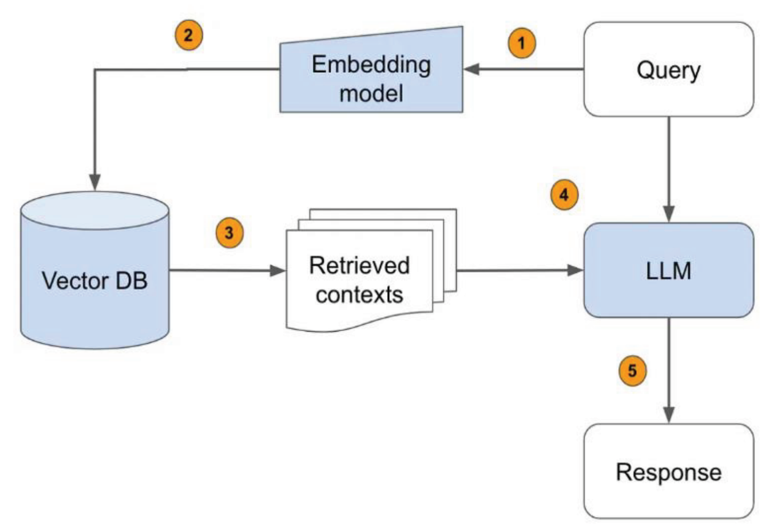
4. RAG Process

Step 1: Data Indexing



1. Incidents are split into chunks of appropriate size.
2. Chunks are encoded into embeddings.
3. Embeddings are stored in vector database.

Steps 2 and 3: Retrieval and Generation



1. Query passed to LLM, encoded into embeddings.
2. Query embedding passed into Vector Database.
3. Similarity search for most similar vectors to query vector in database.
4. Most similar contexts (vectors) passed to LLM.
5. LLM generates response.

Similarity Search Formula: $d(\mathrm{p}, \mathrm{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$

where:
- p and q are two vectors, p_i and q_i are the individual elements of the vectors.
- n is the dimensionality of the vectors.
- $d(\mathrm{p}, \mathrm{q})$ represents the Euclidean distance between vectors p and q.

5. LLMs and Prompt Engineering Techniques Explored

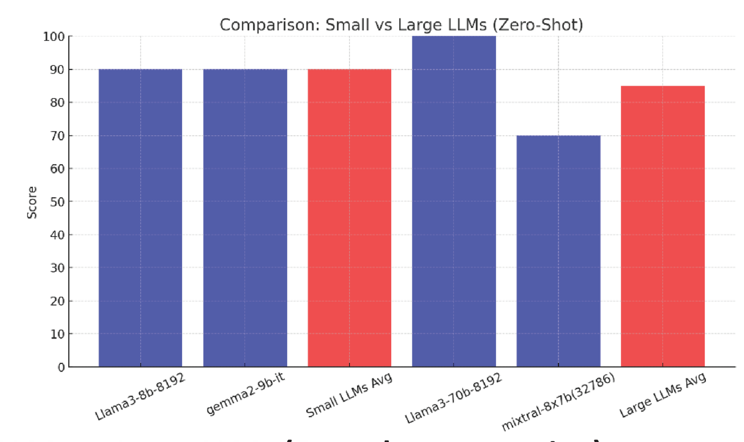
Large Language Models

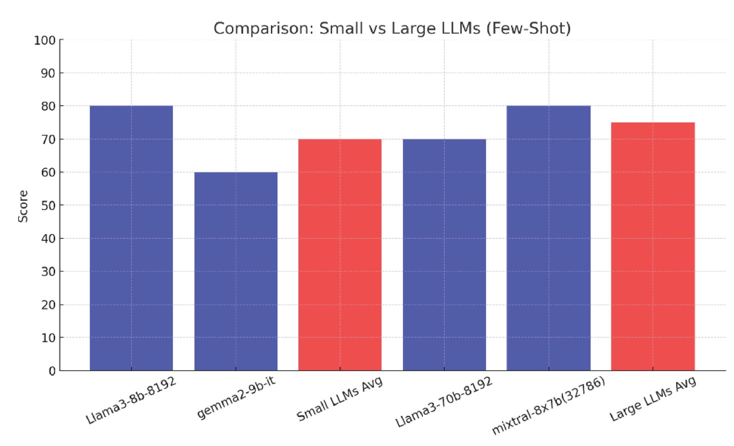
Model	Created By	Size Category	No. of Parameters
Llama3-8b-8192	Meta	Small	8 Billion
Gemma-9b-it	Google	Small	9 Billion
Llama3-70b-8192	Meta	Large	70 Billion
Mixtral-8x7b-32786	Mistral AI	Large	46 Billion

Prompt Engineering Techniques
- **Zero-Shot Prompting:** Using a LLM to perform a task without providing it with any explicit examples or prior demonstrations of the task in the prompt. The LLM relies entirely on its pre-trained knowledge to understand and execute the task based on the instructions provided.
- **Few-Shot Prompting (In-Context Learning):** Providing the LLM with a small number of examples or demonstrations of the desired task within the prompt, before asking it to complete a new, similar task. This could help the model better understand the context and expected output.

6. Results and Discussion

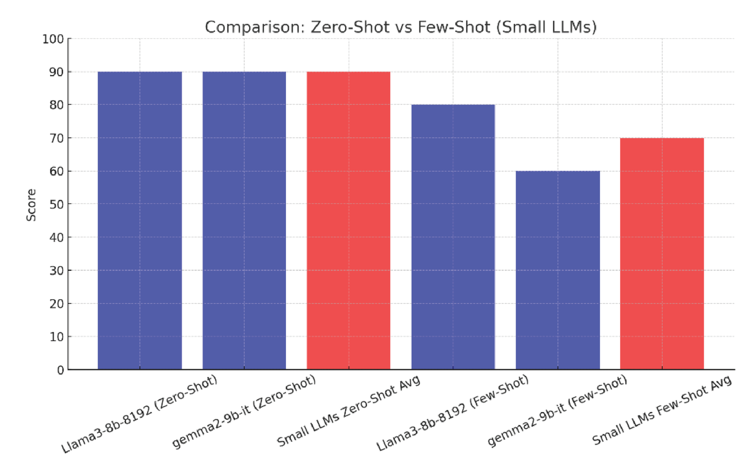
1) Quantitative Evaluation

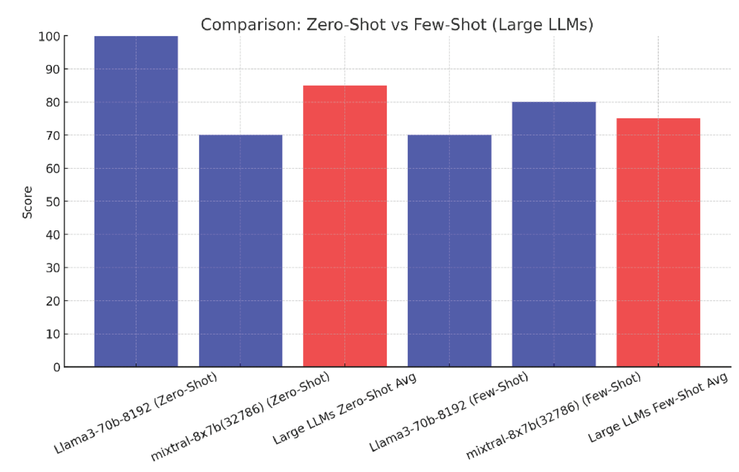




Small LLMs vs Large LLMs (Zero-shot prompting)
Small LLMs perform slightly better than Large LLMs in Zero-shot prompting, with higher accuracy score of 90% compared to 85% respectively. Small LLMs trained on less data, more reliant on information retrieval, hence perform better compared to Large LLMs which may blend contexts with pre-trained knowledge.

Small LLMs vs Large LLMs (Few-shot prompting)
Large LLMs perform slightly better than Small LLMs in Few-shot prompting, with slightly higher accuracy score of 75% compared to 70% respectively. Large LLMs have greater in-context learning ability, so additional examples provided adjust reasoning of LLM based on patterns in examples.





Zero-shot prompting vs Few-shot prompting (Small LLMs)
Small LLMs perform significantly better when zero-shot prompting is used compared to when few-shot prompting is used. Small LLMs have smaller context window, so few-shot prompts take up valuable context space, limited contexts retrieved.

Zero-shot prompting vs Few-shot prompting (Large LLMs)
Large LLMs also perform significantly better when zero-shot prompting is used compared to when few-shot prompting is used. Few-shot prompts are more complex, and the examples presented may cause the LLMs to confabulate, compared to zero-shot prompts which are simpler.

2) Qualitative Evaluation

It was also observed that responses generated by both small and large LLMs showed a decrease in length when Few-shot prompting was used. This was attributed to the increased relevancy of responses when Few-shot prompting was used. Responses generated from Few-shot prompting also had less of a fixed structure and made use of keywords in questions to structure prompt.

7. Conclusion

Both small and large LLMs perform significantly better when Zero-shot prompting was used, compared to when Few-shot prompting was used, with higher accuracy scores being reported for the former. However, despite lower accuracy score, Few-shot prompting yields more relevant and succinct responses from LLMs, which is particularly useful for ICA Officers who do not have time to dissect main information from responses. It was also observed that there was only a marginal difference in performance between small and large LLMs for both prompt engineering techniques used.

8. Future Work

Future work can involve other types of LLMs such as OpenAI's (GPT-Turbo4 and HuggingFace models which were unable to be implemented due to cost and limited memory. This can also be linked to predictive analysis, which can identify trends in certain characteristics of suspicious personnel, allowing for early intervention.