Assignment 1
Evaluation of Large Language Models

Musab bin Jamil – 29409

The code for the evaluation is provided in a separate file below you can find my reviews of each task completed by the model.

Llama

Summarization

The model gave a summary response but it was much more lengthy than other models. It provided additional details although still relevant but they were a bit extra.

Question Answering

It Identified the correct response from sender and was able to identify that the person said customer service did not respond to the customer on time.

Named Entity Recognition (NER)

The model was able to identify the entities and output a list of everything asked for.

Translation (English to French to English)

The model was able to make the translations but it made the email much smaller. It summarized 3 paragraphs to one.

Mistral

Summarization

The response from mistral was much similar to llama it gave a response in points where each had a part of the email summarized. I would prefer a model that gives a 1-line answer as summary.

Question Answering

The answer to question sounded repetitive but it did pick up the correct details.

Named Entity Recognition (NER)

The model waw able to pick up on NERs but it separated the names by spaces rather than commas. It got dates as they were in the email.

Translation (English to French to English)

This model translated the email word for word it seems like. Since it got 3 paragraphs in French and the English translation had same wording as well.

Qwen

Summarization

The model gave a 1-line answer covering the main topic of the email. So, the summary is brief, comprehensive, captures sender's complaint and is clear.

Question Answering

Provided answers are complete, they address the question and have all the details.

Named Entity Recognition (NER)

Although my code was unable to generate this result, I am certain that the model is capable of generating results for NER similar to how other models have done so.

Translation (English to French to English)

My code for this wasn't able to generate a result for French but it gave the same email as a translation for English.

Phi

Summarization

Similar to Qwen the summary was good and concise.

Question Answering

The model ran for quite a while and generated an answer for question 1 but didn't answer question 2.

Named Entity Recognition (NER)

This part of the code gave me quite some trouble but it did not run correctly but the model is capable of generating such responses.

Translation (English to French to English)

The model was able to generate a French response. The English translation part it threw and error but the model is able to create translations.