Institute of Business Administration
Introduction to Text Analytics
Assignment 01 – LLM Response Assessment

Name: Zuha Aqib **ID: 26106**

Name of 3B Model used: "Qwen/Qwen2.5-3B-Instruct"

Name of 7B Model used: "open-thoughts/OpenThinker-7B"

Name of 14B Model used: "microsoft/phi-4"

Evaluate each model's response using the mentioned criteria (on a scale of 1-5, 1 means completely irrelevant response while 5 means perfect response):

Task	Criteria	3B Model	7B Model	14B Model	Comments (if any)
Summarization	Conciseness	5	4	4	3B one was short and to the point while 7B is lightly longer. 13b had repitition and less readabiliy
	Clarity	5	5	5	
Question Answering	Accuracy	4	5	5	
	Completeness	5	5	5	
Keyword Extraction	Completeness	3	5	Didn't do it	
	Categorization	2	5	Didn't do it	
Translation	Fidelity	5	4	4	
	Fluency	5	4	3	
	Consistency	4	4	4	
Time Taken (seconds)	Time taken to do all 4 prompts, and then convert French back to English	25s 20ms + 10s 73ms = **36s**	1m 19s 304ms + 1m 19s 720ms = **2m 40s**	1m 22s 791ms + 23s 863ms = **1m 50s**	

If you were to deploy one of these LLMs at IBA, which model would you choose, and what will be your justification?

I think the Qwen (3B) one would be best, as it is very very fast, in this generation everyone is impatient and have less attention span, and usually do works at the last minute. So we must prioritise time. Secondly, Qwen gave the most accurate and good answer out of all. In some works we require lots of things to be done, and Qwen performed all of them; I have tried multiple 13B

models and all of them skipped half the tasks and did not display, and took multiple minutes to perform. Microsoft PHI itself skipped the middle task. However Qwen did all. Even the French translation was accurate as when translated back to English, it was same.

Why would it benefit IBA? We discussed how time efficiency is crutial, another thing is that it had good text summarization, it included all points and even did a good mimic of the tone of the text, in the required amount of words. This would be beneficial of students (summarizing powerpoints and book chapters) and teachers (summarizing research papers and written material from students). It even answered questions correctly, which teachers can use by giving the LLM text and asking questions to see if the student covered all aspects. I am not sure how admin can benefit, because I did not use any numerical examples on the model to test its calculations (admin could analyze classes and gain insights), but directors/deans could use the model to summarize documents sent from other employees/admins as higher authority have less time to go over extensive things, and could just read bullets or summaries from the model.

However the Keyword Extraction was best in OpenThinker (7B), Qwen did not understand what it was and did not perform correctly.

Criteria Explanation

Summarization:

- Conciseness: Is the summary brief yet comprehensive?
- Clarity: Is the summary clear and understandable?

Question Answering:

- Accuracy: Is the answer factually correct, as per the email?
- Completeness: Is the answer complete, providing all necessary details?

Keyword Extraction:

- Completeness: Are all keywords identified?
- Categorization: Are keywords correctly categorized (person, date, product, etc.)?

Translation (English to French to English):

- Fidelity: Does the back-translation convey the original message accurately?
- Fluency: Is the grammar of the back translation correct and fluent?
- Consistency: Is the tone and terminology consistent in the back translation?