Summary Report:

From our results in the excel sheet, we can deduce that all of the LLM's have performed almost equally well at answering all of these questions, with Llama 3.1 being slightly better than the others. However, this could be as a result of the number of questions only being 5, and more insights may be deduced with a large number of questions, say perhaps 100.

Accuracy:
GPT-4 and Llama 3.1 were the most accurate, with Gemini 1.5 flash being less precise in some areas.

Completeness:
Claude 3.5 Sonnet and Llama 3.1 were the most complete with their answers, with the others often not addressing the full picture.

Relevance:
Claude 3.5 Sonnet, Llama 3.1, and Gemini 1.5 Flash were the best performing, with GPT often adding irrelevant information to its generated output.

It is clear that Llama 3.1 has outperformed the other LLMs in all the categories we tested. Claude 3.5 comes close, and the other three are not too far behind either. However, if we take cost and the LLM's size into consideration, it is indeed ironic how inspite of being free and one of the smaller LLMs, Llama 3.1 was better than larger, more expensive, and better trained models such as GPT4 and Mistral Large. This could be because Llama 3.1 is a newer model, and thus has more enhanced features.