Under

Results and Analysis:

QA

a. MCQ

b. Yes No Maybe

c. Question Answering

Summarization

Question Answering and Summarizations can be kept same. I have some changes for remaining.  
  
Below is the summary of MCQ on MedMCQa dataset:

accuracy valid\_responses total\_questions

groq\_llama-3.3-70b-versatile 0.754902 102.0 102.0

azure\_gpt-4o-mini 0.715686 102.0 102.0

groq\_mistral-saba-24b 0.637255 101.0 102.0

groq\_gemma2-9b-it 0.637255 100.0 102.0

groq\_llama-3.1-8b-instant 0.598039 102.0 102.0

Model Accuracy by Subject:

azure\_gpt-4o-mini groq\_llama-3.3-70b-versatile \

Pediatrics 1.0 0.8

Anaesthesia 0.6 0.8

Dental 0.8 0.4

Radiology 0.8 0.8

Gynaecology & Obstetrics 0.8 1.0

Microbiology 1.0 1.0

Medicine 0.6 0.6

Physiology 0.8 1.0

Pharmacology 0.6 0.4

ENT 0.6 0.8

Biochemistry 0.8 1.0

Skin 1.0 1.0

Pathology 0.8 1.0

Anatomy 0.6 0.8

Psychiatry 0.8 0.6

Surgery 0.4 0.6

Social & Preventive Medicine 0.6 0.8

Ophthalmology 0.6 0.4

Orthopaedics 0.8 0.8

Forensic Medicine 0.6 0.6

groq\_mistral-saba-24b \

Pediatrics 0.8

Anaesthesia 0.2

Dental 0.6

Radiology 0.6

Gynaecology & Obstetrics 0.8

Microbiology 1.0

Medicine 0.2

Physiology 0.6

Pharmacology 0.4

ENT 0.8

Biochemistry 1.0

Skin 0.6

Pathology 1.0

Anatomy 0.6

Psychiatry 0.4

Surgery 0.6

Social & Preventive Medicine 0.4

Ophthalmology 0.6

Orthopaedics 0.8

Forensic Medicine 1.0

groq\_llama-3.1-8b-instant groq\_gemma2-9b-it

Pediatrics 1.0 0.8

Anaesthesia 0.6 0.4

Dental 0.0 0.2

Radiology 0.4 0.6

Gynaecology & Obstetrics 0.8 0.8

Microbiology 0.8 1.0

Medicine 0.6 0.6

Physiology 0.8 0.6

Pharmacology 0.4 0.6

ENT 0.4 0.6

Biochemistry 0.8 1.0

Skin 0.4 0.8

Pathology 0.6 1.0

Anatomy 0.6 0.4

Psychiatry 0.4 0.6

Surgery 0.6 0.4

Social & Preventive Medicine 0.4 0.4

Ophthalmology 0.8 0.4

Orthopaedics 0.8 0.6

Forensic Medicine 0.8 1.0

Top 5 subjects for azure\_gpt-4o-mini:

Skin (5 questions): 100.00%

Microbiology (5 questions): 100.00%

Pediatrics (5 questions): 100.00%

Orthopaedics (5 questions): 80.00%

Biochemistry (5 questions): 80.00%

Top 5 subjects for groq\_llama-3.3-70b-versatile:

Pathology (5 questions): 100.00%

Biochemistry (5 questions): 100.00%

Skin (5 questions): 100.00%

Gynaecology & Obstetrics (5 questions): 100.00%

Physiology (5 questions): 100.00%

Top 5 subjects for groq\_mistral-saba-24b:

Pathology (5 questions): 100.00%

Biochemistry (5 questions): 100.00%

Forensic Medicine (5 questions): 100.00%

Microbiology (5 questions): 100.00%

Orthopaedics (5 questions): 80.00%

Top 5 subjects for groq\_llama-3.1-8b-instant:

Pediatrics (5 questions): 100.00%

Orthopaedics (5 questions): 80.00%

Biochemistry (5 questions): 80.00%

Physiology (5 questions): 80.00%

Forensic Medicine (5 questions): 80.00%

Top 5 subjects for groq\_gemma2-9b-it:

Biochemistry (5 questions): 100.00%

Forensic Medicine (5 questions): 100.00%

Microbiology (5 questions): 100.00%

Pathology (5 questions): 100.00%

Skin (5 questions): 80.00%

=== Medical MCQ Evaluation Summary ===

Number of questions evaluated: 102

Questions per subject: 5

Number of subjects: 21

Model Performance:

azure\_gpt-4o-mini:

Accuracy: 71.57%

Valid responses: 102/102

groq\_llama-3.3-70b-versatile:

Accuracy: 75.49%

Valid responses: 102/102

groq\_mistral-saba-24b:

Accuracy: 63.73%

Valid responses: 101/102

groq\_llama-3.1-8b-instant:

Accuracy: 59.80%

Valid responses: 102/102

groq\_gemma2-9b-it:

Accuracy: 63.73%

Valid responses: 100/102

Best performing model by accuracy: groq\_llama-3.3-70b-versatile

Best performing model by F1 score: groq\_llama-3.3-70b-versatile

Number of questions with model disagreement: 55 (53.9% of questions)  
  
Below given is the summary of MMLU dataset:  
  
**# MMLU Medical MCQ Evaluation Summary**

**## Overview**

This report presents the results of evaluating various language models on medical multiple choice questions from the MMLU dataset.

**## Dataset**

- **\*\*Number of questions evaluated\*\***: 100

- **\*\*Topic\*\***: Medical knowledge (MCQs)

**## Model Performance Summary**

**### Accuracy Metrics**

| model\_provider   | model\_name              |   accuracy |   total\_questions |   accuracy\_pct | display\_name                   |

|:-----------------|:------------------------|-----------:|------------------:|---------------:|:-------------------------------|

| azure            | gpt-4o-mini             |       0.88 |               100 |             88 | gpt-4o-mini (azure)            |

| groq             | llama-3.3-70b-versatile |       0.83 |               100 |             83 | llama-3.3-70b-versatile (groq) |

| groq             | mistral-saba-24b        |       0.78 |               100 |             78 | mistral-saba-24b (groq)        |

| groq             | gemma2-9b-it            |       0.75 |               100 |             75 | gemma2-9b-it (groq)            |

| groq             | llama-3.1-8b-instant    |       0.74 |               100 |             74 | llama-3.1-8b-instant (groq)    |

**## Key Findings**

\* The best performing model was **\*\*gpt-4o-mini\*\*** (azure) with an accuracy of 88.00%

\* The lowest performing model was **\*\*llama-3.1-8b-instant\*\*** (groq) with an accuracy of 74.00%

\* The performance gap between the top and bottom model is 14.00 percentage points

\* The average accuracy across all models was 79.60%

**## Conclusion**

There is a moderate performance gap between the models, suggesting that larger models generally perform better on medical knowledge tasks.

For Yes No Maybe answering on Pubmed Dataset we got:  
  
Model Accuracy Summary:

- azure\_gpt\_4o\_mini: 0.5800

- groq\_llama\_3.3\_70b\_versatile: 0.5600

- groq\_mistral\_saba\_24b: 0.6600

- groq\_llama\_3.1\_8b\_instant: 0.5600

- groq\_gemma2\_9b\_it: 0.6400  
  
  
I believe this happened when LLLMs overconfidently predicted Yes/No instead of maybe. As we reduce the no of questions with Maybe, Accuracy increases