MedLingua at MedArabiQ2025: Zero- and Few-Shot Prompting of Large Language Models for Arabic Medical QA

Fatimah Emad Eldin, Mumina Abukar
Cairo University, The University of South Wales
{12422024441586@pg.cu.edu.eg, 74108361@students.southwales.ac.uk}

Abstract

This paper details the system developed by team MedLingua for the MedArabiQ2025 Shared Task, specifically participating in Track 2, Sub-Task 1: Multiple Choice Question Answering. Our approach centered on evaluating the zero-shot and few-shot capabilities of various Large Language Models (LLMs) on Arabic medical questions, as fine-tuning was not permitted. We systematically tested a range of models, from general-purpose state-of-the-art LLMs like Google's Gemini 2.5 Pro to specialized medical models such as BiMediX2 and MedGemma. Our findings reveal that advanced, general-domain models significantly outperform specialized medical LLMs that are not optimized for Arabic. Our best performing system, using Gemini 2.5 Pro, achieved an accuracy of 78% in the development set and 74% on the blind test set, securing the 3rd place on the official competition leaderboard.

1 Introduction

The MedArabiQ2025 shared task addresses the critical need for robust natural language understanding systems in the Arabic medical domain (Abu Daoud et al., 2025). Our team, MedLingua, participated in Track 2, Sub-Task 1, which focuses on Multiple Choice Question Answering (MCQA). This task is vital for developing clinical decision support systems and educational tools tailored to Arabic-speaking healthcare professionals and students. The primary challenge lies in the complexity of medical language and the relative scarcity of high-quality Arabic medical datasets and models compared to English. Given the constraint that participants could not fine-tune models on the provided data, our core strategy was to leverage the in-context learning abilities of existing LLMs. We employed both zero-shot and few-shot prompting techniques to guide various models toward the correct answer.

Our key finding was the pronounced performance gap between large, multilingual general-purpose models and the available specialized medical LLMs. The former demonstrated superior understanding of the Arabic questions, while many of the latter struggled with the language or failed to adhere to the task's constraints. Our best system achieved 74% accuracy on the blind test set, demonstrating the effectiveness of modern LLMs in this zero-resource fine-tuning scenario. To ensure reproducibility and facilitate future research in Arabic medical question answering, we make all experimental code publicly available on GitHub [1].

2 Background and Related Work

Question answering in the medical domain is a well-established research area (Pampari et al., 2018). However, most work, including the development of specialized models like Palmyra-Med Writer Engineering team (2024) and Med-PaLM (Singhal et al., 2023), has been overwhelmingly focused on English. While models like BiMediX2 (Mullappilly et al., 2024) have emerged to address the bilingual (Arabic-English) need, the field is still nascent. The MedArabiQ benchmark (Abu Daoud et al., 2025) is a crucial step in spurring research in this area. Our work contributes by providing a comprehensive evaluation of how current SOTA generalist and specialist LLMs perform on this new Arabic benchmark without task-specific fine-tuning.

3 Data

3.1 Shared Task Data

The MedArabiQ2025 MCQA sub-task is framed as a classification problem where the system receives a question in Arabic and must return the single

[1] https://github.com/astral-fate/AraHealthQA-2025-MedArabiQA

Proceedings of The Third Arabic Natural Language Processing Conference, pages 126–139

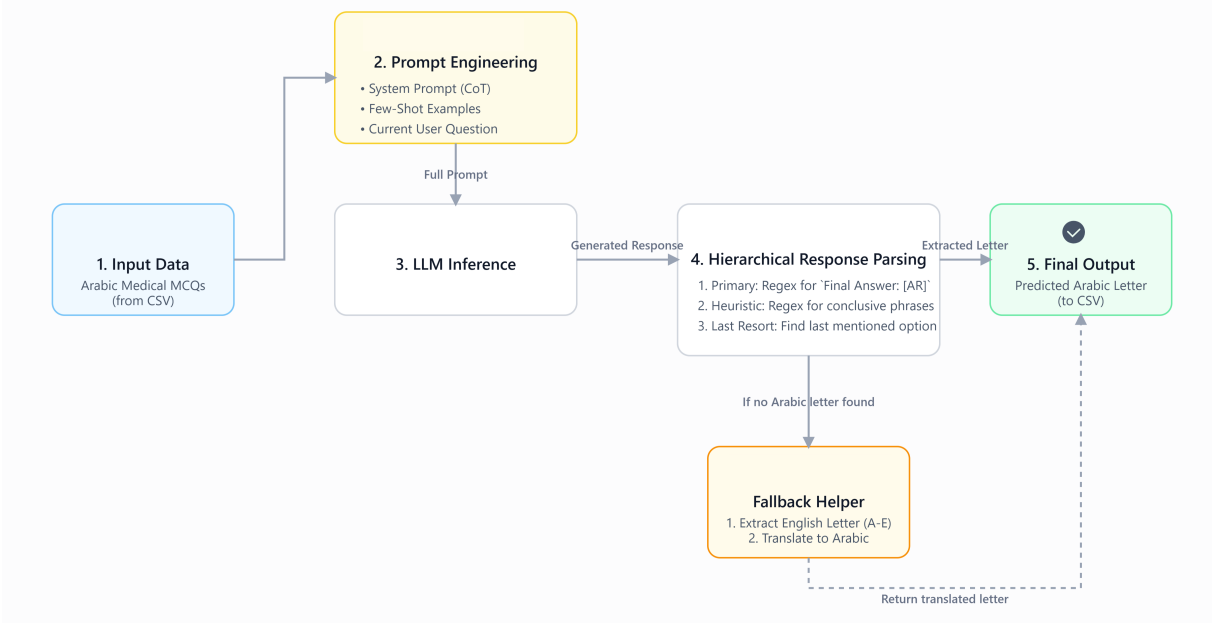


Figure 1: Overview of the system architecture for Arabic Medical QA.

letter corresponding to the correct answer from a list of choices (Alhuzali et al., 2025).

The organizers provided three distinct datasets for model development and validation, each containing 100 questions. The questions were sourced from medical exams and categorized into 12 medical specialties.

3.2 Validation Dataset

The validation data was split into three types, which we used for iterative testing and model selection:

- **Multiple Choice Questions (MCQ):** A standard set of multiple-choice questions.
- **Multiple Choice Questions with Bias (MCQ w/ Bias):** Questions designed with misleading phrasing to test model robustness.
- **Fill-in-the-Blank (FITB) with Choices:** Questions presented in a fill-in-the-blank format.

3.3 Test Dataset

The final evaluation was performed on a blind test set containing 100 questions. This dataset was a combination of all three question types from the validation set and was used to determine the final competition rankings.

4 Methodology

Our approach for Arabic medical question answering (QA) leverages in-context learning through various Large Language Models (LLMs), given the constraint against fine-tuning. The system architecture, designed to process Arabic medical multiple-choice questions (MCQs), is detailed in Figure 1.

4.1 Prompt Engineering and System Architecture

Our methodology centered on carefully structured prompt engineering to guide LLMs in a zero-shot or few-shot setting. The architecture can be broken down into five key stages:

1. **Input Data**: The process begins with loading Arabic medical MCQs from a CSV file.
2. **Prompt Engineering**: A full prompt is dynamically constructed by combining a system prompt, few-shot examples (if applicable), and the current question.
3. **LLM Inference**: The prompt is sent to an LLM for processing.
4. **Hierarchical Response Parsing**: The model's response is parsed using a multi-step process to extract the final answer.
5. **Final Output**: The extracted Arabic letter is saved to an output CSV for evaluation.

4.2 Chain-of-Thought (CoT) and Few-Shot Prompting

A key component of our strategy was the use of Chain-of-Thought (CoT) prompting.

Prompt Type	Prompt Structure
Few-Shot(e.g., MedGemma, Qwen)	`SYSTEM_PROMPT + FEW_SHOT_EXAMPLES + USER_QUESTION`
Zero-Shot (e.g., BioMistral)	`SYSTEM_PROMPT + USER_QUESTION`

Table 1: Comparison of prompt structures for few-shot and zero-shot learning.

We instructed models to first perform a step-by-step reasoning process within a `<thinking>` block before providing the final answer. An example of the Chain-of-Thought (CoT) prompt structure used for few-shot learning is provided in Appendix B (Table 4).

4.3 Zero-Shot vs. Few-Shot Strategies

Our approach involved testing both few-shot and zero-shot prompting strategies to determine the most effective method for each model. The fundamental difference in these approaches lies in the inclusion of examples within the prompt, as illustrated in Table 1.

4.3.1 The Case of BioMistral: When Few-Shots Fail

A notable example was **BioMistral** (Labrak et al., 2024). When provided with few-shot examples in Arabic, its output became nonsensical, generating repetitive, meaningless Arabic words. However, when we switched to a **zero-shot** approach (removing the examples), its behavior changed dramatically. Although it did not produce reasoning in Arabic, it performed the reasoning correctly in English and concluded with the correct Final Answer: format. This highlights that for some models, few-shot examples can confuse rather than guide.

4.4 Model Selection and Implementation

We experimented with two main categories of models:

1. **General-Purpose LLMs:** Models like Google's Gemini 2.5 Pro, Mixtral, Llama 3, and Qwen (Qwen Team, 2025), accessed via APIs (DeepMind AI Studio [2], NVIDIA NIM inference microservices API[3], Groq [4].)

2. **Specialized Medical LLMs:** Models like BiMediX2 (Mullappilly et al., 2024), MedGemma (Sellergren et al., 2025), BioMistral (Labrak et al., 2024), OpenBioLLM (Pal and Sankarasubbu, 2024), and Palmyra-Med Writer Engineering team (2024),

General-purpose LLMs (Gemini, Qwen, etc.) were accessed via APIs from DeepMind, NVIDIA, and Groq. For specialized models, MedGemma, BioMistral, and OpenBioLLM were accessed via Hugging Face; Palmyra-Med via the NVIDIA NIM API; and BiMediX2 was run locally on a Google Colab Pro+ A100 GPU.

5 Results

Our experiments revealed a striking performance gap, with large, general-purpose LLMs consistently outperforming specialized medical models on Arabic medical question answering. Our final submission, using Gemini 2.5 Pro, achieved **74% accuracy on the blind test set**, securing the 3rd place on the official competition leaderboard. Table 3 shows the performance of all 11 models we evaluated on the final blind test set.

For a more granular error analysis of the 100-question blind test set, the manual categorization of each question into 12 medical specialties was performed by co-author Dr. Mumina Abukar, MD, MScPH. This allowed us to precisely identify model weaknesses. Analysis of the categorized test set revealed that certain medical domains were universally more difficult for the models. The detailed error distribution by medical category and the accuracy versus execution time analysis are presented in Appendix A (see Figures 2a and 2b).

The primary sources of errors remained consistent with our development set findings: incorrect medical reasoning and output formatting failures.

5.1 Error Distribution on the Test Set

Table 2 details the error counts for the five highest-scoring models across the five most challenging medical categories, identified by the highest total number of errors across all tested models. Physiology emerged as the most difficult category, where even top models struggled. Notably, Gemini 2.5 Pro demonstrated the most robust performance, registering the lowest error count in three of the five

[2] https://aistudio.google.com/
[3] https://build.nvidia.com/models/
[4] https://console.groq.com/

Category	Gemini	MedGemma	Colosseum	Palmyra-Med	Llama3 70B
Physiology	6	13	12	11	14
Ophthalmology	4	7	5	9	9
Oncology	4	6	7	6	5
Biochemistry	4	4	5	5	6
Neurosurgery	1	4	6	6	7

Table 2: Focused Error Analysis: Error counts for the top 5 performing models in the 5 most error-prone medical categories on the blind test set.

Model	Test Accuracy
Gemini 2.5 Pro	**74%**
Qwen	67%
MedGemma	53%
Colosseum	51%
Palmyra-Med	49%
Llama3 70B	45%
BiMediX2	37%
Mixtral	21%
OpenBioLLM	21%
Biomistral	19%
DeepSeek	17%

Table 3: Performance of all evaluated models on the Blind Test set. Our final submission used Gemini 2.5 Pro.

most challenging categories: Neurosurgery (1 error), and tying for the lowest in Oncology (4 errors) and Biochemistry (4 errors). This highlights its strong reasoning capabilities even in complex domains.

6 Discussion

The pronounced performance gap between large, generalist LLMs and their specialized medical counterparts on the blind test set is the key finding of this work. The superior performance of models like Gemini 2.5 Pro (74%) and Qwen (67%), can be attributed to their advanced multilingual capabilities and vast general knowledge. These features appear to compensate for the lack of specific medical fine-tuning, especially when handling nuanced Arabic medical questions.

Our detailed error analysis of the test set reinforces this conclusion. The annotation of the test set questions into 12 medical specialties was manually performed by co-author Dr. Mumina Abukar, MD, MScPH, leveraging her expertise in the medical field. During this process, it became apparent that some questions, particularly those related to study design and data collection, did not fit precisely within the original 12 medical categories in Appendix I (Table 13) shows examples of such questions, which were categorized as "Physiology" in the original dataset but are better described as "Research Methodology". This potential mismatch could impact the fine-grained error analysis; however, for consistency with the original dataset structure, we adhered to the provided 12 categories for our evaluation.

The specialized models were largely hindered by a "language barrier." For instance, MedGemma's relatively high error rate in Physiology (13 errors, as shown in Table 2) suggests its specialized training did not effectively transfer to the Arabic context. This necessitated a translation-based approach for English-centric models like Palmyra-Med, which introduces potential information loss and likely limited their performance. BiMediX2, the only dedicated bilingual model tested, showed promise but was not competitive with the scale and reasoning power of top-tier generalist models on this task.

This outcome underscores a critical consideration for applying LLMs in specialized, non-English domains: strong foundational language understanding is a prerequisite for effective domain-specific reasoning. The test set results clearly show that Gemini's robust grasp of Arabic allowed it to apply its reasoning capabilities more effectively than models that were technically more specialized in medicine but weaker in the target language.

7 Conclusion

This work evaluated zero-shot and few-shot prompting strategies for Arabic medical question answering using general-purpose and specialized medical large language models. Our best-performing system achieved 74% accuracy on the MedArabiQ2025 blind test set using Gemini 2.5 Pro, securing the 3rd place on the official competition leaderboard.

Results demonstrate that advanced general-purpose models significantly outperformed specialized medical LLMs due to superior multilingual capabilities compensating for lack of domain-specific training.

Key limitations include language barriers hindering specialized models and potential dataset categorization inconsistencies. Future research should prioritize developing medical LLMs specifically trained on high-quality, large-scale Arabic medical corpora to bridge the identified performance gap between general and specialized models.

Acknowledgments

We thank the organizers of the MedArabiQ2025 shared task at New York University Abu Dhabi for creating this valuable benchmark and facilitating research in Arabic medical NLP.

References

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E. Shamout. 2025. MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks. *arXiv preprint arXiv:2505.03427*.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, et al. 2025. AraHealthQA 2025 Shared Task Description Paper. *arXiv preprint arXiv:2508.20047*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv preprint arXiv:2402.10373*.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX2: Bio-Medical EXpert LMM for Diverse Medical Modalities. *arXiv preprint arXiv:2412.07769*.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. Hugging Face repository. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Andrew Sellergren et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.

Writer Engineering team. June 2024. Palmyra-Med-70b: A powerful LLM designed for healthcare. *Writer Engineering Blog*.

A Test Set Analysis: Error Distribution and Performance

B Example Prompt Structure

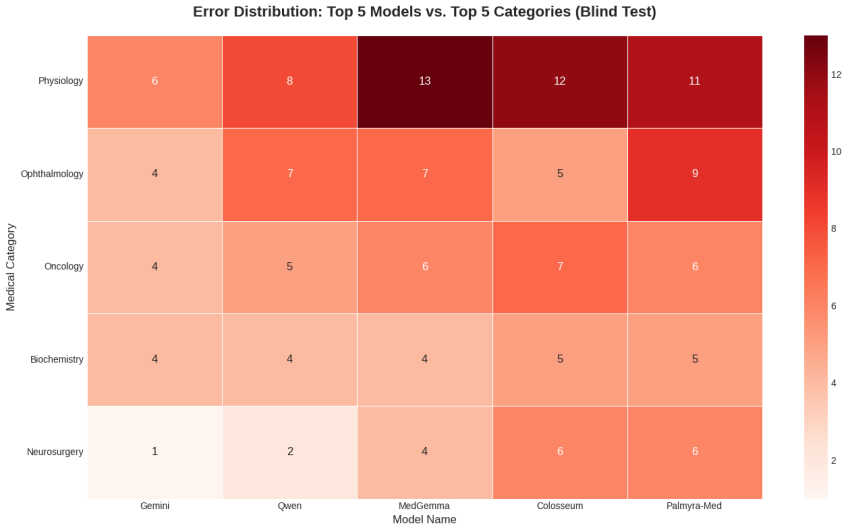
Table 4 illustrates the detailed Chain-of-Thought (CoT) prompt structure that was a key component of our methodology for the few-shot experiments, as referenced in Section B.

C Full Error Distribution on the Blind Test Set

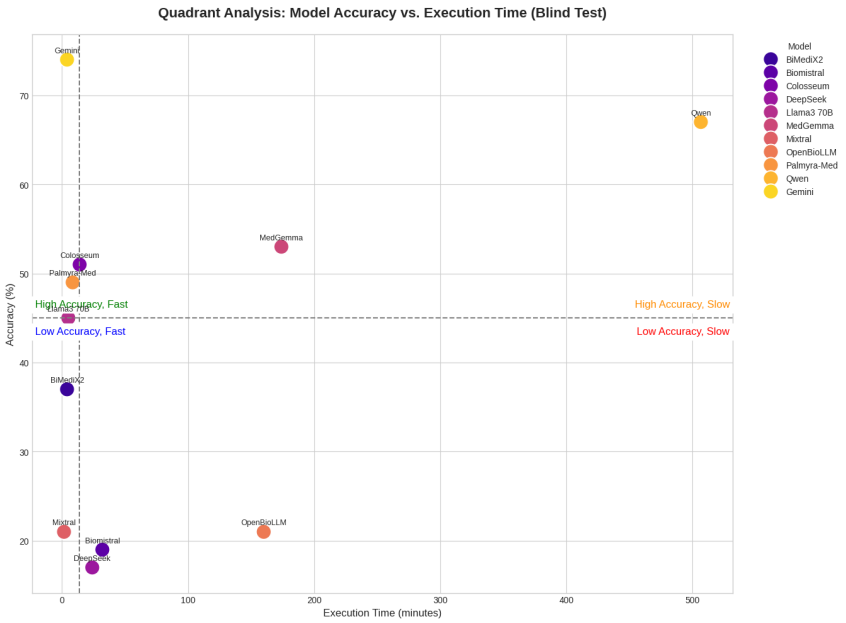
Table 5 provides a comprehensive breakdown of the errors made by each of the 11 models evaluated on the blind test set. The questions were manually classified into 12 distinct medical specialties to facilitate this granular analysis.

D Summary of Model Performance on Development Datasets

This appendix provides a consolidated view of the performance of all evaluated models across the three distinct development datasets: Fill-in-the-Blank (FITB), standard Multiple Choice Question (MCQ), and MCQ with Bias. Table 6 summarizes the development accuracy for each model, highlighting the variance in performance depending on the question format and the presence of intentionally misleading phrasing.



(a) Errors by medical category.



(b) Accuracy vs. Execution Time.

Figure 2: A comparison of error distribution and performance on the blind test set.

E Analysis of the Fill-in-the-Blank (FITB) Task

This appendix presents a detailed Exploratory Data Analysis (EDA) of model performance on the "Fill-in-the-Blank with Choices" dataset. We analyze the overall accuracy, error distribution across medical specialties, and the relationship between model performance and inference time.

E.1 Model Performance Overview

The experiments revealed a wide range of performance. A clear hierarchy emerged, with a distinct group of high-performing models separating from the rest. **Gemini** achieved the highest accuracy at 84.0%, establishing itself as the top performer on this task. It was followed by a competitive tier including **MedGemma** (81.0%), **DeepSeek 70B** (78.0%), and **Colosseum** (75.0%). Conversely, several specialized models like **BioMistral** (15.0%) and **OpenBioLLM** (34.0%) struggled significantly. Table 7 summarizes the final accuracy and execution times for each model.

E.2 Error Analysis by Medical Category

To understand model weaknesses, we analyzed the distribution of errors across medical categories. The results show that certain domains were universally more difficult. The five categories with

Component	Example Content
System Prompt	You are an expert medical professional... Your task is to solve a multiple-choice question in Arabic. First, you will engage in a step-by-step thinking process in a `<thinking>` block... Second, after your reasoning, you MUST provide the final answer on a new line in the format: `Final Answer: [The single Arabic letter]`
User Question Example	املأ الفراغات... في حالة الانصباب الجنبي...

Ideal Assistant Response (with CoT)	`<thinking>`
	١. تحليل السؤال: يسأل السؤال عن دلالة انخفاض أو غياب الرجفان اللمسي... ٢. تقييم الخيارات: أ. تراكم السوائل؛ عزل الصوت... ٣. الاستنتاج: الخيار الأكثر دقة هو أن تراكم السوائل هو ما يسبب عزل الصوت...
	`</thinking>`
	Final Answer:
	أ

Table 4: Illustration of the Chain-of-Thought (CoT) prompt structure used in our few-shot experiments.

the highest total error counts were **OBGYN**, **Pulmonology**, **Cardiovascular System**, **Gastroenterology**, and **Neurology**. This suggests the questions in these fields may contain more complex terminology or require more nuanced clinical reasoning. Table 8 details the error counts for the top-performing models in these challenging categories.

E.3 Accuracy vs. Execution Time Analysis

The relationship between inference time and accuracy provides critical insights into model efficiency, as illustrated in the quadrant analysis in Figure 3c. We observe distinct performance archetypes:

1. **High Accuracy, Fast**: **Gemini** is the clear standout, occupying the top-left quadrant with the highest accuracy (84%) and a fast execution time. **DeepSeek 70B** (78%), **Colosseum** (75%) and **Palmyra-Med** (66%) also demonstrate strong efficiency.

2. **High Accuracy, Slow**: **MedGemma** resides in this category, achieving a high accuracy of 81% but requiring the longest execution time.

3. **Low Accuracy, Slow**: **BioMistral** is a notable example here, combining the lowest accuracy (15%) with a long execution time.

This analysis indicates that while more processing time can be beneficial, model architecture and optimization are paramount for achieving both speed and accuracy.

F Analysis of the Multiple Choice w/ Bias Task

This appendix presents a detailed Exploratory Data Analysis (EDA) of model performance on the "Multiple Choice with Bias" dataset. The objective is to identify which models were most resilient to the introduced bias and to pinpoint the medical categories where models struggled the most.

F.1 Model Performance Overview

The introduction of biased phrasing created a clear performance hierarchy among the models. Gemini 2.5 Pro demonstrated exceptional resilience to bias, achieving a top score of 75.0% and clearly separating itself from the other models. It was followed by Qwen (CoT), which also performed robustly with an accuracy of 68.0%. A competitive middle

Category	Gemini	Qwen	MedGemma	Colosseum	Palmyra-Med	Llama3 70B	BiMediX2	Mixtral	OpenBioLLM	Biomistral	DeepSeek
Biochemistry	4	4	4	5	5	6	9	10	9	8	10
Embryology	0	1	1	2	2	2	2	2	1	2	1
Histology	2	5	3	4	5	4	5	3	5	2	5
Microbiology	2	2	2	3	3	3	4	4	4	2	2
Neurosurgery	1	2	4	6	6	7	6	9	10	9	10
OBGYN	1	1	1	1	1	1	1	3	2	1	1
Oncology	4	5	6	7	6	5	5	11	9	10	9
Ophthalmology	4	7	7	5	9	9	11	10	11	10	12
Pediatrics	1	1	1	1	0	1	1	1	1	1	1
Pharmacology	0	0	0	2	2	2	4	3	4	4	4
Physiology	6	8	13	12	11	14	14	22	21	20	22
Pulmonology	1	1	1	1	1	1	1	1	2	2	1

Table 5: Full error distribution for all models across 12 medical categories on the 100-question blind test set.

Model's Name	Fill in the Blank (Dev Acc)	Multiple Choice Question (Dev Acc)	Multiple Choice w/ Bias (Dev Acc)
Gemini 2.5 Pro	84%	78%	75%
qwen/qwen2-32b	83%	70%	68%
google/medgemma-27b-it	81%	55%	53%
deepseek-r1-distill-llama-70b (CoT)	78%	62%	53%
colosseum_355b_instruct_16k	75%	50%	45%
llama-3.3-70b-versatile	66%	57%	40%
palmyra-med-70b / 32k	66%	55%	35%
BiMediX2	52%	25%	31%
mixtral-8x22b-instruct-v2	40%	30%	19%
OpenBioLLM	34%	24%	18%
BioMistral	15%	19%	23%

Table 6: Comprehensive development accuracy results across the three development datasets.

Model	Accuracy (%)	Total Errors	Time (mins)
Gemini 2.5 Pro	84.00	16	50.00
MedGemma	81.00	19	176.07
DeepSeek 70B	78.00	22	25.00
Colosseum	75.00	25	14.72
Llama3 70B	69.00	31	27.20
Llama3 70B	66.00	34	27.33
Palmyra-Med	66.00	34	13.95
BiMediX2	52.00	48	10.07
Mixtral	40.00	60	15.90
OpenBioLLM	34.00	66	10.78
BioMistral	15.00	85	47.77

Table 7: Final performance summary for the Fill-in-the-Blank task.

model_name Category	Gemini 2.5 Pro	MedGemma	DeepSeek 70B	Colosseum	Llama3 70B
OBGYN	2	4	4	5	10
Pulmonology	4	5	5	5	9
Cardiovascular System	3	1	4	2	9
Gastroenterology	2	1	2	4	8
Neurology	1	3	1	2	7

Table 8: Error counts for top models in the five most challenging categories on the FITB task.

tier emerged, led by DeepSeek 70B (Groq) and MedGemma (Local), which tied at 53.0%.

F.2 Error Analysis by Medical Category

The five categories with the highest total error counts were **Embryology**, **Histology**, **Physiology**, **Biochemistry**, and **Microbiology**. This suggests that questions in these foundational science fields may be harder to answer correctly when potentially misleading information is present. The heatmap in Figure 4b shows that Gemini 2.5 Pro had the fewest errors in four of these five most difficult categories.

F.3 Accuracy vs. Execution Time Analysis

The quadrant analysis in Figure 4c highlights significant differences in efficiency. Gemini 2.5 Pro is the clear standout, occupying the "High Accuracy, Fast" quadrant and demonstrating the best balance of speed and performance. Qwen (CoT) falls into the "High Accuracy, Slow" category, delivering strong results but at a significant time cost. The remaining models form a cluster of lower-accuracy options, with DeepSeek 70B (Groq) offering the best performance among the faster, less accurate models.

Model	Accuracy (%)	Total Errors	Time (mins)
Gemini 2.5 Pro	75.00	25	4.00
Qwen (CoT)	68.00	32	81.68
DeepSeek 70B (Groq)	53.00	47	15.82
MedGemma (Local)	53.00	47	180.00
Colosseum	45.00	55	13.60
Llama3 70B (CoT)	40.00	60	20.25
Palmyra-Med	35.00	65	10.77
BiMediX2 (vLLM)	31.00	69	0.53
BioMistral (Fallback)	23.00	77	41.75
Mixtral	19.00	81	14.47
OpenBioLLM 8B (Local)	18.00	82	10.37

Table 9: Final performance summary for the MCQ with Bias task, based on the updated data.

model_name Category	Gemini 2.5 Pro	Qwen (CoT)	DeepSeek 70B (Groq)	MedGemma (Local)	Colosseum
Embryology	2	5	8	7	8
Histology	1	5	8	6	9
Physiology	3	6	7	9	9
Biochemistry	3	3	4	6	7
Microbiology	1	2	3	3	1

Table 10: Updated error counts for the new top 5 models in the five most challenging categories on the biased dataset.

G Analysis of the Multiple Choice Question (MCQ) Task

This appendix provides a detailed EDA of model performance on the standard "Multiple Choice Question" dataset. We examine the overall accuracy, error distribution, and the trade-offs between accuracy and processing time.

G.1 Model Performance Overview

The standard MCQ task revealed a clear performance hierarchy. Gemini 2.5 Pro established it-

self as the top-performing model with an impressive accuracy of 78%. It was followed by a tier of other strong models including Qwen (70%), DeepSeek (62%), Llama3 70B (57%), and both Palmyra-Med (55%) and MedGemma (55%). In contrast, some specialized models like Biomistral (19%) and OpenBioLLM (24%) struggled significantly.

G.2 Error Analysis by Medical Category

Some medical specialties were consistently more challenging for all models. The five categories accumulating the most errors were **Physiology**, **Histology**, **Embryology**, **Biochemistry**, and **Microbiology**. This indicates that questions in these foundational medical sciences likely require more complex reasoning or contain more specialized terminology. The error distribution for the top-performing models in these categories is detailed in Table 12.

G.3 Accuracy vs. Execution Time Analysis

The quadrant analysis of accuracy versus execution time in Figure 5c reveals four distinct performance profiles:

1. **High Accuracy / Fast**: This quadrant is led by the top performer, **Gemini**. Other strong models like **Qwen**, **DeepSeek**, **Llama3 70B**, and **Palmyra-Med** also fit here, offering high accuracy with efficient processing times.

2. **High Accuracy / Slow**: **MedGemma** stands alone in this category, achieving a respectable accuracy of 55% but requiring significantly more computational time (over 160 minutes).

3. **Low Accuracy / Fast**: Models like **Mixtral**, **BiMediX2**, and **OpenBioLLM** delivered results quickly but with lower accuracy scores.

4. **Low Accuracy / Slow**: **Biomistral** was the least efficient, combining low accuracy with a relatively slow execution time.

H Challenges in Manual Test Set Annotation

As mentioned in the Discussion, the manual categorization of the blind test set revealed that some questions did not align well with the provided 12 medical specialty categories. Table 13 lists five questions originally classified as "Physiology" that co-author Dr. Mumina Abukar, MD, MScPH, leveraging her expertise in the medical field, identified as belonging to "Research Methodology." This highlights a potential area for refinement in future iterations of the benchmark to ensure that the categories accurately reflect the question content, thereby improving the validity of category-based error analyses.

Model	Accuracy (%)	Total Errors	Time (mins)
Gemini 2.5 Pro	78.00	22	5.00
Qwen	70.00	30	40.00
DeepSeek	62.00	38	24.00
Llama3 70B	57.00	43	18.00
MedGemma	55.00	45	165.00
Palmyra-Med	55.00	45	8.00
Colosseum	50.00	50	24.00
Mixtral	30.00	70	2.00
BiMediX2	25.00	75	4.00
OpenBioLLM	24.00	76	12.00
Biomistral	19.00	81	33.00

Table 11: Final performance summary for the MCQ task. Total errors are based on a dataset size of 100 questions.

Category	Gemini	Qwen	DeepSeek	Llama3 70B	Palmyra-Med
Physiology	4	5	7	8	7
Histology	1	4	3	8	5
Embryology	1	2	8	8	7
Biochemistry	3	3	3	5	6
Microbiology	1	4	3	2	6

Table 12: Error counts for the top 5 models in the five most challenging categories on the MCQ task.

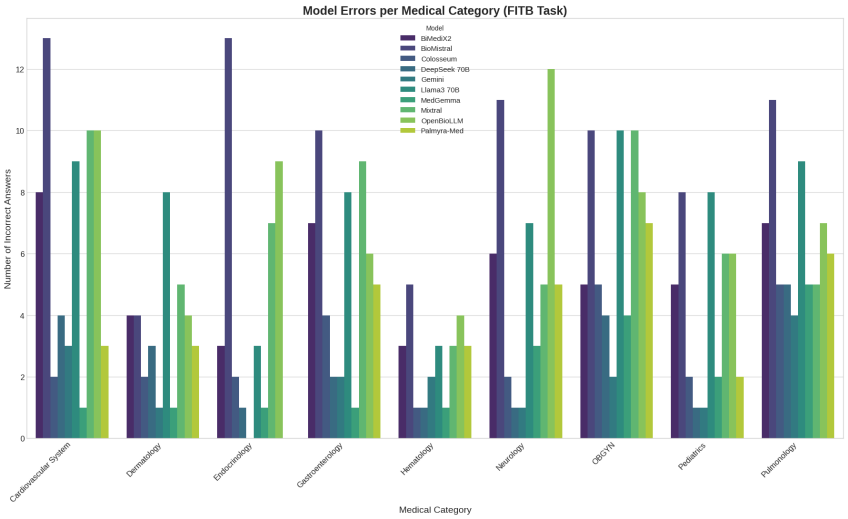
co-author Dr. Mumina Abukar, MD, MScPH, leveraging her expertise in the medical field, identified as belonging to "Research Methodology." This highlights a potential area for refinement in future iterations of the benchmark to ensure that the categories accurately reflect the question content, thereby improving the validity of category-based error analyses.

I Challenges in Manual Test Set Annotation

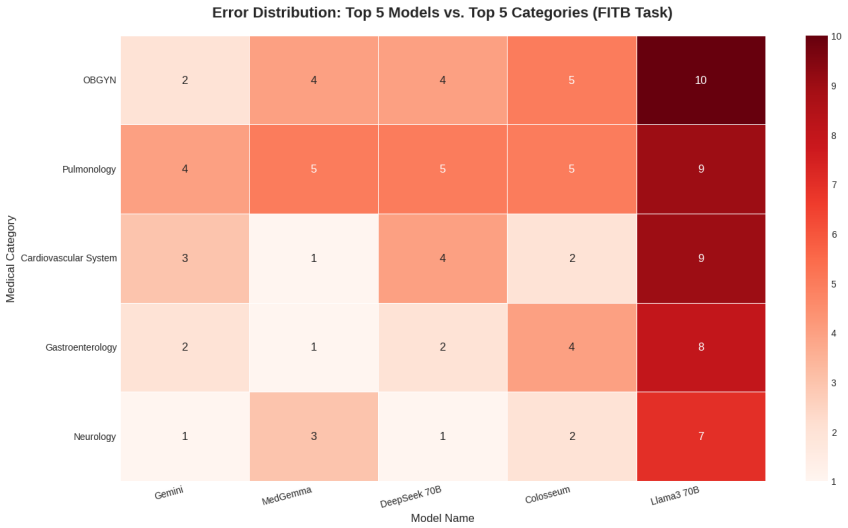
As mentioned in the Discussion, the manual categorization of the blind test set revealed that some questions did not align well with the provided 12 medical specialty categories. Table 13 lists five questions originally classified as "Physiology" that co-author Dr. Mumina Abukar, MD, MScPH, leveraging her expertise in the medical field, identified as belonging to "Research Methodology." This highlights a potential area for refinement in future iterations of the benchmark to ensure that the categories accurately reflect the question content, thereby improving the validity of category-based error analyses.

ID	Question	Original Category	Expert's Proposed Category
34	فيما يتعلق بالعينة العشوائية البسيطة، يُعد ..... ليس من خصائصها. أ. أبسط أنواع العينات ب. يتم اختيار الأفراد بإجراء القرعة ج. قيام طبيب بإجراء دراسة على مرضى مراجعين له مثال عليها د. تستخدم في حالة تجانس المجتمع	Physiology	Research Methodology
49	أساليب جمع البيانات تتضمن: أ. الاتصالات الهاتفية ب. المواقع الاجتماعية ج. استخدام الإنترنت د. كل ما سبق صحيح	Physiology	Research Methodology
54	من أساليب العينة العشوائية البسيطة . أ. اجراء قرعة عندما يكون حجم العينة صغيرا ب. استخدام جداول الأرقام العشوائية ج. عندما يكون حجم العينة كبيرة د. ا+ـ	Physiology	Research Methodology
77	من مزايا طريقة المواجهة أو المقابلة الشخصية ما عدا: أ. الحصول على إجابات دقيقة ومراقبة ردود ب. ارتفاع نسبة المستجيبين ج. كلفتها عالية د. تستخدم في المجتمعات التي ترتفع فيها نسبة الأمية	Physiology	Research Methodology
85	ال ..... هي من أقدم الطرق المستخدمة لجمع البيانات لمراقبة الكثير من الظواهر لا سيما التي يصعب السؤال عنها:	Physiology	Research Methodology

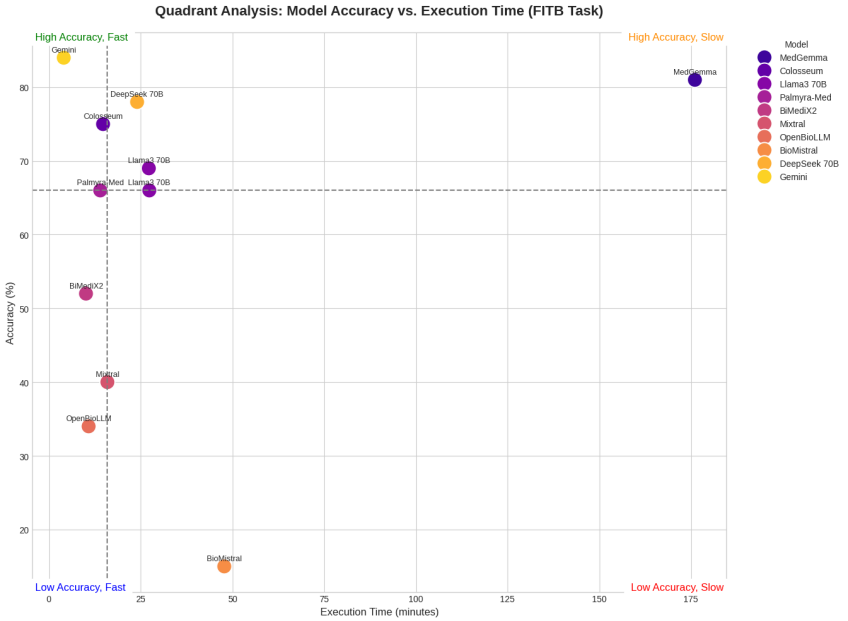
Table 13: Examples of questions from the blind test set with proposed category corrections. These questions were originally categorized under Physiology.



(a) Errors by medical category.

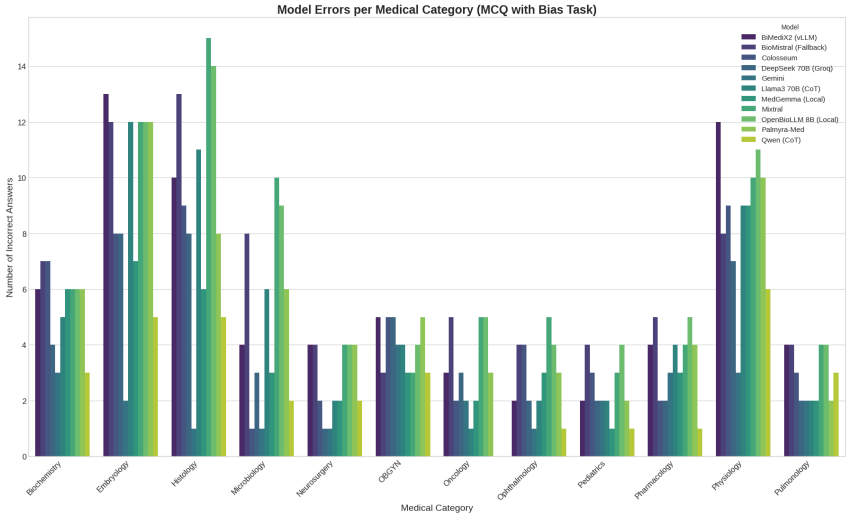


(b) Error heatmap for top models.

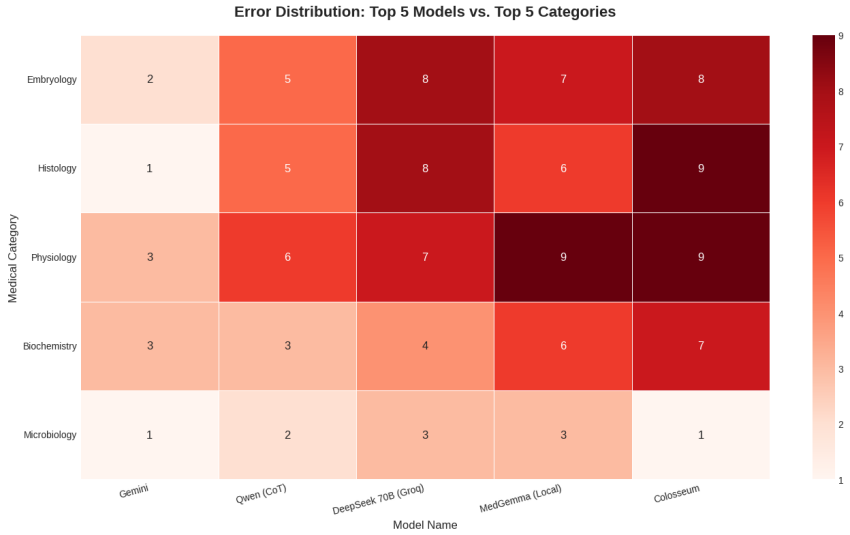


(c) Accuracy vs. Execution Time.

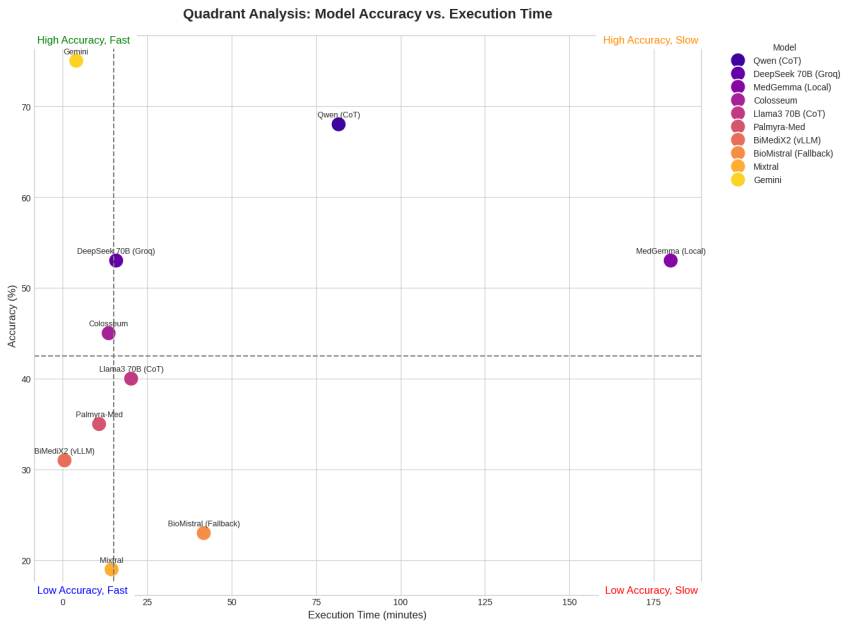
Figure 3: Detailed performance analysis for the Fill-in-the-Blank (FITB) task.



(a) Errors by medical category, based on the updated model performance.

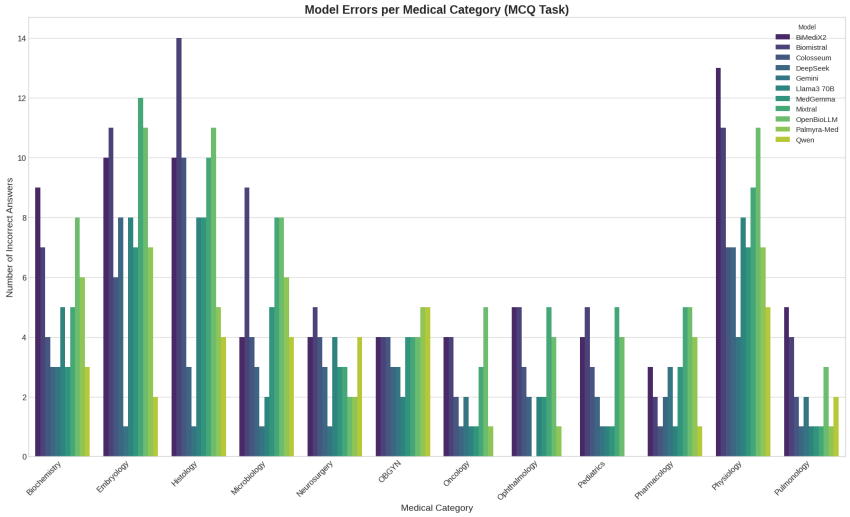


(b) Error heatmap for the new top 5 models.

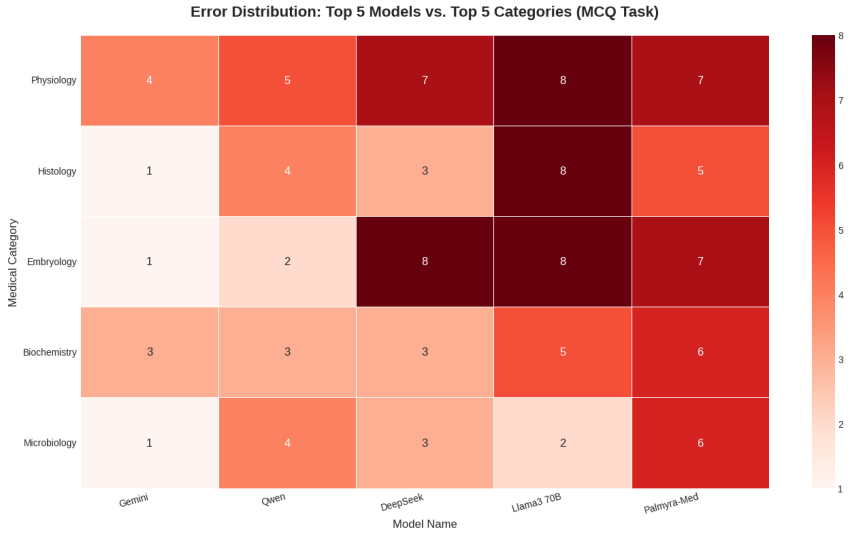


(c) Accuracy vs. Execution Time, including Gemini.

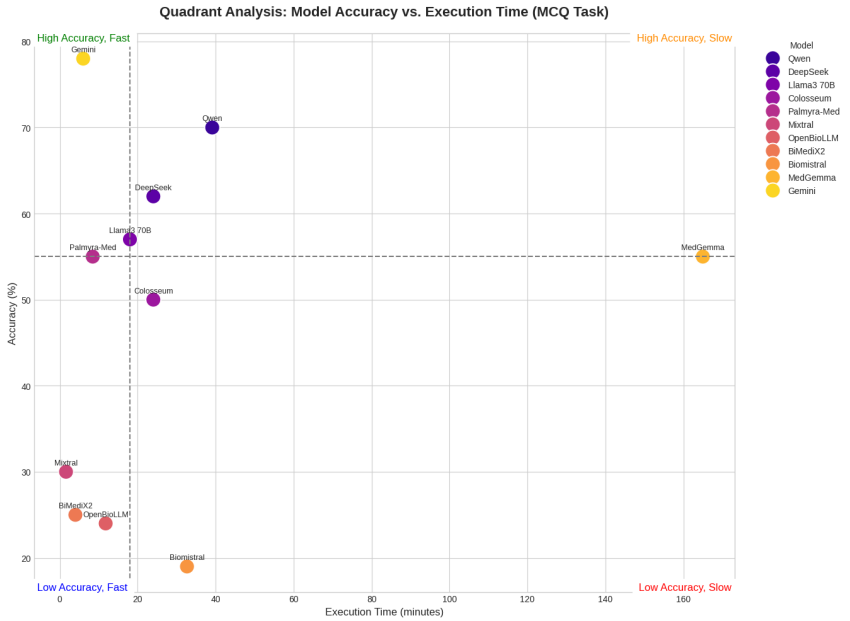
Figure 4: Detailed performance analysis for the Multiple Choice with Bias (MCQ w/ Bias) task using the latest data.



(a) Errors by medical category.



(b) Error heatmap for top models.



(c) Accuracy vs. Execution Time.

Figure 5: Detailed performance analysis for the standard Multiple Choice Question (MCQ) task.