Reliability in AI-Assisted Critical Care: Assessing Large Language Model Robustness and Instruction Following for Cardiac Arrest Identification

Anonymous Author(s) **Ugurcan Vurgun** Sy Hwang Sunil Thomas
Affiliation
Address
email
Ashley Batugo Ana Acevedo Aarthi Kaviyarasu Benjamin S. Abella
Oscar J.L. Mitchell Danielle L. Mowery

Abstract

This study systematically evaluates the performance, robustness, and instruction-following capabilities of large language models (LLMs) in identifying in-hospital cardiac arrest (IHCA) events. We assessed 51 open-source LLMs—comprising 36 general-purpose models and 15 medical-specific models—against GPT-4o, serving as a benchmark. Our analysis focused on model accuracy, robustness, and adherence to clinical instructions, with robustness measured using confidence intervals derived from non-parametric bootstrapping across several runs. While GPT-4o set a high standard with consistent performance across metrics, several open-source models demonstrated competitive results (e.g., Mistral-Nemo-Instruct-2407: F1: 0.84±0.05, Balanced Accuracy: 0.84±0.04), albeit with greater variability. Medical-specific models showed strong recall, but often exhibited wider confidence intervals, indicating potential challenges in maintaining consistent performance. Instruction-following evaluation revealed that some general-purpose models excelled in adhering to clinical guidelines (e.g., unsloth/Meta-Llama-3.1-8B-Instruct: 99.0%), while certain medical models struggled with consistency. Our findings emphasize the potential of LLMs in critical care settings, highlighting the need to balance accuracy and instruction-following capabilities for reliable clinical deployment.

1 Introduction

1.1 Background on IHCA and importance of accurate detection

In-hospital cardiac arrest (IHCA) is a leading cause of death and disability, with almost 300,000 patients experiencing IHCA annually in the United States alone [1, 2]. While short- and long-term survival after IHCA have both improved gradually over the past several decades, overall survival remains low, with 20-30% of patients surviving to hospital discharge [3, 4, 5]. Patients who do survive frequently experience significant physical disability and neurological injury [3, 6]. Multiple interventions have been shown to improve outcomes after IHCA, and high-quality post-arrest care, including coronary angiography, targeted temperature management, expert neuroprognostication, and the avoidance of premature withdrawal of life-sustaining therapies, have been associated with reductions in both morbidity and mortality after IHCA [7, 8]. However, given the complex, multi-disciplinary, and time-sensitive nature of post-arrest care, adherence to post-arrest care guidelines is low, and variable implementation likely contributes to wide inter-hospital variation in survival

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

following IHCA [9, 10, 11, 12]. Unsurprisingly, quality improvement (QI) aimed at optimizing post-arrest care has been identified as a key intervention to improve survival from IHCA [1].

While individual care teams can readily identify IHCA at the bedside, notification of QI groups or post-arrest care teams after a patient has been successfully resuscitated relies on human processes that often fail, especially for IHCA events occurring at night or on weekends. Traditional retrospective methods used to identify IHCA patients include administrative billing codes, such as the International Classification of Diseases-9 (ICD-9), and/or the use of manually assembled QI or research registries. These methods are limited for use in patient care and QI efforts as there are often delays and inaccuracies in coding and documentation [13, 14]. For example, the use of ICD-9 codes to identify cardiac arrest patients has been found to lack both sensitivity and specificity, with one large registry study demonstrating that over 36% of IHCA events were not identified using a combination of billing and procedure codes [13, 14]. Inaccurate or delayed identification of patients in the immediate post-IHCA period can further hamper the implementation of efforts designed to improve the quality of post-IHCA care and IHCA patient outcomes. As such, the accurate identification of patients after IHCA represents a modifiable factor to improve QI processes and reduce morbidity and mortality after IHCA.

1.2 Overview of LLMs and their potential in healthcare

Large Language Models (LLMs) have demonstrated significant potential in healthcare, particularly in processing and understanding complex medical texts [15]. Recent advancements have led to LLMs capable of performing a wide range of tasks, including medical diagnosis, treatment recommendation, and clinical documentation analysis [16, 17]. In the context of IHCA detection, LLMs offer several advantages: they can interpret unstructured clinical notes to extract relevant information [18], analyze the broader context of a patient's medical history [15], process large volumes of data rapidly for real-time or near-real-time detection [16], apply consistent criteria across cases to reduce human variability and bias [17], and continuously update with new data and guidelines to stay current with medical knowledge.

However, the application of LLMs in critical care settings like IHCA detection also presents challenges, including the need for high accuracy, robustness to variations in clinical documentation, and the ability to follow specific medical instructions consistently.

1.3 Study Objectives and Hypotheses

The primary goal of this study is to systematically evaluate the performance, robustness, and instruction-following capabilities of various LLMs in the context of IHCA detection. Specifically, the objectives are to:

1. Assess the accuracy of multiple open-source LLMs in detecting IHCA events.
2. Evaluate the robustness of these models by analyzing performance consistency across different runs.
3. Investigate the ability of LLMs to follow instructions.
4. Compare the effectiveness of general-purpose LLMs with models specifically fine-tuned for medical applications.
5. Benchmark the performance of open-source LLMs against GPT-4o.

The study is grounded in the following hypotheses: While we expect GPT-4o to exhibit superior performance, we hypothesize that several open-source models will achieve competitive results. We also anticipate that medical-specific models may deliver higher precision in IHCA detection, but could face challenges in maintaining consistent adherence to guidelines.

2 Methods

This retrospective, observational study was reviewed and approved by the University of Pennsylvania Institute Review Board.

2.1 Data Collection and Preprocessing

We collected data from the Hospital of the University of Pennsylvania, focusing on adult ($\geq$18 years old) inpatient encounters from June 2018 to March 2022. The dataset included patients with reported clinical emergency or rapid response calls, identified from a quality improvement database hosted on REDCap, an encrypted, HIPAA-compliant, and Penn Medicine-approved data management software. We extracted discharge summaries, progress notes, and assessments from the electronic health record (EHR) and de-identified them using an adapted version of PHIlter [19].

Our preprocessing pipeline involved several key steps:

1. **Note type selection:** We prioritized discharge summaries, progress notes, plan of care, and assessments based on their likelihood of containing relevant IHCA information.
2. **Note sampling:** For each patient, we selected up to four notes, prioritizing them in the enumerated order listed above. This approach ensured a balanced representation of information while focusing on the most informative note types.
3. **Labeling:** Notes were manually reviewed by clinical experts with emergency medicine experience to determine true IHCA labels. Positive IHCA was defined as the loss of pulses followed by the delivery of chest compressions.

For both GPT-4o and the open-source models, we developed a specialized instruction prompt:

```
INSTRUCTION_PROMPT = """
Given the patient's notes summary, assess whether the patient had an
in-hospital cardiac arrest (IHCA) event. IHCA refers to a cardiac arrest
that occurs in an inpatient setting where the patient is admitted and
receiving care. Cardiac arrests occurring in outpatient settings, such as
the hospital's cafeteria or during a visit, do not count as IHCA. Cardiac
arrest refers to the sudden loss of cardiac function leading to a lack of
a pulse or signs requiring that the patient receives chest compressions,
or ACLS.

Steps to determine IHCA:
1. Identify if there is a mention of cardiac arrest.
2. Determine if the cardiac arrest occurred in an inpatient setting.
3. Confirm if the patient received chest compressions.

Respond with 'positive' or 'negative' and give your rationale.
"""
```

This prompt structure was designed to provide clear, consistent instructions to both GPT-4o and the open-source models, enabling fair comparison across different model architectures and sizes.

3 Methods

3.1 Model Selection and Implementation

In this study, we evaluated 52 LLMs, including 51 open-source LLMs and GPT-4o, which served as a benchmark. The open-source models included both general-purpose (36 models) and medical-specific (15 models) variants, enabling a comprehensive comparison of their capabilities.

3.2 Experimental Design

The experimental design employed a zero-shot learning approach. To enhance reasoning capabilities, we used a chain-of-thought prompting strategy that provided structured instructions for identifying IHCA events. Each model processed the same set of patient notes with the same prompt, allowing for a fair comparison (The temperature parameter was 0.2 for all models).

3.3 Evaluation Metrics

Model performance was evaluated using several metrics, including F1 score to measure the balance between precision and recall, balanced accuracy to account for class imbalance, and both precision and recall to assess the models' ability to correctly identify IHCA cases. To assess robustness, we derived confidence intervals through a non-parametric bootstrapping method with 100,000 resamples, providing a quantification of the uncertainty in our performance estimates without assumptions about the underlying data distribution.

3.4 Instruction Following Assessment

The models' ability to follow clinical instructions was evaluated by analyzing their responses to the provided prompt. We calculated an instruction-following score based on whether the models' predictions began with "positive" or "negative" within the first five tokens of their responses. This score was determined by the proportion of correct responses relative to the total number of predictions, offering insights into the models' reliability in adhering to specific formats in clinical settings.

3.5 Model Performance Comparison

We evaluated the performance of 52 different language models, including GPT-4o, on IHCA detection.

3.5.1 Dataset and Testing Methodology

The testing methodology differed significantly between GPT-4o and the open-source models. GPT-4o was evaluated on a balanced sample dataset of 21 patients due to API cost limitations, while the open-source models underwent a progressive evaluation, starting with (randomly sampled) smaller subsets of 10 and 100 documents, moving to 1,000 documents, and finally testing on the full dataset for the most promising models (F1 Score > 50%). All models were implemented using the HuggingFace Transformers library and executed on a Microsoft Azure Databricks cluster equipped with NVIDIA A100 GPUs, with an estimated total of approximately 720 GPU hours dedicated to this study.

4 Results

4.0.1 Dataset Characteristics

Our study analyzed 2,674 medical notes from 684 unique patients, with an uneven distribution of IHCA cases: 63.43% positive and 36.57% negative, emphasizing the need for metrics like balanced accuracy and F1 score, which account for uneven distributions. Note lengths had a maximum of 27,671 tokens, with an average of 2,334 tokens and a median of 1,911 tokens. For consistent evaluation, all LLMs were provided access to the first 7,000 tokens of each document, ensuring fair comparison between models with limited context windows (1K - 10K tokens) and those with capacities over 128K tokens.

4.1 F1 Score

In IHCA detection, a high F1 score indicates that the model can accurately identify cardiac arrest cases without generating excessive false positives or missing true events. This balance is crucial in a clinical setting where both over-identification (leading to unnecessary interventions) and under-identification (missing critical events) can have serious consequences. GPT-4o demonstrated the highest F1 score among all models (shown in Figure 1, see Appendix for other key plots), achieving 0.90 ± 0.01. Among the open-source models, several exhibited competitive performance:

- *mistralai/Mistral-Nemo-Instruct-2407*: 0.84 ± 0.05
- *princeton-nlp/gemma-2-9b-it-SimPO*: 0.82 ± 0.07
- *THUDM/glm-4-9b-chat*: 0.79 ± 0.00
- *meta-llama/Meta-Llama-3.1-8B-Instruct*: 0.79 ± 0.05

Medical models like *ruslanmv/Medical-Llama3-8B* and *aaditya/Llama3-OpenBioLLM-8B* also showed competitive F1 scores (0.78 ± 0.01 and 0.76 ± 0.02, respectively).

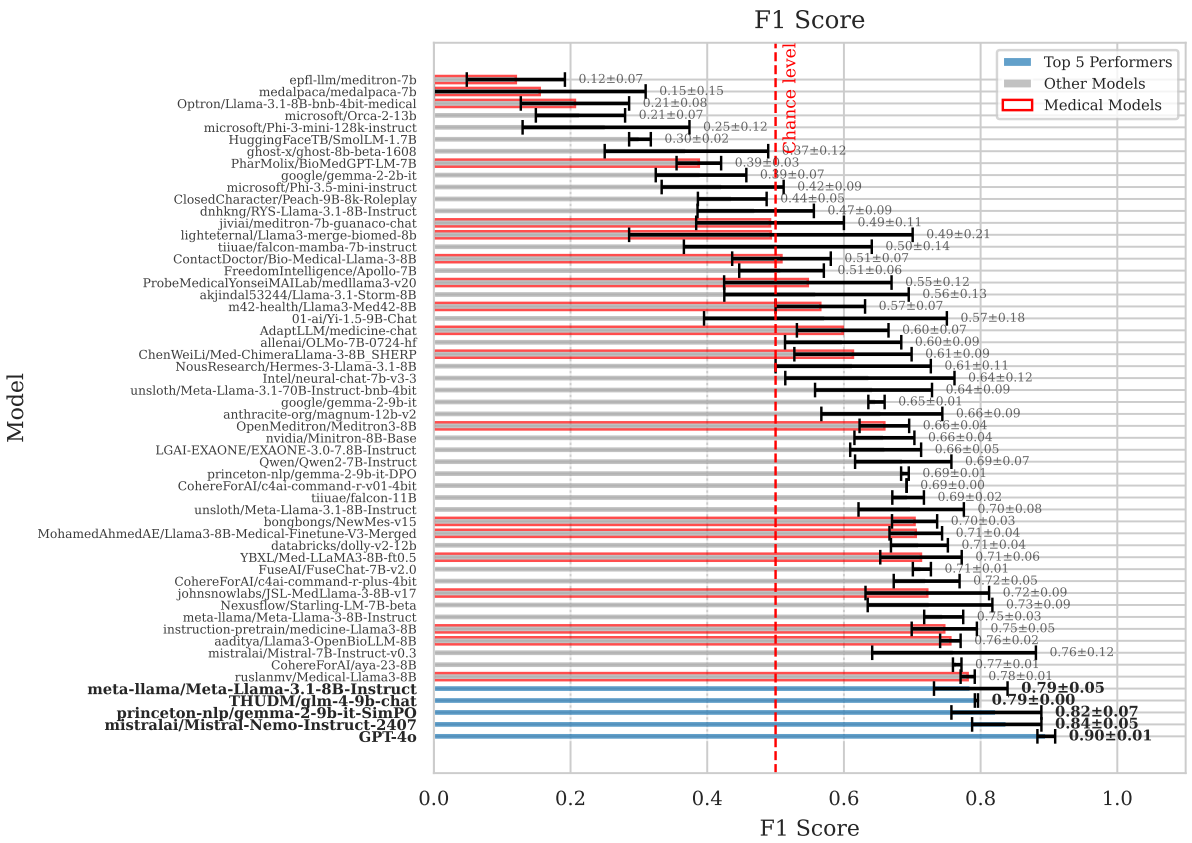

Figure 1: F1 score comparison of LLMs. This plot displays the F1 scores across models, providing insight into their precision and recall balance in IHCA detection. The confidence intervals were calculated with non-parametric bootstrapping across several runs of each model. The red dashed line represents chance level for the binary classification task.

4.2 Balanced Accuracy

Balanced accuracy is particularly important in our study due to the potential class imbalance in IHCA cases. GPT-4o led with a score of 0.90 ± 0.01, followed by:

- *princeton-nlp/gemma-2-9b-it-SimPO*: 0.84 ± 0.06
- *mistralai/Mistral-Nemo-Instruct-2407*: 0.84 ± 0.04
- *mistralai/Mistral-7B-Instruct-v0.3*: 0.76 ± 0.07
- *unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit*: 0.74 ± 0.05

Medical models such as *johnsnowlabs/JSL-MedLlama-3-8B-v17* (0.74 ± 0.03) exhibited slightly lower balanced accuracy, with similar confidence intervals.

4.3 Precision and Recall

In the context of IHCA detection:

- Precision (positive predictive value) indicates the proportion of correctly identified IHCA cases among all cases flagged as positive by the model. High precision minimizes false alarms, reducing unnecessary interventions and resource allocation.
- Recall (sensitivity) represents the proportion of actual IHCA cases correctly identified by the model. High recall is critical for ensuring that genuine IHCA events are not missed, as failing to detect a previous cardiac arrest could have severe consequences.

The trade-off between precision and recall is particularly important in IHCA detection, where the cost of false negatives (missed IHCA events) may be considered higher than that of false positives.

4.3.1 Precision

Several open-source models demonstrated exceptionally high precision:

- *unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit*: 0.99 ± 0.01
- *princeton-nlp/gemma-2-9b-it-SimPO*: 0.98 ± 0.02
- *microsoft/Phi-3.5-mini-instruct*: 0.97 ± 0.03
- *m42-health/Llama3-Med42-8B*: 0.97 ± 0.03
- *NousResearch/Hermes-3-Llama-3.1-8B*: 0.96 ± 0.04

Medical models like *m42-health/Llama3-Med42-8B* demonstrated strong precision, with narrow confidence intervals similar to general-purpose models.

4.3.2 Recall

Models with high recall include:

- *GPT-4o*: 0.97 ± 0.03
- *CohereForAI/aya-23-8B*: 0.97 ± 0.02
- *THUDM/glm-4-9b-chat*: 0.95 ± 0.05
- *ruslanmv/Medical-Llama3-8B*: 0.94 ± 0.02
- *meta-llama/Meta-Llama-3-8B-Instruct*: 0.89 ± 0.10

Medical models like *ruslanmv/Medical-Llama3-8B* demonstrated strong recall, with confidence intervals similar to general-purpose models.

4.4 Model Robustness and Variability

Robustness, as indicated by narrow confidence intervals, is crucial in clinical applications. A model with consistent performance across different patient populations and clinical scenarios is more reliable for real-world deployment. Wide confidence intervals suggest that a model's performance may vary significantly depending on the specific characteristics of the input data, which could be problematic in diverse clinical settings.

The confidence intervals provide valuable insight into the robustness of the models. Narrow intervals, as observed in GPT-4o (F1: 0.90 ± 0.01, Balanced Accuracy: 0.90 ± 0.01), indicate consistent performance across different samples, suggesting high reliability. *Mistral-Nemo-Instruct-2407* (F1: 0.84 ± 0.05, Balanced Accuracy: 0.84 ± 0.04) shows slightly wider intervals, suggesting some variability, but still relatively consistent performance. In contrast, wider intervals, such as those found in *princeton-nlp/gemma-2-9b-it-SimPO* (F1: 0.82 ± 0.07, Balanced Accuracy: 0.84 ± 0.06), indicate greater variability and potential sensitivity to specific data samples.

Some medical models, like *MohamedAhmedAE/Llama3-8B-Medical-Finetune-V3-Merged* (F1: 0.71 $\pm$ 0.12), exhibited even wider confidence intervals, reflecting a lower degree of reliability in different contexts. This suggests that these models may be less dependable when applied to varied datasets, showing more significant performance fluctuations across different samples.

4.5 Instruction Following Performance

In a clinical context, the ability to follow specific instructions is paramount. This metric assesses how well a model adheres to the given prompt structure, which is crucial for integration into existing clinical workflows and for ensuring that the model's outputs are in a format that is immediately useful to healthcare providers, e.g., in our use case, consistently indicating "positive" or "negative" in its response. Instruction following capabilities varied widely among the models (see Figure 2):

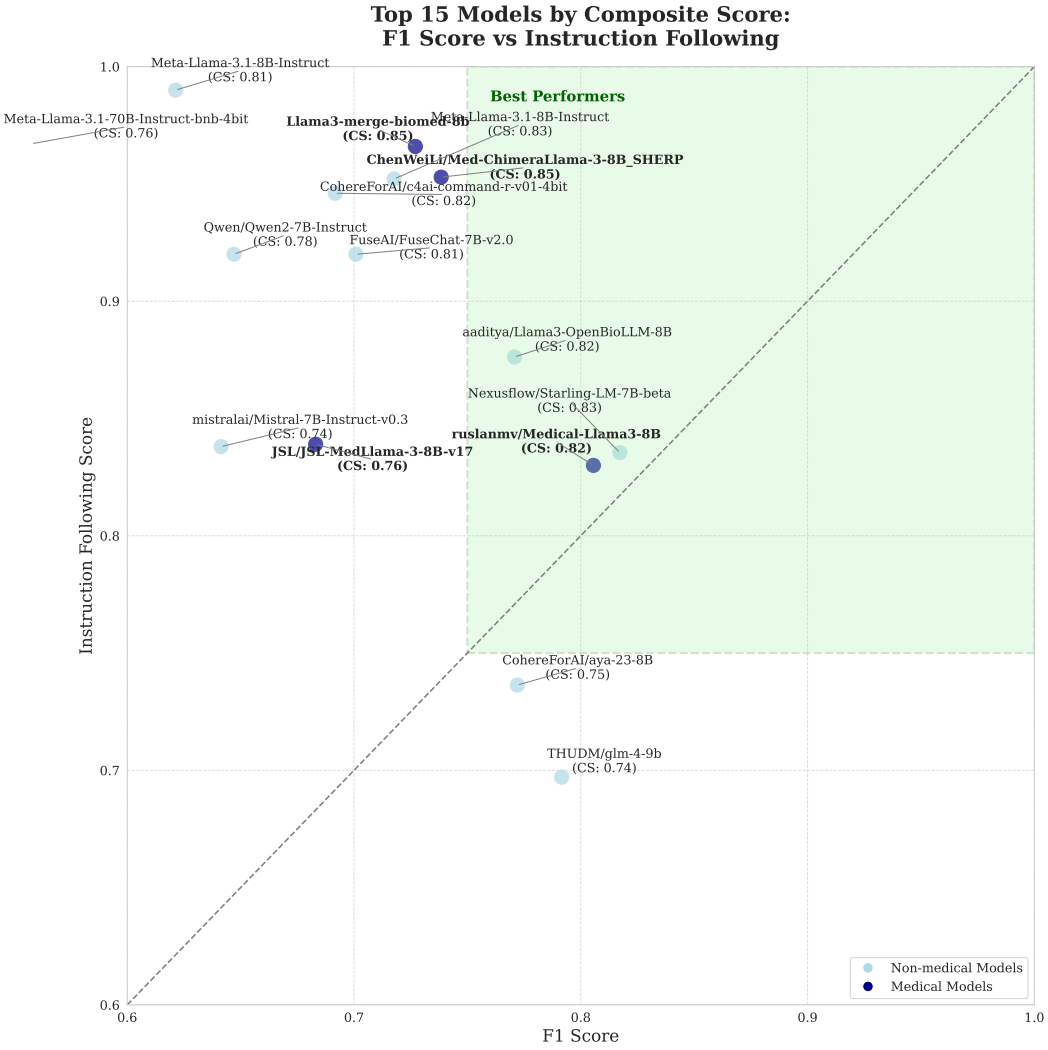

Figure 2: Top 15 models by composite score (CS). This plot illustrates the relationship between F1 score and instruction-following ability for the top 15 models, with the composite score representing the average of these two metrics. Medical model labels are shown in bold.

- General-purpose models like *unsloth/Meta-Llama-3.1-8B-Instruct* and *unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit* excelled, achieving scores of 99.0% and 96.7%, respectively.
- Some medical models performed well, with *lighteternal/Llama3-merge-biomed-8b* and *ChenWeiLi/Med-ChimeraLlama-3-8B_SHERP* scoring 96.6% and 95.3%, respectively.
- However, other medical models struggled, with *MohamedAhmedAE/Llama3-8B-Medical-Finetune-V3-Merged* scoring as low as 32.6%.

This wide range in instruction following performance highlights the importance of this metric in assessing model reliability for clinical applications.

4.6 Overall Performance

GPT-4o demonstrated superior performance across all metrics, with high scores and low variability. Several open-source models, particularly larger variants like *Mistral-Nemo-Instruct-2407* and *princeton-nlp/gemma-2-9b-it-SimPO*, showed competitive performance, often achieving F1 scores and balanced accuracy within 10% of GPT-4o's performance.

The variability in performance, as indicated by the confidence intervals, suggests that while some models (e.g., GPT-4o and *ruslanmv/Medical-Llama3-8B*) are highly consistent, others may be more sensitive to the specific characteristics of the input data. This variability is an important consideration for real-world deployment, where consistent performance across diverse patient populations is crucial.

Furthermore, the evaluation of instruction-following capabilities highlights the importance of model responsiveness in clinical contexts. While some medical models lag behind in instruction-following performance, surprisingly, top-performing general-purpose models demonstrate both strong task performance and reliable instruction adherence, making them potentially more versatile in real-world applications.

5 Discussion

5.1 Interpretation of Results

Our comprehensive evaluation of 52 large language models, including GPT-4o and 51 open-source alternatives (15 medical-specific and 36 general-purpose), reveals several key insights into their potential for IHCA detection:

- **Performance of Open-Source Models:** While GPT-4o demonstrated superior performance (F1: 0.90 ± 0.01, Balanced Accuracy: 0.90 ± 0.01), several open-source models showed competitive results. Notably, *mistralai/Mistral-Nemo-Instruct-2407* (F1: 0.84 ± 0.05) and *princeton-nlp/gemma-2-9b-it-SimPO* (F1: 0.82 ± 0.07) achieved performance within 10% of GPT-4o, suggesting robust IHCA detection capabilities in accessible alternatives.
- **Impact of Medical Fine-tuning:** Medical-specific models showed mixed results. Some, like *ruslanmv/Medical-Llama3-8B* (Recall: 0.94 ± 0.02), demonstrated strong performance in certain metrics, but often exhibited wider confidence intervals compared to top general-purpose models. This suggests that fine-tuning on medical text can enhance domain-specific capabilities, but may not always translate to superior overall performance in IHCA detection.
- **Variability within Model Families:** We observed significant performance variability among models from the same family. Llama variants showed F1 scores ranging from 0.79 ± 0.05 (*meta-llama/Meta-Llama-3.1-8B-Instruct*) to 0.49 ± 0.21 (*lighteternal/Llama3-merge-biomed-8b*). Similarly, Gemma models ranged from 0.98 ± 0.02 to 0.75 ± 0.02 in precision. This highlights the impact of specific fine-tuning approaches and the importance of careful model selection.
- **Instruction Following Capabilities:** General-purpose models often outperformed medical-specific models in instruction following, with top performers like *unsloth/Meta-Llama-3.1-8B-Instruct* achieving near-perfect scores (99.0%). This suggests that structured prompt following may be more related to general language understanding than domain-specific training.
- **Robustness and Consistency:** Confidence intervals revealed significant differences in model robustness. Larger general-purpose models often demonstrated consistent performance across data samples, while others, including several medical models, showed wider variability. This shows the importance of considering both peak performance and consistency when evaluating models for clinical applications.

5.2 Clinical Implications

The high performance of several LLMs in IHCA detection suggests significant potential for improving the efficiency and accuracy of critical care event identification in clinical settings. Accurate detection of patients who have suffered IHCA during their hospitalization could greatly enhance QI efforts by ensuring more comprehensive capture of events for review and analysis. Additionally, high-precision models have the potential to reduce the substantial workload of reviewing individual patient records for IHCA for applications in both research and QI. The competitive performance of some open-source models also indicates that effective AI-assisted IHCA detection could be implemented more widely, even in resource-constrained healthcare settings.

Furthermore, models with high recall and strong instruction-following capabilities could be integrated into real-time monitoring systems, enabling faster responses to patients with prior IHCA events. How-

ever, the variability in performance and instruction-following capabilities among models illustrates the importance of careful selection and rigorous validation before clinical deployment.

5.3 Limitations

Despite the promising results, our study has several limitations. The study was conducted on data from a single institution, which may limit the generalizability of findings to other healthcare settings with different patient populations or documentation practices. While we used a substantial dataset, an even larger and more diverse dataset could provide more robust estimates of model performance and generalizability. Although we evaluated a wide range of models, the rapidly evolving field of LLMs means that newer models may have become available since the study's completion. The inherent biases in medical documentation practices could influence model performance and potentially perpetuate existing healthcare disparities if not carefully addressed. Additionally, our evaluation focused on IHCA detection from text alone, without incorporating other clinical data sources that might be available in real-world settings.

5.4 Future Directions

We propose several directions for future research. Enhancing model robustness is crucial, with a focus on developing techniques to improve the consistency of open-source models, particularly in handling diverse patient populations and documentation styles. Explainable AI should be a priority, focusing on improving the quality and consistency of model explanations to enhance interpretability, facilitate chart review, and build clinical trust. Multi-modal integration presents an exciting avenue, exploring the integration of text-based LLMs with other data modalities such as vital signs and lab results for more comprehensive IHCA detection.

Prospective validation studies will be essential to evaluate the real-world impact of LLM-assisted prior IHCA detection on patient outcomes and quality improvement processes. Ethical AI development must be at the forefront, investigating methods to mitigate potential biases in model training and deployment to ensure equitable performance across diverse patient groups. Lastly, optimizing instruction following capabilities is crucial, with efforts directed towards developing specialized training techniques to enhance the instruction-following capabilities of medical-specific models without compromising their domain expertise.

6 Conclusion

This evaluation of 52 large language models for IHCA detection demonstrates the significant potential of these AI technologies in this critical care application. While GPT-4o exhibited superior performance and robustness, several open-source alternatives, both general-purpose and medical-specific, showed promising results that approach clinical viability. The study revealed that fine-tuning on medical text does not guarantee superior performance, with some general-purpose models outperforming their medical counterparts in certain aspects, particularly in instruction following. The variability in model performance, both within and across model families, emphasizes the complexity of applying LLMs to specialized medical tasks. It highlights the need for careful model selection and rigorous evaluation that considers not only peak performance but also consistency, robustness, and adherence to clinical instructions.

As the field of AI in healthcare continues to evolve rapidly, our findings provide valuable insights for researchers, clinicians, and health system leaders considering the implementation of LLMs for critical care applications. The potential for improving IHCA detection and, consequently, patient outcomes is significant. However, the path to clinical deployment requires addressing challenges in model performance, consistency, and ethical considerations.

Future work should focus on enhancing the reliability and interpretability of these models, conducting prospective validations in diverse clinical settings, and developing frameworks for responsible AI integration in healthcare. With continued research and development, LLMs could become powerful tools in the ongoing effort to improve the quality and efficiency of critical care, ultimately contributing to better patient outcomes in IHCA and beyond.

References

[1] Lars W. Andersen, Mads J. Holmberg, Katherine M. Berg, Michael W. Donnino, and Asger Granfeldt. In-Hospital Cardiac Arrest: A Review. *JAMA*, 321(12):1200–1210, 2019.

[2] Lingling Wu, Bharat Narasimhan, Kirtipal Bhatia, Kam S. Ho, Chayakrit Krittanawong, Wilbert S. Aronow, Patrick Lam, Salim S. Virani, and Salpy V. Pamboukian. Temporal Trends in Characteristics and Outcomes Associated With In-Hospital Cardiac Arrest: A 20-Year Analysis (1999–2018). *Journal of the American Heart Association*, 10(23), 2021.

[3] Saket Girotra, Brahmajee K. Nallamothu, John A. Spertus, Yan Li, Harlan M. Krumholz, and Paul S. Chan. Trends in Survival after In-Hospital Cardiac Arrest. *New England Journal of Medicine*, 367(20):1912–1920, 2012.

[4] Dhaval Kolte, Sahil Khera, Wilbert S. Aronow, Chandrasekar Palaniswamy, Marjan Mujib, Chul Ahn, Sei Iwai, and Gregg C. Fonarow. Regional Variation in the Incidence and Outcomes of In-Hospital Cardiac Arrest in the United States. *Circulation*, 131(16), 2015.

[5] Sebastian Wiberg, Mathias J. Holmberg, Michael W. Donnino, Jesper Kjaergaard, Christian Hassager, Lise Witten, Katherine M. Berg, Ari Moskowitz, Lars W. Andersen, Anne Grossestreuer, Dana Edelson, Joseph Ornato, Mary Ann Peberdy, Matthew Churpek, Michael Kurz, Monique Anderson Starks, Paul Chan, Saket Girotra, Sarah Perman, and Zachary Goldberger. Age-dependent trends in survival after adult in-hospital cardiac arrest. *Resuscitation*, 151:189–196, 2020.

[6] Merel Schluep, Sanne E. Hoeks, Marleen Blans, Berber van den Bogaard, Annelies Koopman-van Gemert, Charlotte Kuijs, Christine Hukshorn, Nynke van der Meer, Matthijs Knook, Tim van Melsen, Rene Peters, Peter Perik, Kim Simons, Gerard Spijkers, Willem Vermeijden, Evelien J. Wils, Robert Jan Stolker, and Rik Endeman. Long-term survival and health-related quality of life after in-hospital cardiac arrest. *Resuscitation*, 167:297–306, 2021.

[7] Ashish R. Panchal, Jason A. Bartos, José G. Cabañas, Michael W. Donnino, Ian R. Drennan, Karen G. Hirsch, and Peter J. Kudenchuk. Part 3: Adult Basic and Advanced Life Support: 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation*, 142(16_suppl_2), 2020.

[8] Jacob C. Jentzer, Michael Scutella, Francis Pike, Justin Fitzgibbon, Nicole M. Krehel, Lindsey Kowalski, Clifton W. Callaway, Jon C. Rittenberger, Jason C. Reynolds, Gregory W. Barsness, and Cameron Dezfulian. Early coronary angiography and percutaneous coronary intervention are associated with improved outcomes after out of hospital cardiac arrest. *Resuscitation*, 123:15–21, 2018.

[9] Raina M. Merchant, Robert A. Berg, Lin Yang, Lance B. Becker, Peter W. Groeneveld, and Paul S. Chan. Hospital Variation in Survival After In-hospital Cardiac Arrest. *Journal of the American Heart Association*, 3(1), 2013.

[10] Rohan Khera, Andrew Humbert, Brian Leroux, Graham Nichol, Peter Kudenchuk, Damon Scales, Andrew Baker, Mike Austin, Craig D. Newgard, Ryan Radecki, Gary M. Vilke, Kelly N. Sawyer, George Sopko, Ahamed H. Idris, Henry Wang, Paul S. Chan, and Michael C. Kurz. Hospital Variation in the Utilization and Implementation of Targeted Temperature Management in Out-of-Hospital Cardiac Arrest. *Circulation: Cardiovascular Quality and Outcomes*, 11(11), 2018.

[11] Ankur Gupta, Tharanga Perera, Ananda Ganesan, Trish Sullivan, Dennis H. Lau, Kevin C. Roberts-Thomson, Anthony G. Brooks, and Prashanthan Sanders. Complications of catheter ablation of atrial fibrillation: a systematic review. *Circulation: Arrhythmia and Electrophysiology*, 6(6):1082–1088, 2013.

[12] Kjetil Sunde, Marius Pytte, D. Jacobsen, Arne Mangschau, Lars P. Jensen, Cathrine Smedsrud, Trond Draegni, and Petter A. Steen. Implementation of a standardised treatment protocol for post resuscitation care after out-of-hospital cardiac arrest. *Resuscitation*, 73(1):29–39, 2007.

[13] Rohan Khera, John A. Spertus, Monique A. Starks, and et al. Administrative Codes for Capturing In-Hospital Cardiac Arrest. *JAMA Cardiology*, 2(11):1275–1277, 2017.

[14] Christopher DeZorzi, Brendan Boyle, Ahmed Qazi, Kanav Luthra, Rohan Khera, Paul S. Chan, and Saket Girotra. Administrative billing codes for identifying patients with cardiac arrest. *Journal of the American College of Cardiology*, 73(12):1598–1600, 2019.

[15] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):1–13, 2021.

[16] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023.

[17] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT, 2023.

[18] Yikuan Li, Sheng Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):1–12, 2020.

[19] Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. Protected Health Information filter (Philter): Accurately and securely de-identifying free-text clinical notes. *NPJ Digital Medicine*, 3(1):57, 2020.

7 Appendix

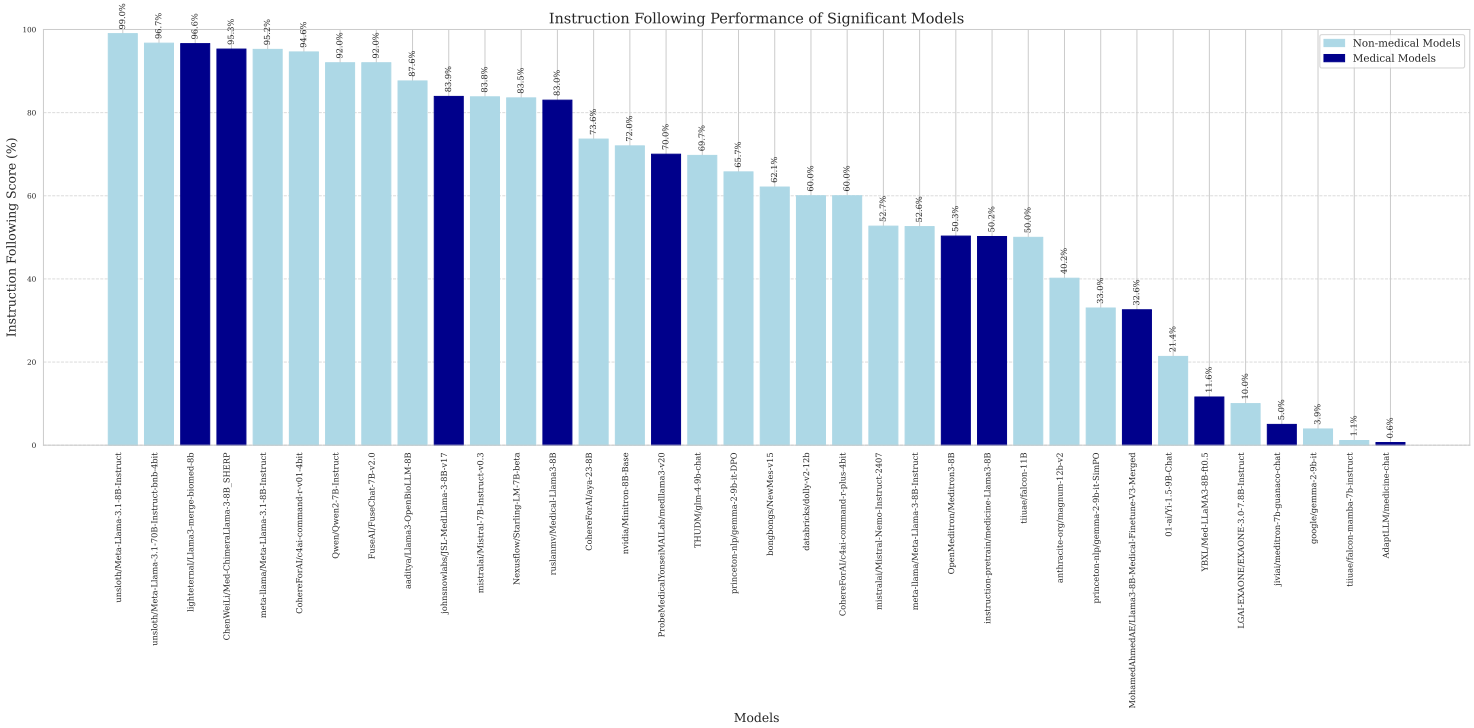

Figure A1: Instruction Following Performance of Significant Models. This bar chart displays the instruction following scores for models that showed statistically significant performance. The scores represent the percentage of responses that correctly followed the given instructions. Models are ordered by their instruction following ability, with medical models highlighted in dark blue.

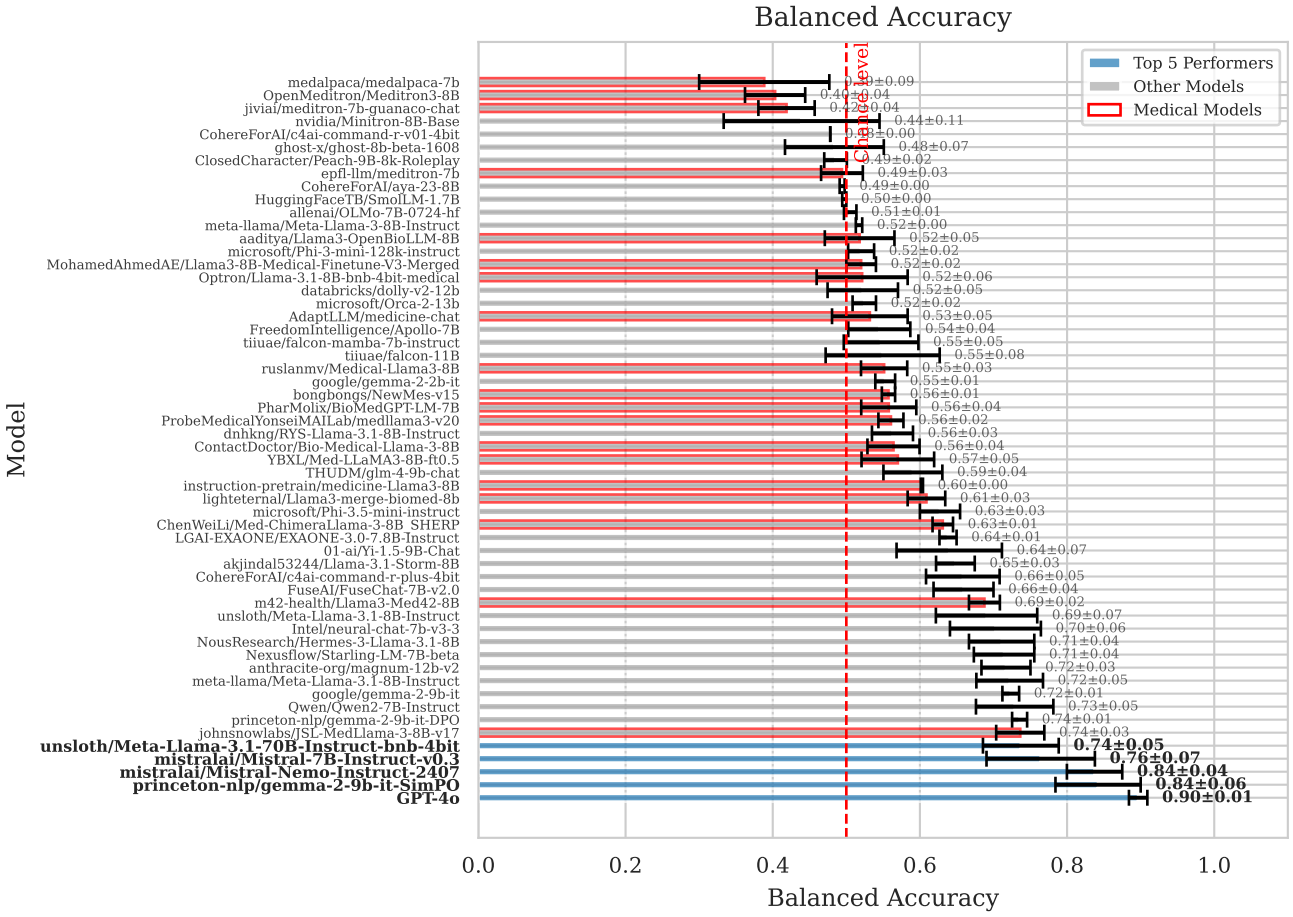

Figure A2: Balanced Accuracy Comparison of Models. This bar chart compares the balanced accuracy scores of various models, including both medical and non-medical models. The top 5 performers are highlighted, and medical models are distinguished by color. Error bars represent confidence intervals calculated through bootstrapping.

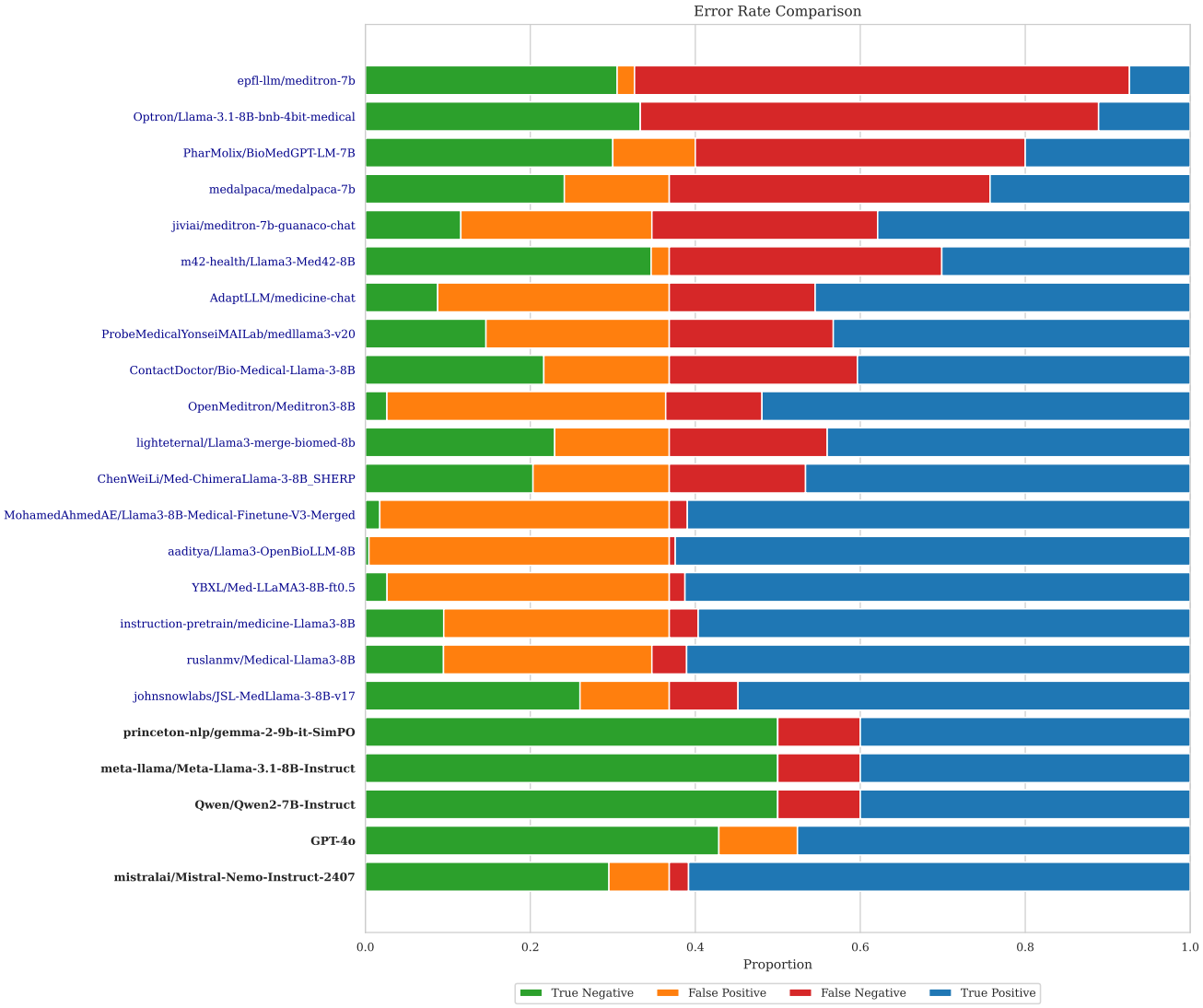

Figure A3: Error Rate Comparison of Top Models. This stacked bar chart illustrates the distribution of true negatives, false positives, false negatives, and true positives for the top-performing models. It provides a comprehensive view of each model's performance in terms of correct and incorrect predictions.

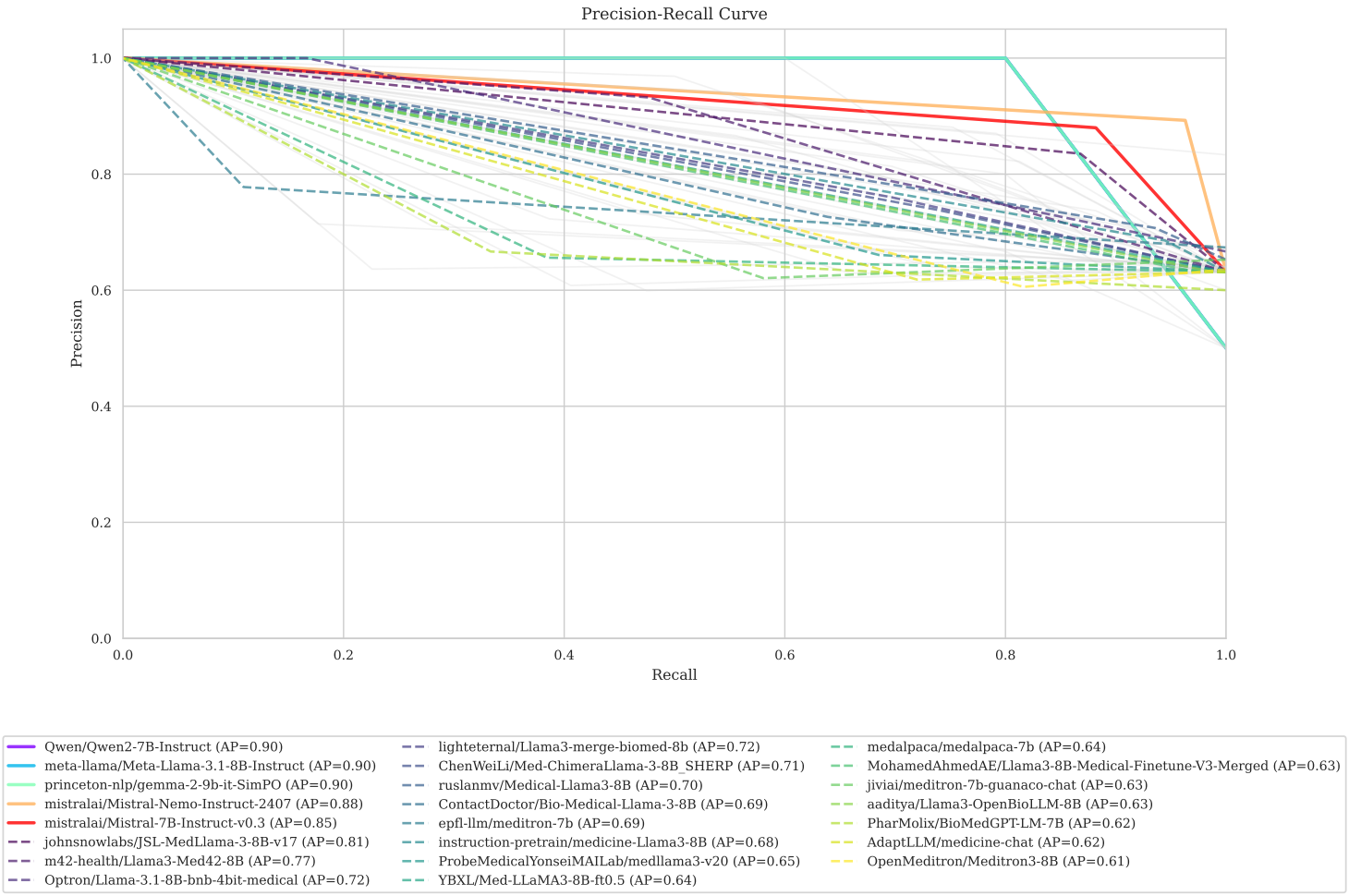

Figure A4: Precision-Recall Curves for Model Comparison. This plot shows the precision-recall curves for various models, allowing for a comparison of their performance across different classification thresholds. The area under each curve (AP) is indicated in the legend, with higher values suggesting better overall performance. Some models are not shown the legend due to space limitations.

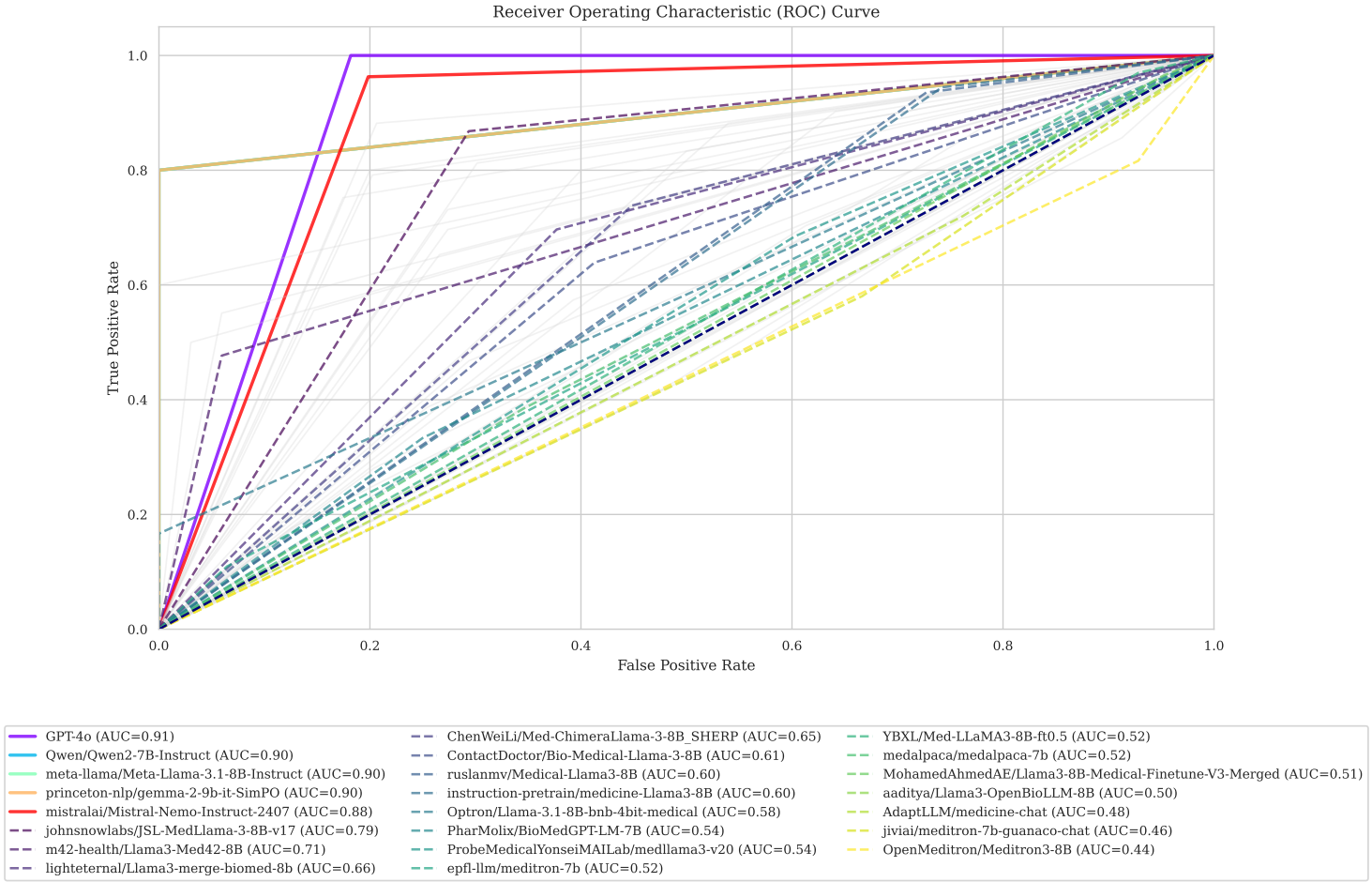

Figure A5: Receiver Operating Characteristic (ROC) Curves for Model Comparison. This plot displays the ROC curves for various models, illustrating their ability to distinguish between positive and negative cases across different classification thresholds. The area under each curve (AUC) is provided in the legend, with higher values indicating better discriminative power. Some models are not shown on the legend due to space limitations.