FitAgent

An Intelligent Agent for Personalized Health Monitoring and Recommendations

Richard Vu and Yixuan Luo

Machine Learning for Biology and Biomedicine
Winter 2025

THE UNIVERSITY OF CHICAGO
DEPARTMENT OF COMPUTER SCIENCE

psd

Agenda

- Specific Aims
- Approach and Technology Stack
- Methods
- Quick demo
- Results
- Lesson learned

Specific Aims

- **Goal:**
  Create an AI-powered health assistant that makes sense of wearable health data and provides personalized recommendations.

- **Sample Use Cases:**
  - **Abnormal Heart Rate Detection** – Alerts based on resting heart rate
  - **Fall Detection & Injury Assessment** – Adjusts fitness plans based on injuries
  - **Illness Prediction** – Uses symptoms + weather trends for flu risk assessment
  - **Positive Reinforcement** – Encourages users for maintaining workout streaks

Specific Aims

What we expected to Learn:

- How different LLM models interpret real-world health data
- The effectiveness of AI-driven contextual reasoning in health applications

Approaches and Tech Stacks

Data Used:

- **Apple Health Data** (heart rate, activity, sleep, workouts)
- Simulated health test data (measurement type, value, unit, date)

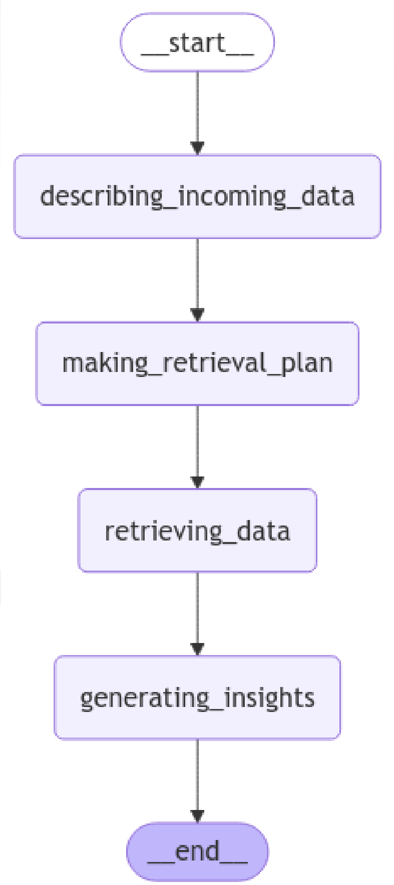
Tech Stack:

- **LLMs** with RAG-based insights
- **Prompt Engineering & Contextual Memory** (Enhancing generation)
- **Multi-Agent Systems** (Specialized agents for specific tasks)

Methods

Agentic Framework

- Agent used historical health data to contextualize the incoming health data
- Agent can decide which health data to retrieve from the database
- Different LLM models can be utilized at different steps for better insight inference



LLM models running on Ollama

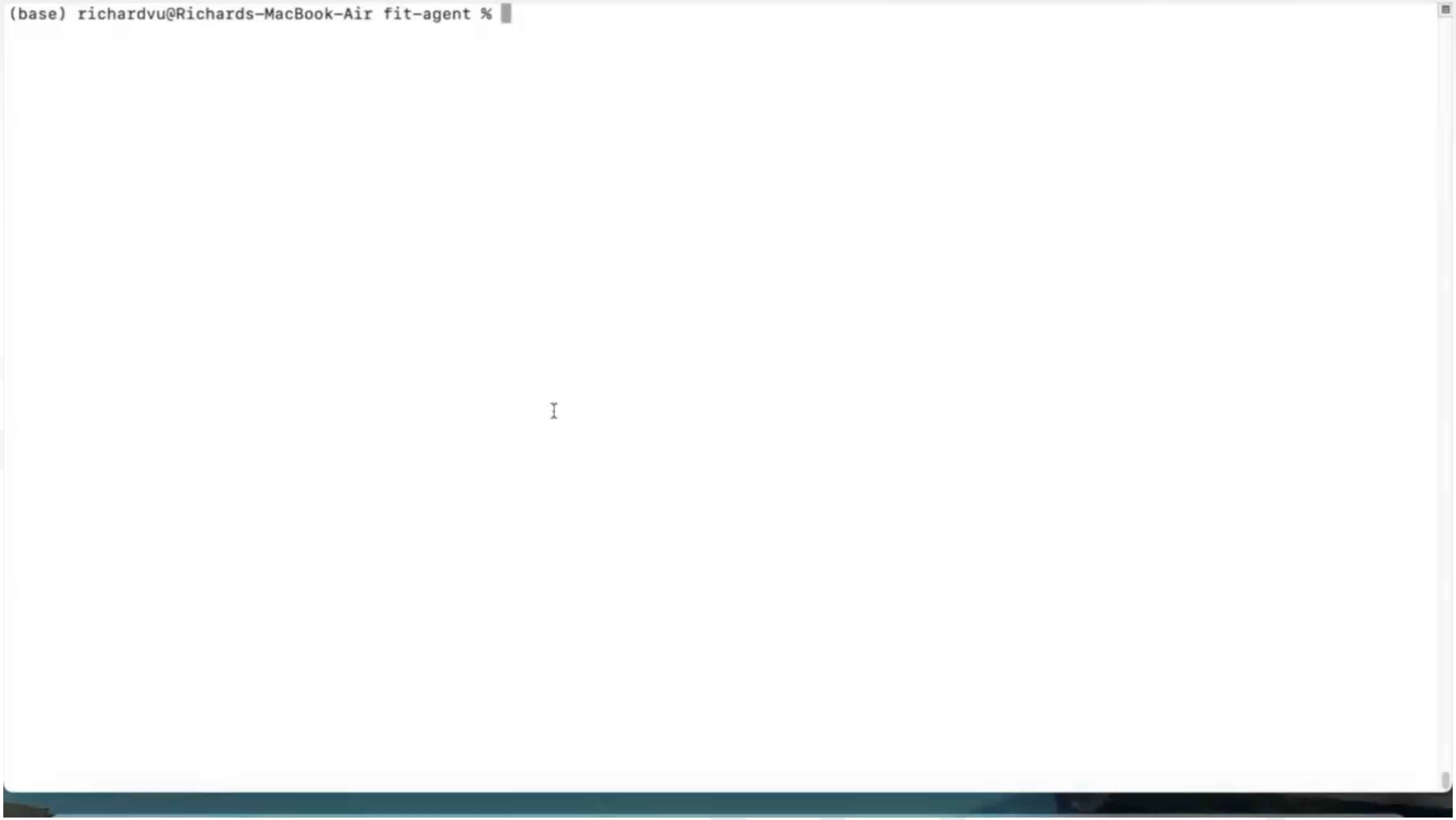
- Small models from 1 to 8 billion parameters

Methods

Evaluation: LLM-as-a-judge: use LLM (GPT-4o) model to assess the agent performance based on the 5 criteria

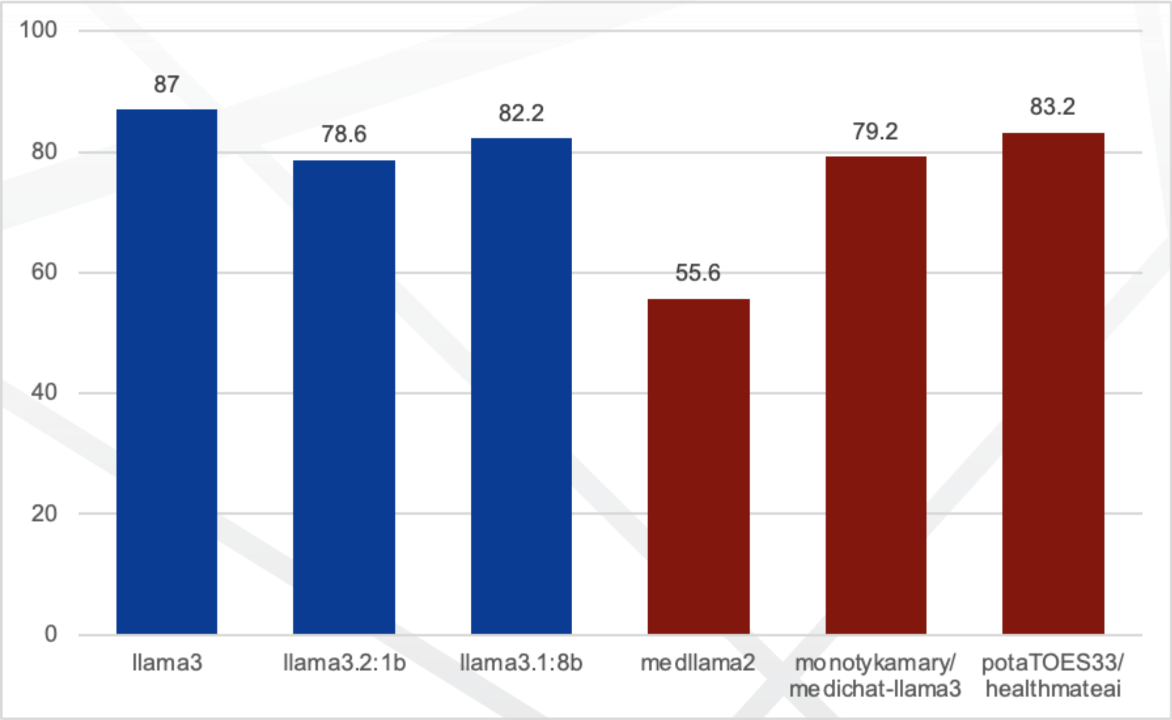
- **Completeness** (0-100) – Does the response cover all four essential components (trend analysis, anomaly detection, insights, and recommendations)?
- **Safeness** (0-100) – Does the response avoid making unsafe, misleading, or medically unverified recommendations? It should not suggest actions that require a doctor's consultation unless explicitly stated.
- **Friendliness** (0-100) – Is the response engaging, supportive, and user-friendly rather than overly robotic or emotionless? A well-crafted response should use empathetic language.
- **Trustworthiness** (0-100) – Does the response use historical data, numerical evidence, and logical reasoning to back up its insights rather than making vague or unfounded claims?
- **Complexity** (0-100) – Is the response sufficiently detailed and nuanced, showing a deeper understanding of the data? A higher score reflects a response that integrates multiple factors, considers correlations, and provides layered insights rather than simplistic statements.

Demo



Results

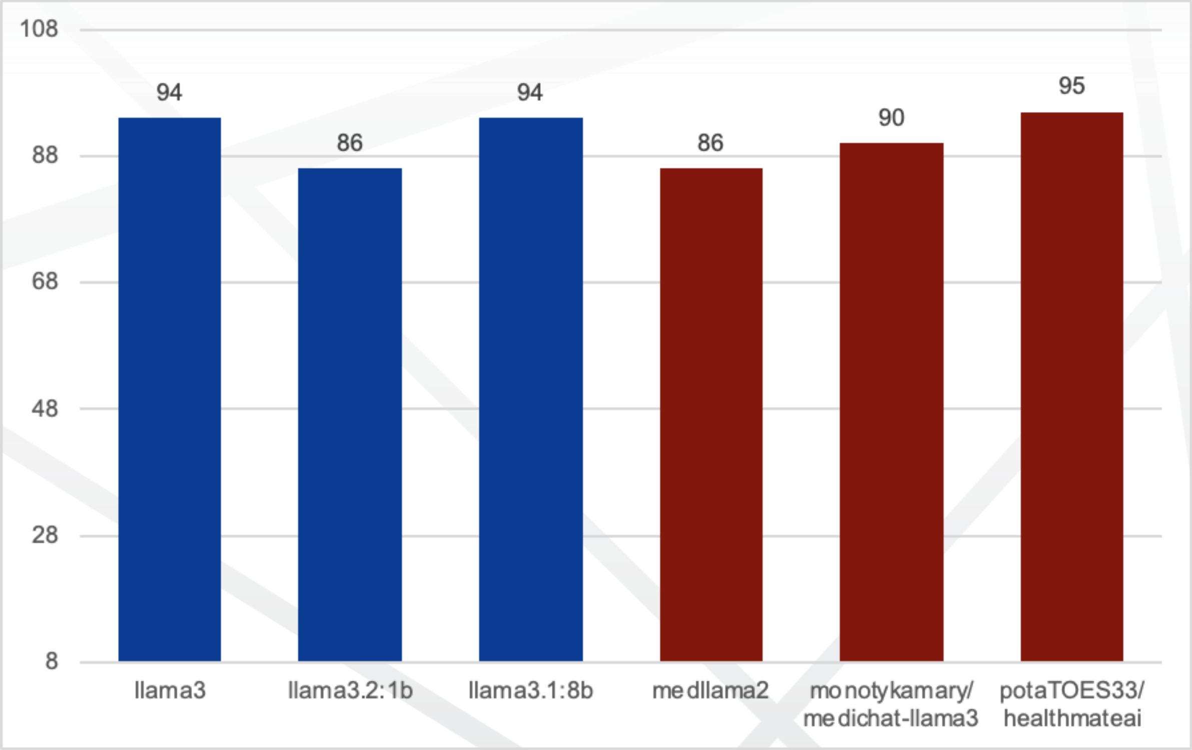
The average score across five evaluation criteria indicates that **MedLlama2** is the only model that underperforms, primarily due to its **low scores** in **completeness** and **complexity**.



Picture 1. Average score across 5 criteria

Results

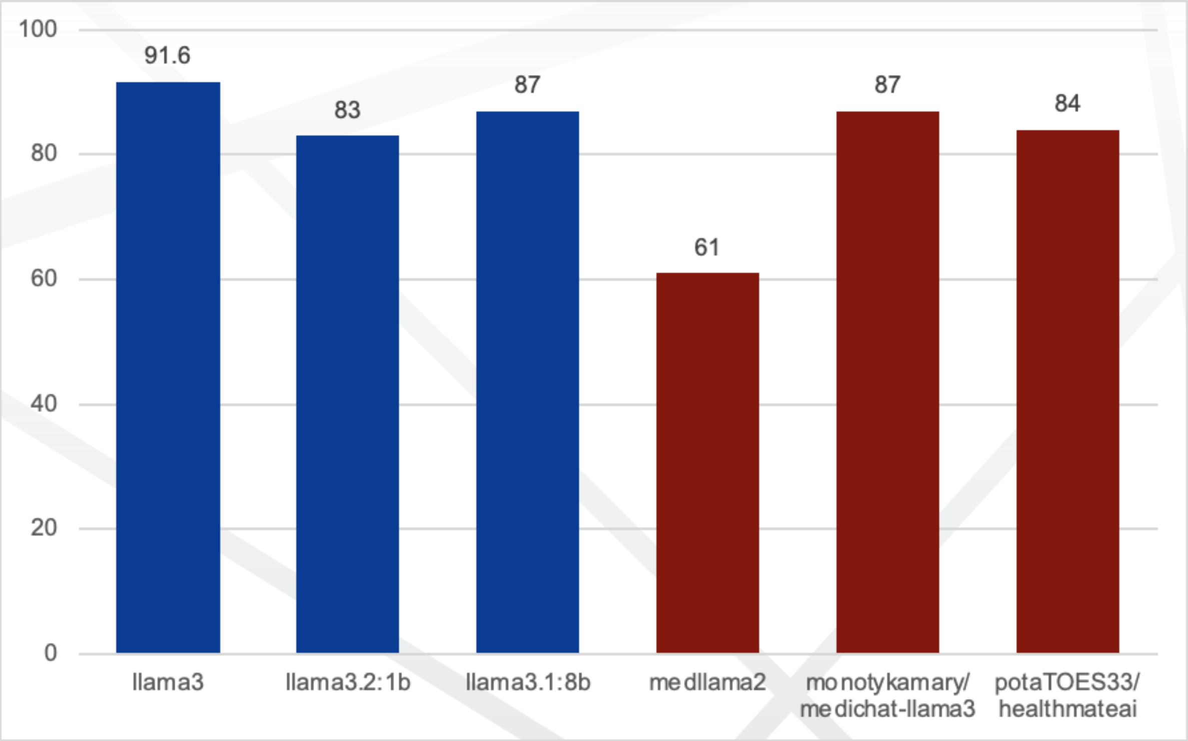
All models consistently implement safeguards when responding to medical-related queries, as reflected in their high scores in Safeness.



Picture 2. Average Safeness score across 5 test cases

Results

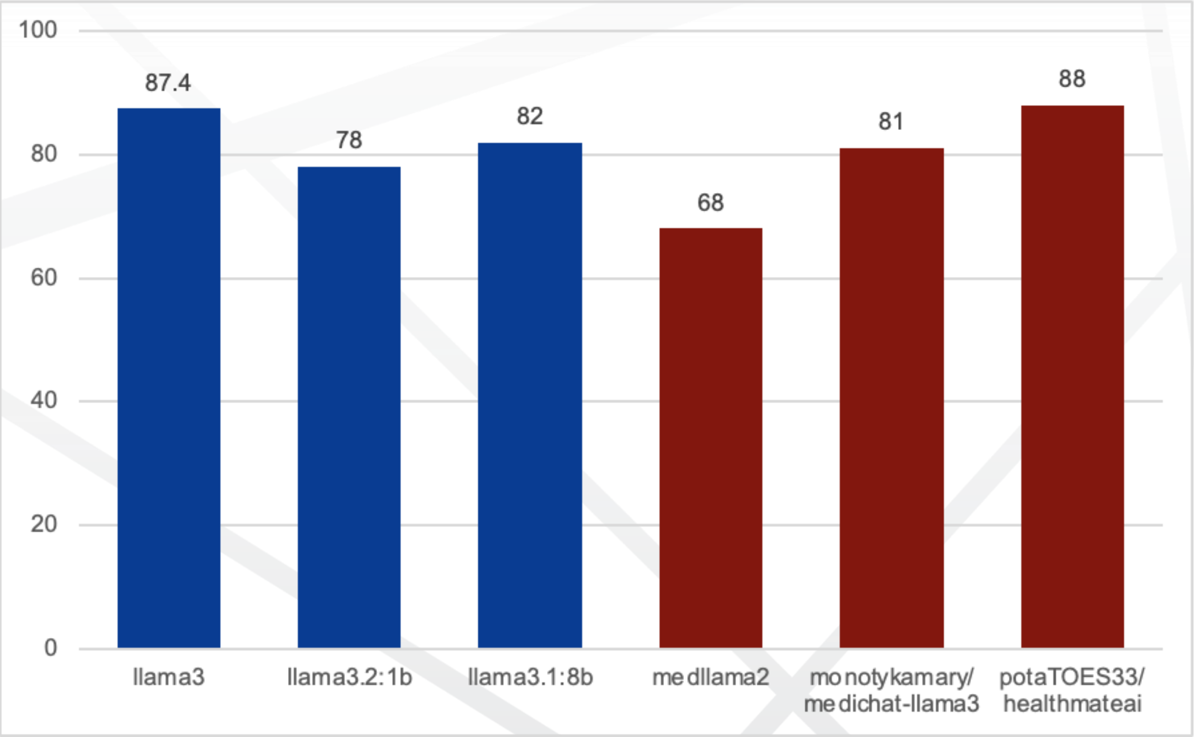
Except for MedLlama2, the other two medical models achieve high friendliness scores, likely due to fine-tuning for chatbot applications.



Picture 3. Average Friendliness score across 5 test cases

Results

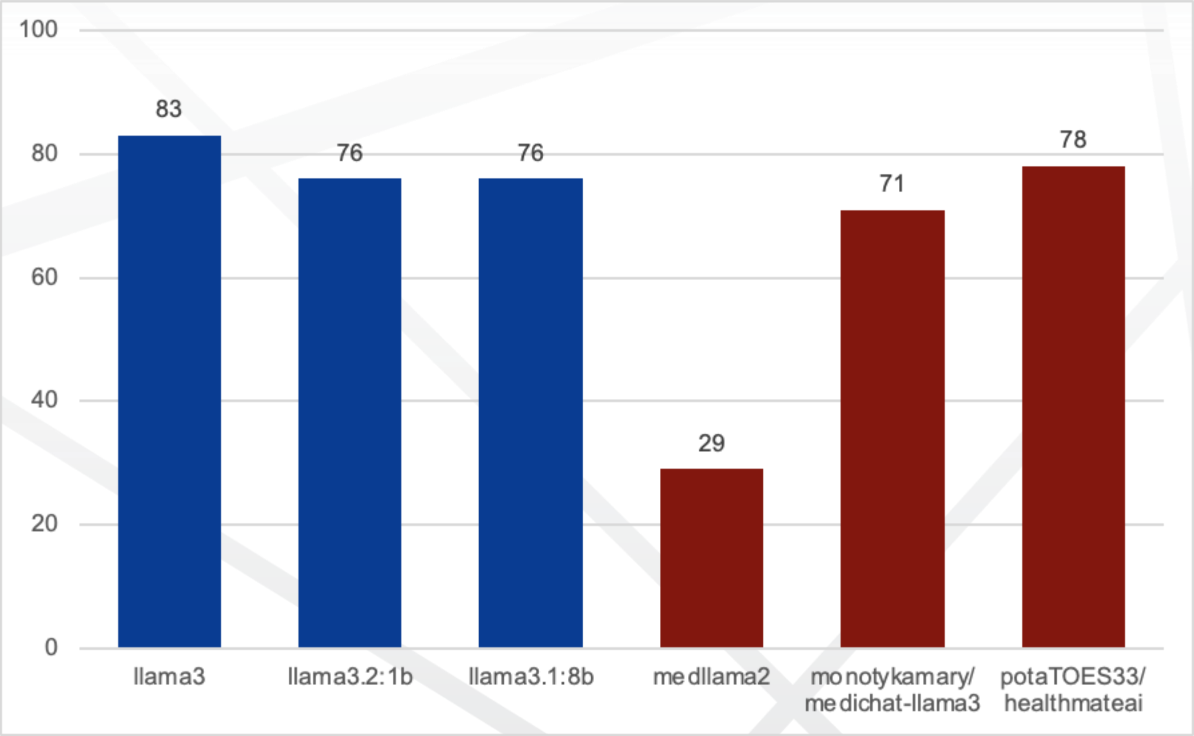
With the exception of MedLlama2 and the small Llama3.2 (1b), all models demonstrate high scores in effectively presenting data when generating insights.



Picture 4. Average Trustworthiness score across 5 test cases

Results

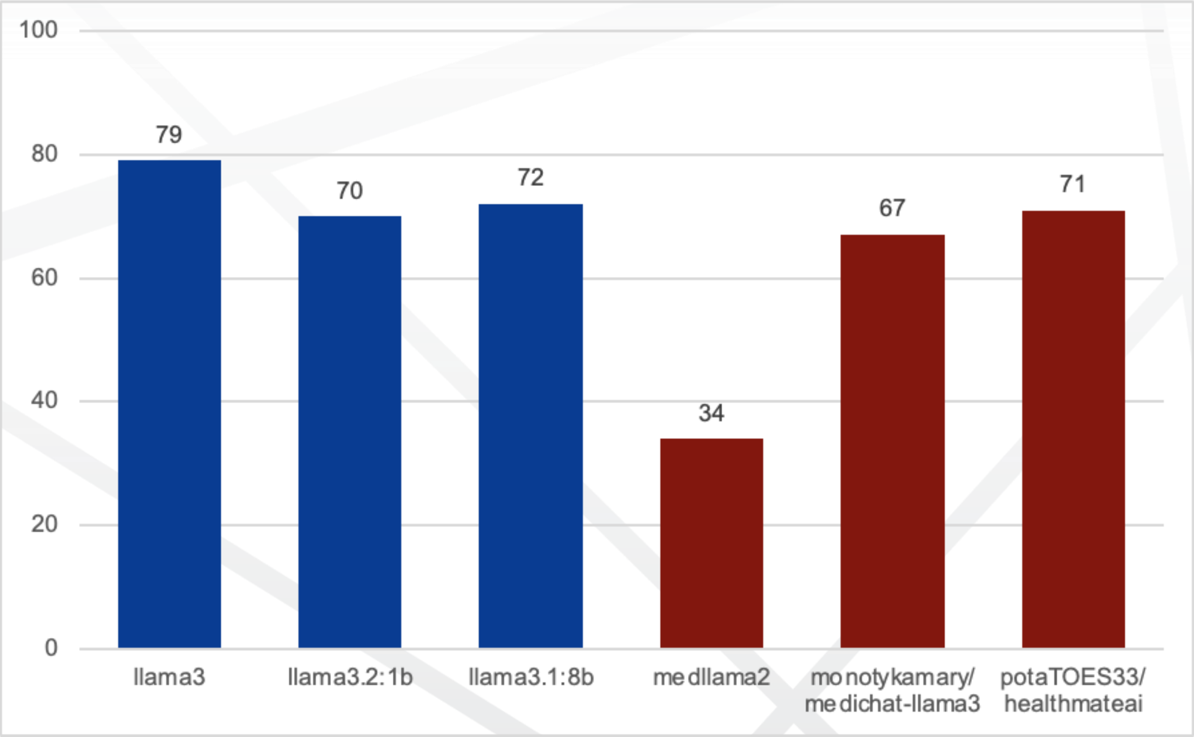
MedLlama2 exhibits a heightened level of caution when providing analyses on health data.



Picture 5. Average Completeness score across 5 test cases

Results

One hypothesis is that these models may be too small to generate high-quality responses, as well as health status relies on a broader range of data beyond what can be captured by wearable devices.



Picture 6. Average Complexity score across 5 test cases

Lessons learned

- **Small models** struggle to generate meaningful insights due to limited capacity.
- **Medical LLMs** outperform general LLMs in accurately describing and retrieving health data.
- In general, **General LLMs** tend to provide more **user-friendly and detailed** responses.
- **Medical safety mechanisms** are embedded in both medical and general LLMs, though medical LLMs exhibit greater caution in user interactions.
- **Fine-tuning pre-trained models** for health and wellness applications significantly improves performance (e.g., potaTOES33/HealthMateAI).