The viability of running a medically oriented LLM on low powered hardware

Derek Liu

Research Question:

How viable is it to run a medically oriented large language model (LLM) on lower-powered hardware while maintaining accuracy and efficiency?

Abstract:

This project explores how well a medically focused large language model (LLM) can run on lower-powered hardware without losing accuracy or efficiency. Using LM Studio and the Meta LLaMA 3BDistilled Medical Model, the experiment was run on a MacBook Air with an M3 chip and 16 GB of RAM. The model was tested in two modes: high-performance and low-performance. Factors such as response time, token processing speed, and accuracy were measured and compared. The results showed that while high-performance mode was faster and slightly more accurate, low-performance mode still delivered similar accuracy with only slower response times. These findings suggest that it is possible and practical to run medical AI models locally on less powerful devices, making them more accessible for personal or educational use.

Background:

Artificial intelligence (AI) is becoming more popular in the medical field, especially with tools called large language models (LLMs). These models can understand and generate text, which makes them useful for answering medical questions, explaining health topics, and even summarizing patient records. Recent studies show that advanced models like GPT-4 and Med-PaLM 2 can score very high on medical exams—sometimes even better than real doctors. For example, Med-PaLM 2 scored 86.5% on a test called MedQA, and GPT-4 reached around 90% on clinical reasoning questions, showing that these tools can understand complex medical content.

Smaller models like BioGPT and PubMedGPT have also done well, especially on more specific tasks. BioGPT, for instance, scored over86% on a medical Q&A test and worked well with data from research papers. These smaller models are useful because they don't need super powerful computers, increasing accessibility for students or people in areas with limited technology.

AI models have also been tested for writing summaries of patient notes or giving simple health advice. Some AI-written summaries were rated just as good—or better—than the ones written by doctors, but researchers still warn that AI can make mistakes or leave things out. Because of this, experts say AI should be used with human supervision, especially in real medical situations.

This project looks at whether a smaller medical AI model, the MetaLLaMA 3B Distilled Medical Model, can still run effectively on a basic laptop like a MacBook Air. If it stays accurate and fast even without high-performance hardware, it could help make medical AI more accessible for everyday use, especially in schools, clinics, or places without strong internet.

LLaMA by Meta LM Studio

Acknoledgements:

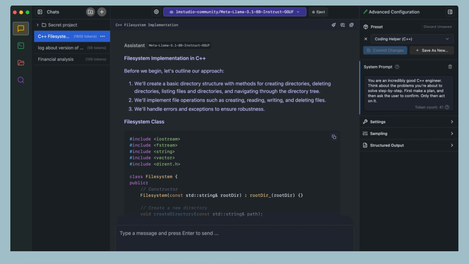
I would like to express my profound appreciation to my research teachers, Jean Armstrong and Jaime Polk, in addition to Andrew from Da Vinci's Donuts in Alpharetta for the immensely helpful guidance and mentoring from the very beginning to the end of this research project.

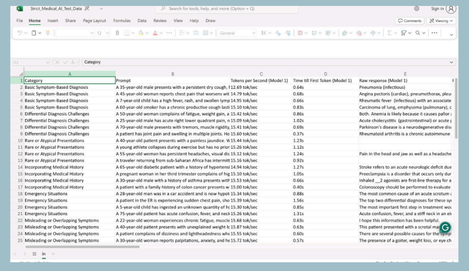
Materials:

For this project, I used LM Studio to run the Meta LLaMA 3B Distilled Medical Model, which is a smaller version of a large language model designed to answer medical questions. I chose this model because it's made to be accurate with medical information but also small enough to run on regular computers. I ran it on a MacBook Air with anM3 chip and 16GB of RAM, which is not a super powerful machine, but it's something a lot of students or people might have. I tested the model in two different performance settings using LM Studio. One was "half hash rate," where the computer wasn't using all of its power, kind of like when you're trying to save battery or keep it cool. The other was "full hash rate," where the computer used all of its available processing power. I asked the model a series of medical questions in both modes and compared how long it took to answer and whether the answers stayed the same. This helped me figure out if it's possible to use advanced medical AI on lower-powered devices while still getting accurate and quick responses.

Methodology:

To test how well the AI model worked under different performance settings, I used LM Studio to run the Meta LLaMA 3B Distilled Medical Model on a MacBook Air with an M3 chip and 16 GB of RAM. I ran the model twice—once in high-performance mode and once in low-performance mode. For each test, I asked the model the same set of medical questions. After getting the responses, I copied all the results into an Excel spreadsheet. Then, I measured and recorded how long the model took to give the first part of its answer (response latency), how fast it processed tokens (tokens per second), and how accurate the answers were. I gave each result a score based on how correct and clear the answers were and compared the scores and times between the two performance modes.





Problem Statement:

The increasing reliance on large language models in the medical field presents a challenge due to their high computational demands, which typically require expensive, high-powered hardware. This study aims to address this issue by testing the viability of deploying medical LLMs on lower-end hardware, evaluating performance trade-offs, and identifying solutions for more accessible and sustainable AI-driven healthcare.

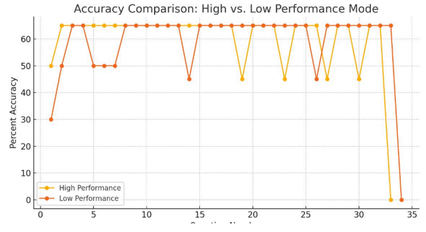
Hypothesis and Criterias for Success:

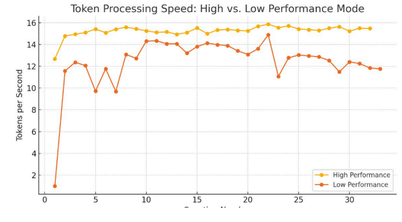
I believe that changing the performance setting will affect how fast the model responds, but it will not change how accurate the answers are. This is because the same medical language model is being used in both cases: only the amount of computing power is different. In this experiment, the independent variable is the hash rate setting (half or full), and the dependent variables are the speed of the model's response and the accuracy of its answers. My hypothesis was that the model will respond slower on the half hash rate setting, but the accuracy of the medical information it gives will stay the same.

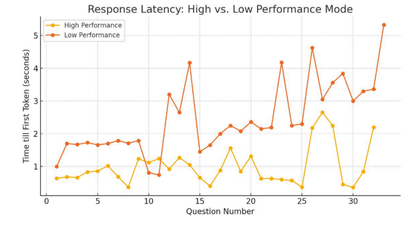
The test will be considered successful if the model gives accurate medical answers in both performance modes and the difference inresponse time between half and full hash rate can be clearly measured.

Results:

After testing the AI model in both high-performance and low-performance modes, the results showed that the high-performance mode was faster and slightly more accurate. On average, it responded in about 0.999 seconds, while the low-performance mode took around 2.462 seconds. Token processing speed was also better in high-performance mode, with 15.25 tokens per second compared to 12.42 tokens per second in low-performance. For accuracy, high-performance mode scored 60.15%, and low-performance scored 59.12%, showing only a small difference which should account for margin of error.







Conclusion:

In this project, I tested how well a medically focused AI model could run on lower-powered hardware by comparing its performance in two different modes: high-performance and low-performance. The results showed that while the model responded faster and was slightly more accurate in high-performance mode, the accuracy in low-performance mode was still very close. This means that the AI model can still be used effectively on deviceswith limited computing power, like a MacBook Air, especially for non-urgent tasks. Although the slower response time in low-performance mode might not be ideal for real-time situations, the model's reliable accuracy makes it a viable option for local, offline medical support when high-end hardware isn't available. This proves that smaller, distilled models can be useful even on everyday devices.

Next Steps:

Now that this project has shown a medical AI model can run on lower-powered hardware with similar accuracy, there are several ways to expand the research. One next step is to test the model with a wider variety of medical questions, including more complex or urgent scenarios, to better evaluate its accuracy and reliability. This would help demonstrate the capability of medical AI across different scenarios.

Another step would be to compare it with other models, like larger versions of LLaMA or different medical AIs such as Med-PaLM, to see how they perform under the same conditions. It would also be helpful to run the tests on other devices—like older laptops, tablets, or phones—to see how the hardware affects speed and accuracy.

Future tests could also use automated tools to score the model's answers, which would allow for larger-scale testing and more consistent results. Lastly, gathering feedback from real users such as medical students or professionals could help determine how useful the model is in real-life situations. These steps would help show the full potential of using local AI for medical support and learning on everyday devices.

Citations:

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal, 8(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095