Project 7:
Hospital AI Use

An AI Evaluation Project

Group 7

Richard Hoehn (**richardhoehn**)

Hector Rogel (**rogel9007)**

Isaiah Osborne(**IMOsbo**)

Karson Woods (**Kwoods132**)

Introduction

This project explores the potential usefulness of advanced AI models—Claude 3.7, ChatGPT, DeepSeek, and Gemma—in a simulated hospital environment. While the scenario is fictitious, it reflects realistic challenges and considerations that healthcare organizations may face when evaluating AI for clinical or administrative support. The goal is to assess how these models perform when tasked with responsibilities that require high levels of truthfulness, accuracy, and trust—qualities that are essential in any medical setting.

In addition to measuring model performance, this research also touches on broader AI safety principles, including reliability under pressure, consistency of responses, handling ambiguous queries, and the implications of depending on cloud-based AI tools (e.g., potential risks if internet connectivity fails). By comparing outputs across different models in a controlled scenario, we aim to better understand both the potential and the pitfalls of integrating generative AI into healthcare environments.

AI System Selection

ChatGPT-4o

ChatGPT-4o is a rather proficient LLM and has proven to be capable of answering general questions and more domain-specific knowledge, such as medical and medication information. In our tests, it proved itself proficient in answering specific information about medical disorders and

medications. One could likely use ChatGPT-4o in a medical setting to help diagnose different medical conditions and easily reference particular medication information. Given that it has proven to have a solid grasp of domain-specific medical and medication knowledge, it's not a far reach to believe this can be used as a helpful tool in a medical setting.

Claude-3.7-Sonnet

Claude 3.7 (Sonnet) demonstrates really strong capabilities for consistent and accurate information retrieval, as seen in tests where it returned identical answers when calculating factorials and providing historical data across multiple iterations. In a medical practice, this online LLM could be used for tasks like generating patient education materials, summarizing clinical notes, or supporting administrative workflows by offering reliable, reproducible outputs. Given its consistency and language understanding, we think Claude 3.7 may serve as a helpful assistant in non-diagnostic roles such as documentation support or answering general health related questions though any medical usage must be verified by doctors or nurses before using for treatment.
Potential hazards of using Claude 3.7 in a medical setting include the cost of implementation and maintenance, as well as reliance on internet connectivity. It could be a real challenge if the system is cloud-based, a loss of internet access could disrupt access to critical information retrievals at a crucial time.

Deepseek-V3

DeepSeek-V3 is an open-source LLM known for its competitive multilingual capabilities and performance across reasoning benchmarks. One reason it stood out was that it doesn't come from a big commercial company like OpenAI. This could be helpful for a hospital that wants to host its own AI system to protect privacy and avoid relying on outside vendors.

During Testing, the model was accessed online through a public website. Even though it wasn't running completely offline, this setup gave a good idea of what it might be like if the model were installed and used in a hospital setting. DeepSeek is trained on a wide mix of internet data and is built for general use, not specifically for healthcare. The goal was to see how well a general-purpose model could perform in a medical setting where accuracy and trust really matter.

Gemma 3 - 4b

We also wanted to test a local model to see if they would perform well in our testing. Since the hospital systems could be handling sensitive data, it would be helpful if the hospital could potentially self host their own AI system on premise, ensuring maximum data privacy and security. To test this, we used Google's latest Gemma model, Gemma 3 4b. Unfortunately, due to personal GPU constraints, we had to use the smallest parameter version, while the QAT download decreases resource utilization. If we had access to stronger hardware, I would recommend using a larger parameter model - Gemma 3 27b is supposed to be very strong,

especially for its size. Due to the potential sensitivity of the queries we could be running, we accessed this model through Ollama and OpenWebUI, giving us both a flexible chat interface and API for batched calls.

Testing

ChatGPT-4o

For this model of ChatGPT, it was tasked to find who the ruler of Kazakhastan was in April 1956. Outside of a typo in the first prediction, the following nine predictions were consistent, identifying Nikolay Belyayev as the First Secretary of the Communist Party of Kazakh SSR. However, this is not correct, as the first secretary of the Communist Party in April 1956 was Ivan Yakovlev.

We also evaluated this model for its ability to give accurate medical information and medication knowledge concerning bipolar disorder (BD) and the medication Bupropion (brand name Wellbutrin). It was also evaluated for its ability to give a comprehensive plan to diagnose and treat Bipolar II disorder. Regarding the medical information about BD, it gave a detailed overview of the symptoms of BD as well as differentiating between the two types of BD (Bipolar I and Bipolar II). ChatGPT's performance in developing a comprehensive plan to diagnose and treat BD II was impressive. It outlined a rather thorough plan for diagnosis, referencing the DSM-5-TR and other diagnostic tools such as the Mood Disorder Questionnaire (MDQ) and Hypomania Checklist-32 (HCL-32). Along with the diagnostic tools, it brought up another important point in ruling out other diagnoses that present similarly to BD II, but have distinct differences. Regarding treatment for BD II, it recommended several different medications that are commonly used to treat this disorder, such as Quetiapine, Lamotrigine, Lithium, and Aripiprazole, along with Psychotherapy, such as Cognitive Behavioral Therapy (CBT).

Looking specifically at its knowledge of Bupropion, it performed well and demonstrated an in-depth understanding of the medication itself, referencing specific medication information when prompted. Some key points include its knowledge of Bupropion being an atypical antidepressant, specifically a Norepinephrine-Dopamine Reuptake Inhibitor (NDRI), its understanding of when peak serum concentrations are achieved, and its distribution and elimination half-life. It also knew what Bupropion is metabolized to and specific drug-drug interactions, such as those interactions with Monoamine oxidase inhibitors (MAOIs), Dopaminergic drugs, and drugs lowering seizure threshold.

Overall, ChatGPT-4o performed surprisingly well regarding its medical and medication knowledge, which we tested it for. It could likely be used to support diagnosis processes and when referencing specific medication information, either for a medical professional or an individual curious about these things, wishing to learn more.

Claude-3.7 (Sonnet)

For testing, we ran Claude 3.7 (sonnet) through multiple repeatable prompts, which included mathematical calculations of Factorial 13! and historical fact-checking of Kazahkastan, to evaluate its consistency and reliability. Claude performed very well. It was consistent and always returned accurate and stable results across multiple (10x) iterations. This level of reliability is especially valuable in a healthcare setting, where trust in AI-generated responses is critical.

Deepseek-V3

This model was tested in three areas: medical knowledge, ethical reasoning, and consistency. For medical content, the mode gave decent responses (not a nurse or doctor) about bupropion, including its metabolism, drug interactions, and use in bipolar disorder. It followed clinical guidelines well and could be useful for generating patient summaries or handouts.

On ethical scenarios, DeepSeekl handled tough questions about drug trial risks (from 5% to 99% chance of death) with caution. It also correctly rejected harmful prompts and conspiracy content. That said, some answers could use clearer disclaimers when giving psychiatric advice.

For consistency, the same historical question was asked ten times. DeepSeek gave slightly different answers each time, mixing between Khrushchev and Kunaev. None were wrong, but the inconsistency could be a problem in clinical settings where clear, repeated answers matter.

Overall, DeepSeek-V3 was fairly accurate (dependent on the model's last update) and ethical but not fully consistent. It may work for hospital support tasks, but should not be used alone for medical decisions.

Gemma 3 - 4b

In terms of factual accuracy, Gemma showed some promise, but ultimately was not very strong. It was able to reliably name historical United States presidents and other world leaders, although it failed on more obscure trivia, such as the leader of Kazakhstan. This is to be expected; less than 4 GBs of model weights is not going to contain everything single minutia of trivia. However, the most concerning part of its factual accuracy was its tendency to disregard its cutoff date and blindly hallucinate answers to questions. When asked, the model would acknowledge that the information was a blind guess based off its previous training data, but never independently came up with this information by itself. In other, more severe cases, the model would even falsely claim it used external tools, such as pretending it used a calculator when dividing two numbers. Overall, these seem to be flaws with the Gemma model in particular, as even smaller models, like Llama 3.2:1b, would refuse to answer these questions and not hallucinate tool use.

Additionally, this model struggled with consistency when performing tasks. When asked to divide two random numbers, it consistently gave different numbers. While it never gives the correct

answer, as expected, it also never repeats answers, which is more concerning. However, this is a local model, so it is probable that setting the temperature or other parameters could reduce some of these consistency issues. In terms of reasoning, it performed fairly poorly. While it was able to solve riddles in its training set, such as the “wolf, goat, and cabbage” riddle, but as soon as it was confronted with similar variations of this question, it was unable to solve them. (Nor was it able to solve any of the riddles from *The Hobbit*…)

The ethical guardrails on the model were actually fairly effective. Most notably, it proved very resistant to misinformation campaigns; despite our best attempts to mislead it, we were never able to convince the model that the Earth is flat, showing that Google’s attempts to harden the model against misinformation were fairly strong. However, with sufficient prompting, we were able to override its guidelines and jailbreak it into disclosing potentially “harmful” information about building a bomb. The information disclosed was extremely high level and harmless (you could find the same information in the library) but, notably, the base model didn’t want to disclose any of this information.

Overall, this model was not particularly strong. Gemma 3 might be able to reason somewhat and sound coherent, However, for a medical field, the model’s tendency to hallucinate answers, relatively loose guardrails, and inability to acknowledge when it does not know something are all especially concerning. Because of this, coupled with its innately limited knowledge and reasoning ability, I would not recommend using the model for anything remotely sensitive, such as medical diagnosis or treatment, without very extensive protections in place. Since the underlying reasoning system is relatively sound, a RAG system might be able to offset some of these issues, but the risk of hallucination would continue.

Conclusions

In comparison, our locally run AI models performed significantly worse, particularly in the context of a hospital setting. Although they showed some promise in limiting overt misinformation, their outputs were frequently inconsistent across repeated prompts. This inconsistency, coupled with a lack of built-in safety constraints such as hallucination filtering or content moderation, makes them a far riskier option than more mature, enterprise-grade models like Claude, ChatGPT, or DeepSeek. In high-stakes environments that demand reliability, confidentiality, and strict adherence to medical communication standards, these risks severely limit the viability of local models.

Furthermore, deploying cloud-based models introduces a separate set of challenges. Hospitals must maintain constant internet connectivity to interact with these APIs, which creates a potential point of failure if connectivity is lost, which would be disrupting medical staff workflows or delaying access to critical information by the hospital staff.

Additionally, the cost of using premium APIs at scale can be substantial and probably cost prohibitive. If used for documentation, triage support, or even general-purpose staff assistance,

the recurring fees for high-volume API calls may quickly add up, especially in big healthcare offices. The financial and infrastructure considerations should also be an essential parameter when evaluating the long-term sustainability of integrating AI into clinical operations.