



adam fendri

Medical Chatbot



<https://github.com/adam-fendri/MedicalChatbot>





adam fendri

02

Project Overview

This chatbot is based on a fine-tuned version of Gemma2-2B that came out two weeks ago with connected to a rag system and two other models(NER model & Summarizer model)

🔍 <https://github.com/adam-fendri/MedicalChatbot> 🎧



adam fendri

Problem Statement

LLM's require a lot of computational power mainly powerful GPU's with a lot of VRAM for fine-tuning



Methodology



Creating scripts
& training other
models



Fine-tuning the
main model on
[runpod.io](#)



Connecting the
models & testing

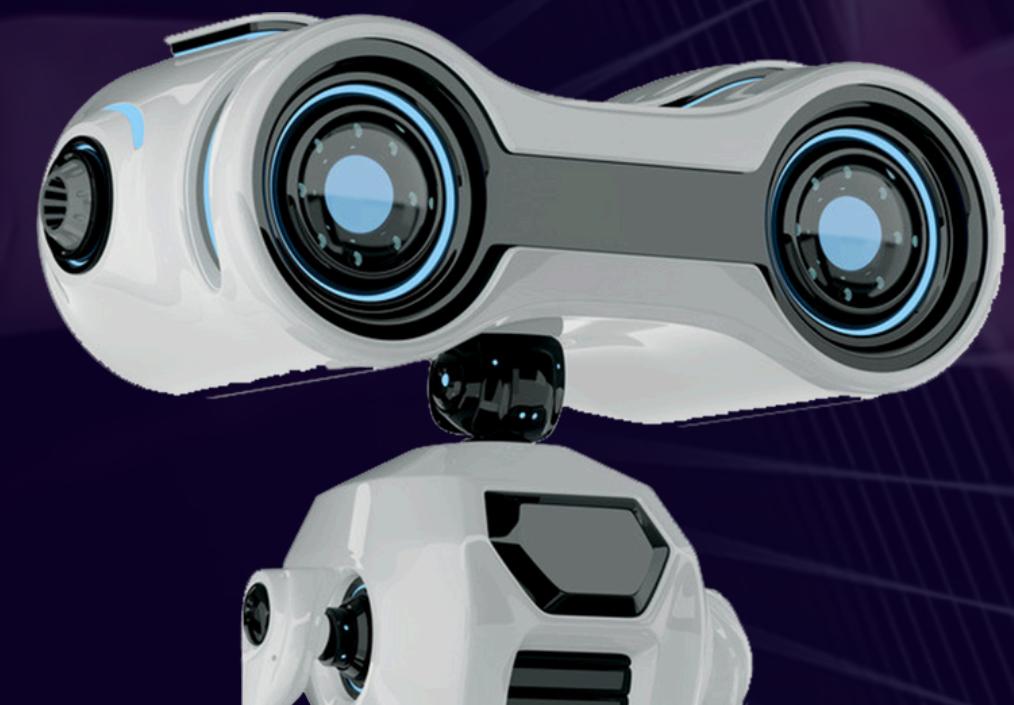


adam fendri

OS

NER Model

Created a custom dataset for my task with proper tagging with the help of chatGPT4, and trained Roberta-Large on it. Ended up with over 90% F1 score





adam fendri

06

BART & T5

For conversation summary, I used an already existing model T5 or BART for my case without any training because of the lack of data.

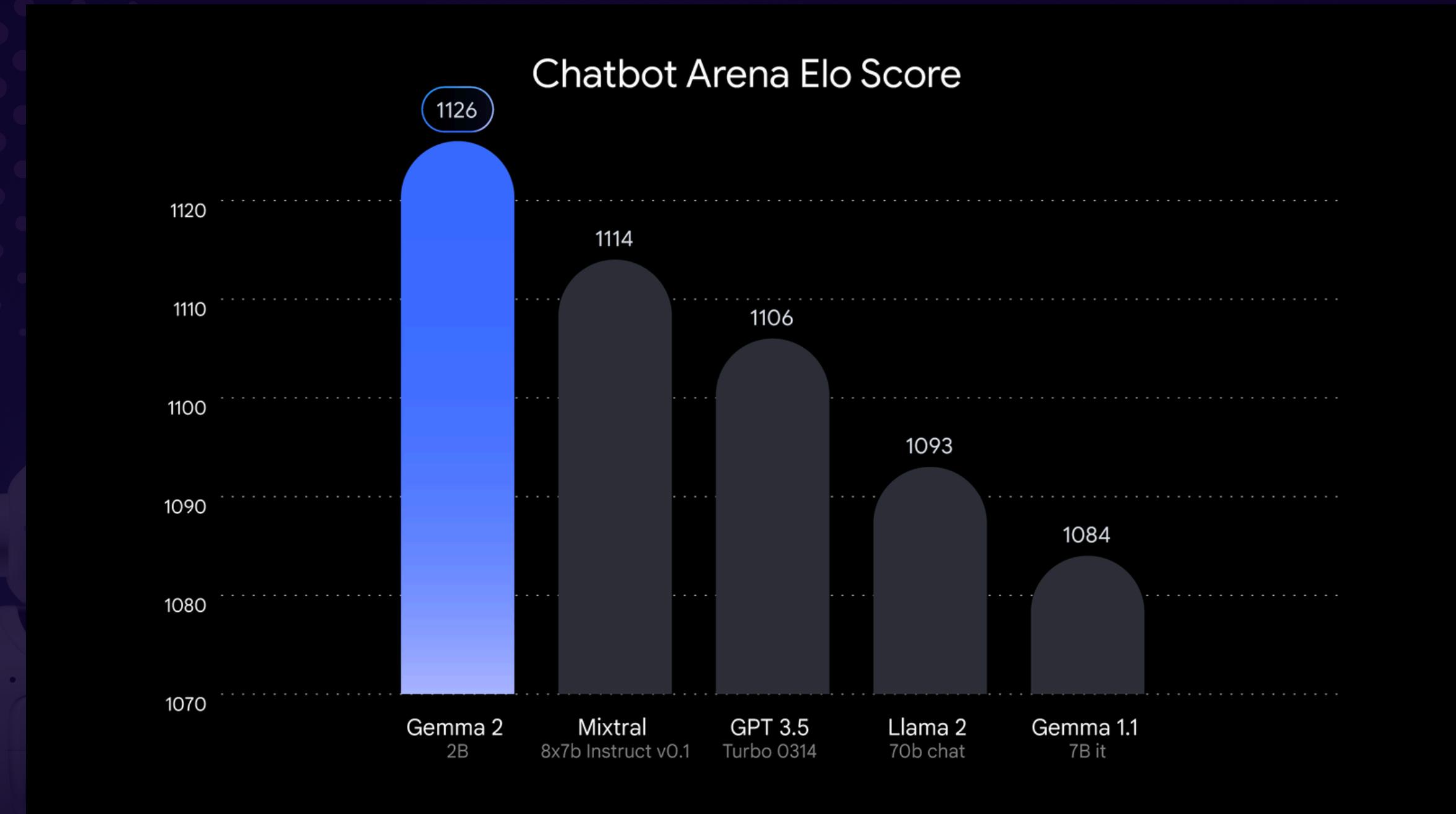


<https://github.com/adam-fendri/MedicalChatbot>



Gemma2-2B

OP



2 weeks ago , Google released a 2Billion version of their open source model Gemma2, punching way over its weight with a way smaller size compared to competition

Runpod.io

OP

Select an instance

GPU CPU Secure Cloud Network Volume Any Filter

Filter GPUs by VRAM

Any 16 24 48 80 144 240 336 432 564 640 752 960 1536

AMD

MI300X \$3.99/hr
3.39/hr
192 GB VRAM 8 max
283 GB RAM • 24 vCPU Low

NVIDIA Latest Gen

H100 SXM \$3.99/hr
3.39/hr
80 GB VRAM 8 max
125 GB RAM • 16 vCPU High

H100 PCIe \$3.29/hr
2.80/hr
80 GB VRAM 8 max
176 GB RAM • 16 vCPU High

L40S \$1.03/hr
0.88/hr
48 GB VRAM 8 max
62 GB RAM • 16 vCPU High

RTX 6000 Ada \$1.03/hr
0.88/hr
48 GB VRAM 8 max
62 GB RAM • 14 vCPU High

RTX 4090 \$0.69/hr
0.59/hr
24 GB VRAM 8 max
30 GB RAM • 8 vCPU High

H100 NVL \$3.69/hr
3.14/hr
94 GB VRAM 8 max
180 GB RAM • 16 vCPU Medium

RTX 2000 Ada \$0.28/hr
0.21/hr
16 GB VRAM 8 max
31 GB RAM • 6 vCPU Medium

L40 \$0.99/hr
0.84/hr
48 GB VRAM 8 max
250 GB RAM • 16 vCPU Low

Runpod.io

OP

Your Pod Templates

Official

-  **RunPod Pytorch 2.1**
runpod/pytorch:2.1.0-py3.10-cuda11.8.0-devel-ubuntu2... [🔗](#)
-  **RunPod Pytorch 2.2.0**
runpod/pytorch:2.2.0-py3.10-cuda12.1.1-devel-ubuntu2... [🔗](#)
-  **RunPod Pytorch 2.1.1**
runpod/pytorch:2.1.1-py3.10-cuda12.1.1-devel-ubuntu22... [🔗](#)
-  **RunPod Pytorch 2.0.1**
runpod/pytorch:2.0.1-py3.10-cuda11.8.0-devel-ubuntu2... [🔗](#)
-  **RunPod Pytorch 1.13.1**
runpod/pytorch:1.13.0-py3.10-cuda11.7.1-devel-ubuntu2... [🔗](#)
-  **RunPod Desktop**
runpod/kasm-docker:cuda11 [🔗](#)
-  **RunPod VS Code Server**
runpod/vscode-server:0.0.0 [🔗](#)
-  **RunPod Disco Diffusion**
runpod/discoart:web [🔗](#)

Fine tuning results



Step	Training Loss	Validation Loss
1000	2.485000	2.294738
2000	2.303400	2.258224
3000	2.262700	2.226064
4000	2.240100	2.198877
5000	2.228300	2.270314
6000	2.196600	2.201096
7000	2.196000	2.249150
8000	2.180900	2.180279
9000	2.175700	2.135563
10000	2.177300	2.134840
11000	2.162700	2.143120
12000	2.154300	2.131050
13000	2.145900	2.096179
14000	2.145700	2.141682
15000	2.151500	2.190103
16000	2.142800	2.105605
17000	2.147500	2.121175
18000	2.135800	2.153626
19000	2.115200	2.106123



adam fendri

06

RAG System

I used a FAISS database for storing the data and the vectors, and for embeddings I used the HuggingfaceEmbedding function



<https://github.com/adam-fendri/MedicalChatbot>



Final product

Question: I weigh 70 kg, my height is 175 cm. I am 25 years old. Is it considered healthy for my age?

Answer:

Hi, Thanks for writing in. You are a healthy young man. You have a normal weight and height. You are not overweight and your BMI is 22.5. This is a normal weight for your height and age.

NER Entities Detected: {

```
"age": ".",
"height": "height is",
"weight": "70 kg"
```

Question: I weigh 70 kg, my height is 175 cm. I am 25 years old. Is it considered healthy for my age?

Answer:

Hi, Thanks for writing in. You are a healthy young man. You have a normal weight and height. You are not overweight and your BMI is 22.5. This is a normal weight for your height and age.

Conversation Summary: I weigh 70 kg, my height is 175 cm. I am 25 years old. Is it considered healthy for my age? answer based on the provided context. if the context is not available, provide a general answer.



adam fendri

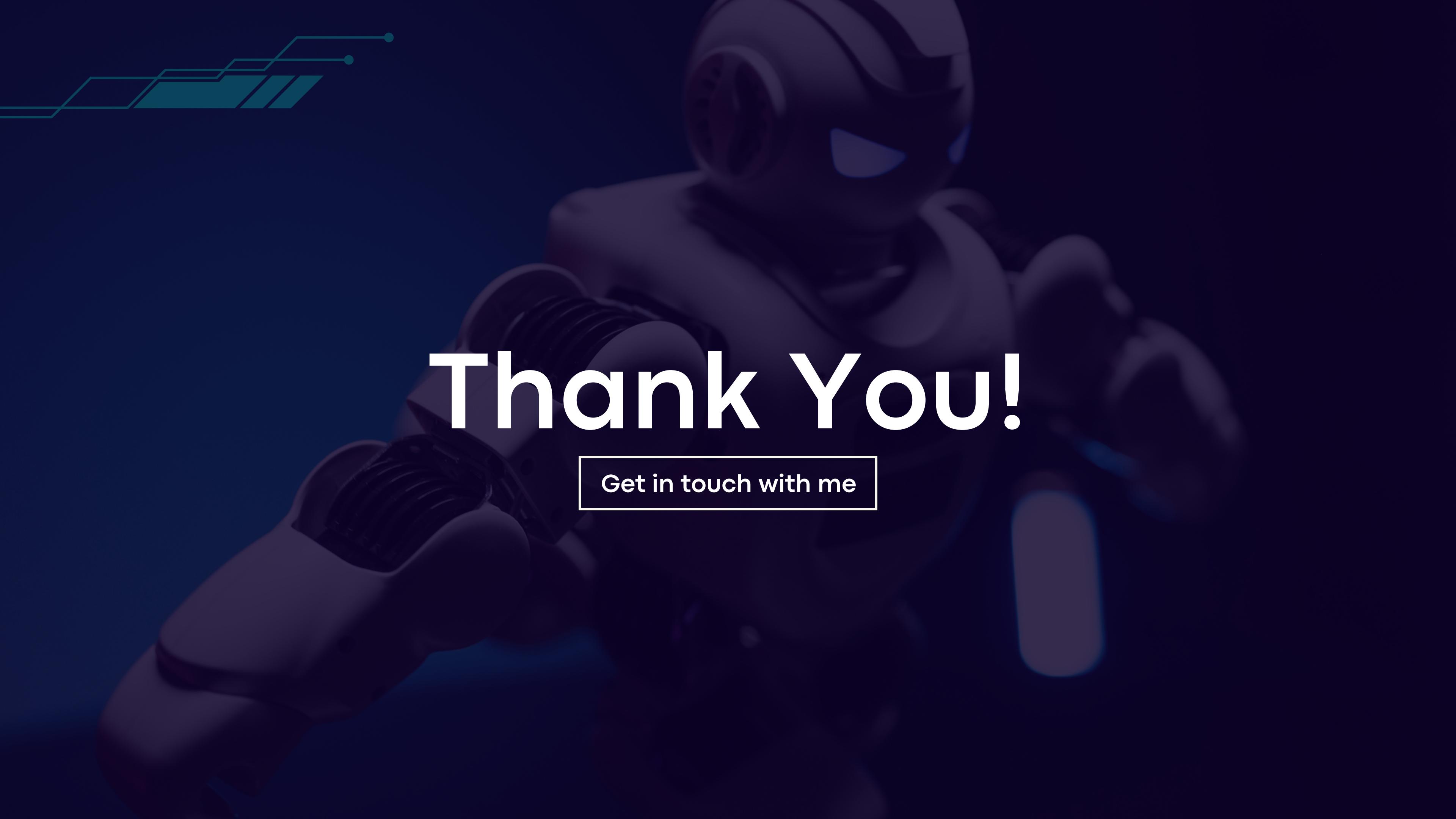
Future Directions

- Improve the NER model by augmenting the dataset, since 1000 sample is not a lot & use a better model like Dberta Large
- Switch to the latest Llama3.1 8B instead of Gemma2 2b and fine tune the LLM with more data
- Use a good source of information for the RAG system since it's the main thing that affects how the LLM responds



<https://github.com/adam-fendri/MedicalChatbot>





Thank You!

Get in touch with me