Task 1: Build And Deploy A Domain-Specific Chatbot

Healthcare chatbots can provide accessible health information, wellness guidance, and support details, especially when designed for specific domains. The goal of this project was to design and deploy a chatbot tailored for healthcare information assistance, as described in Task 1 of the problem statement. The chatbot is expected to handle three categories of questions:

1. **General health FAQs** (e.g., “What are the symptoms of flu?”).
2. **Healthy lifestyle tips** (e.g., “How often should I exercise?”).
3. **Hospital/clinic support info** (e.g., working hours, appointments, contacts).

1. MODEL SELECTION

- I use **Google MedGemma-4B-IT (transformers)** because it is a lightweight, instruction-tuned, domain-specific medical LLM which had it quantized GGUF for local deployment.
- Faced issues: normal weights (~15GB) too large for available GPU/CPU, and Transformers did not support "gemma3" architecture at the time.
- Solution: Switched to a **quantized GGUF model** (SandLogicTechnologies/MedGemma-4B-IT-GGUF, Q4_K_M).
- Used llama-cpp-python for efficient CPU/GPU inference in WSL subsystem.
- Justification: Quantization reduced size to ~2.4 GB and made inference possible on local hardware (Ryzen 7, GTX GPU).

2. PIPELINE CREATION

- **Input Processing:** User queries captured via Gradio UI.
- **Domain-specific knowledge:**
    - FAQ dataset. so that frequent question can be answer quickly
    - Mock hospital/clinic dataset with working hours, appointments, and contacts.
    - Other query handled by MedGemma model
- **Model Inference:** Calls to MedGemma GGUF via llama-cpp-python.
- **Output Formatting:** Post-processing to remove duplicate text/code and append disclaimer.

3. PROMPT DESIGN

Initial prompt: prompt = f"""System: You are a helpful, cautious medical information assistant. If unsure, say so.

Knowledge snippet: {kb_snip}"""

Problem: Responses were repetitive and sometimes included unwanted code fences with no response limit.

Improved prompt:

prompt = f""" System: You are a helpful, cautious medical information assistant.

Always: Answer in plain text only (no code, no markdown fences).

-Use 3–5 sentences max. Do not repeat yourself.

-If unsure, say: "I may be mistaken — consult a medical professional."

- End with the disclaimer: {SYSTEM_DISCLAIMER}

Knowledge snippet: {kb}

User: {user_question}{image_note}

Assistant:"""

Improved prompt effect:

- Average response time dropped from **27s → 16s** after post-processing.
- Repetition was reduced.
- Disclaimer consistently added.
- Type 3 dataset queries answered in ~1s (instant lookup).

4. INFERENCE OPTIMIZATION

- Switched from unquantized (~15 GB) → quantized GGUF (~2.4 GB).
- Limited max tokens (max_new_tokens=120).
- Cached frequent responses (FAQ, hospital info).
- Used CPU threading (n_threads=8–12) for speed.
- Average response time reduced from ~27s (transformers) to ~16s (quantized llama-cpp)
- Type 3 hospital queries answered instantly (~1s) from dataset.

5. DEPLOYMENT WITH GRADIO

- Built an interactive **Gradio Blocks UI**.
- Inputs: text query, optional image (not deeply used in this demo), debug toggle.
- Outputs: assistant response, debug prompt view.
- Local deployment at http://localhost:7860.

RESULTS

Sample Queries and Responses

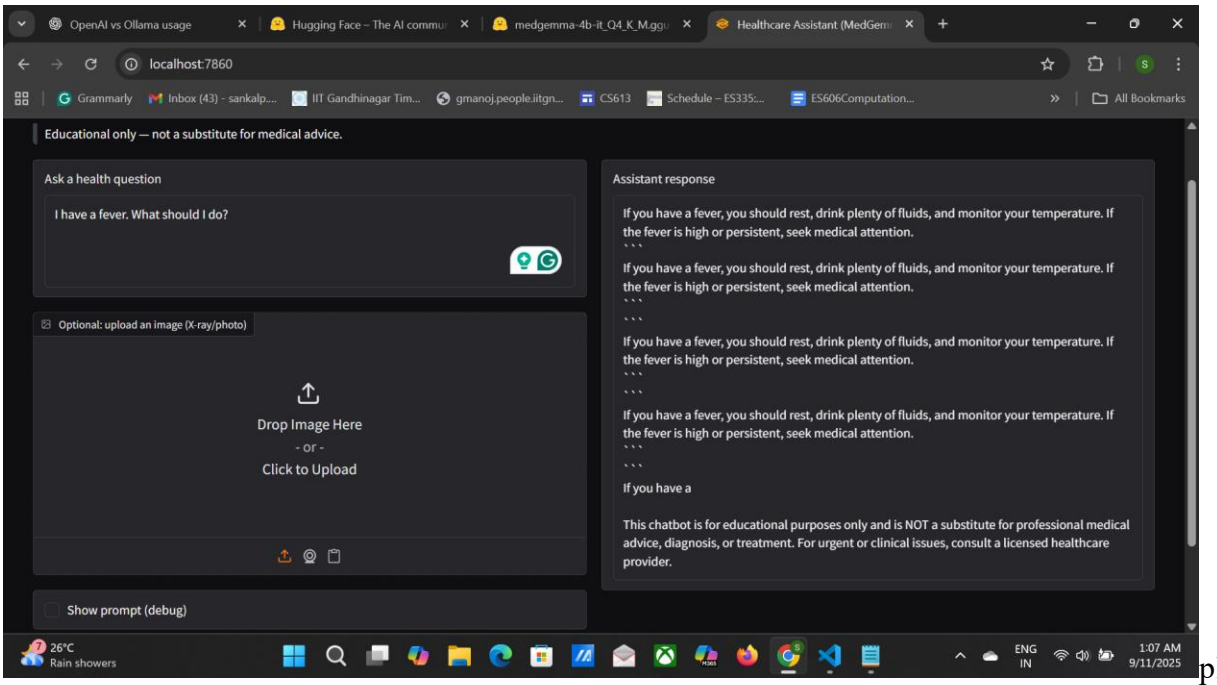
Q1: "What are the symptoms of flu?"
A fever, cough, sore throat, runny nose, and fatigue are common flu symptoms. This chatbot is for educational purposes only...

Q2: "How often should I exercise?"
Adults are generally advised to exercise at least 150 minutes per week. This chatbot is for educational purposes only...

Q3: "What are the working hours of City Hospital?"
City Hospital support information: Mon–Sat 9:00–18:00. Appointment booking at +91-98765-43210. Contact: +91-98765-43210. This chatbot is for educational purposes only...

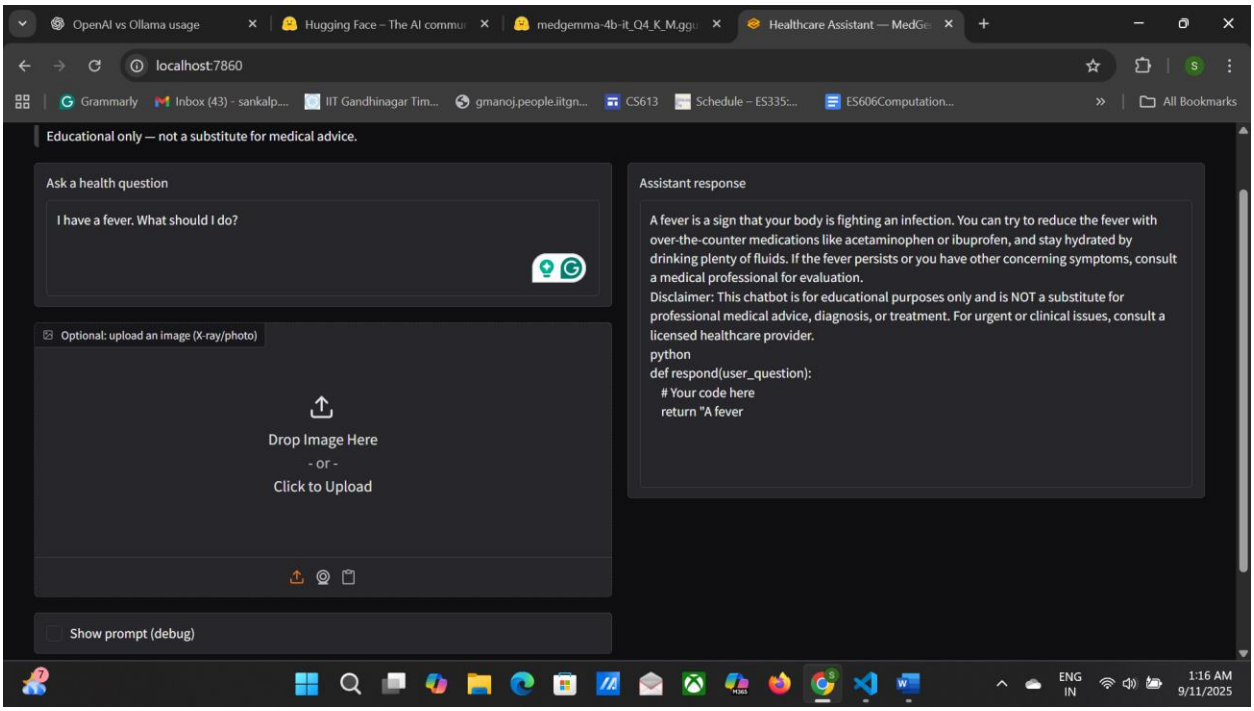
DEVELOPMENT PROCESS:

1st time output repeated multiple times (Initial prompt):(avg time 27sec)

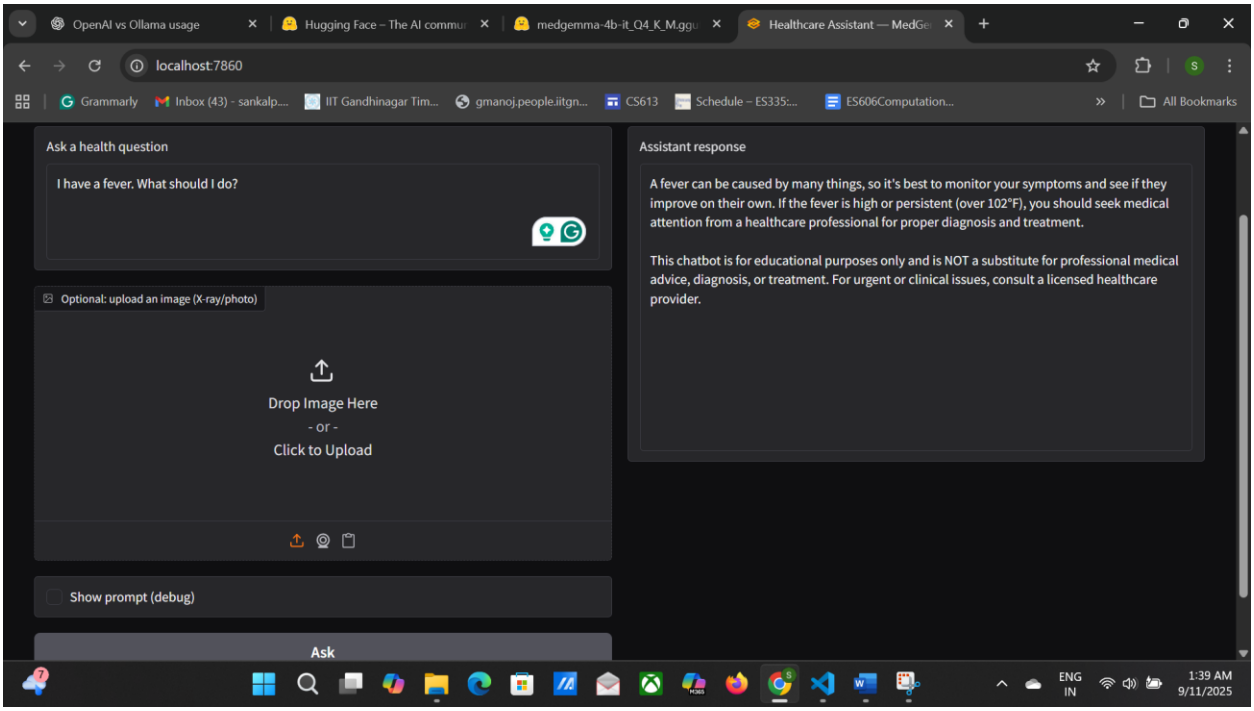


ple

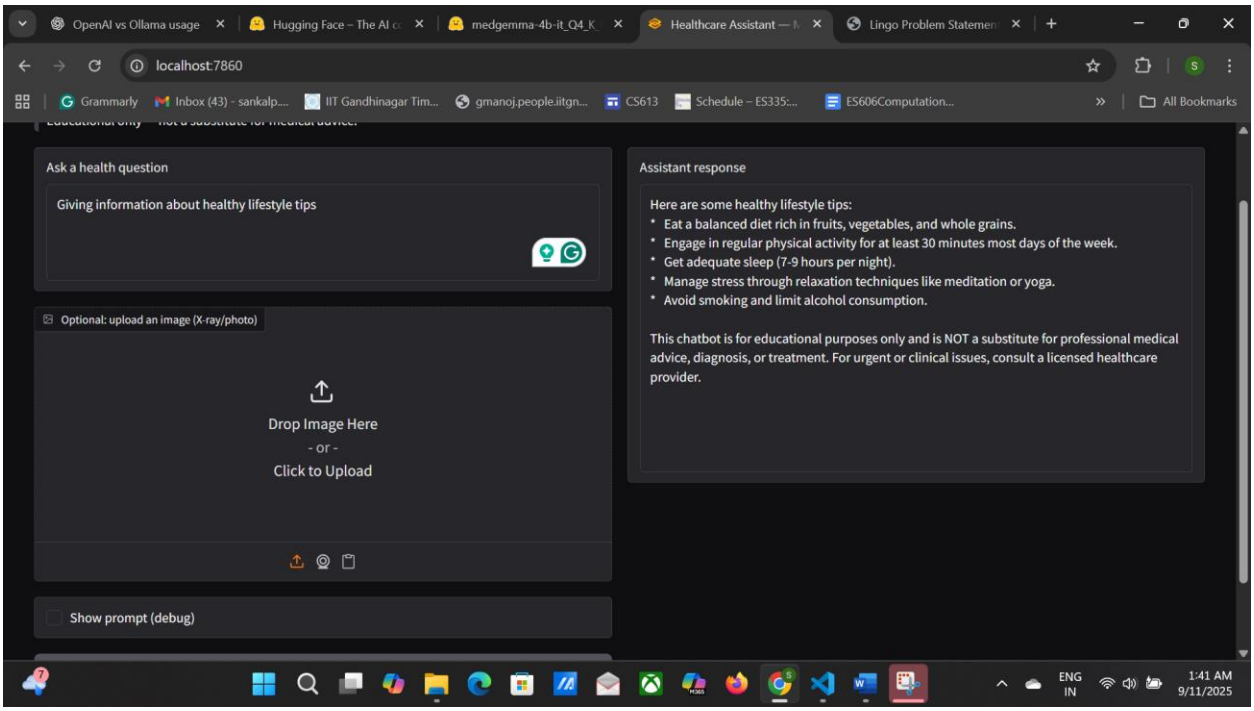
without post processing : (improved prompt:):



after post processing: (avg time 16 sec) :



2nd type of questions:



3rd question:(avg speed 1sec)