NLP Thesis Selection Task Report

Task-1

1)Model Selection

Before selecting the final model, we evaluated several transformer-based architectures for question-answering tasks. BERT was considered due to its strong performance in classification and extractive QA, but its encoder-only, bidirectional design limits its ability to generate free-form answers, and its large size makes real-time responses slower. Distilled BERT offered a smaller, faster alternative while retaining much of BERT's performance, but it still faced the same limitations in generative QA and required significant adaptation for domain-specific medical knowledge. FLAN-T5, being instruction-tuned and generative, could handle a wide variety of tasks with excellent accuracy, yet its very large size made real-time deployment computationally expensive and inefficient.

Chosen Model – Student-Teacher Setup: To meet the medical QA domain's needs for accuracy and responsiveness, we use a hybrid student-teacher approach: distilled T5 (t5-base) as the student and FLAN-T5-large as the teacher. The student is lightweight, fine-tuned on Med Quad, and handles routine queries quickly, ensuring low latency.

Handling Complex Queries: The teacher model serves as a fallback for nuanced or out-of-domain questions. FLAN-T5-large, being instruction-tuned and generative, provides detailed and accurate answers, ensuring no query is left unresolved.

Domain-Specific Knowledge: Fine-tuning both models on Med Quad equips them with medical terminology, abbreviations, and context-specific knowledge, guaranteeing relevant and reliable responses.

2)Pipeline Working

Hybrid Chatbot Pipeline:

Input Processing: Receives a user query and checks for basic validity ,Handles greetings separately with a predefined response.

Medical Keyword Filter: Ensures the question is health-related using a set of medical keywords. On-medical queries return an educational disclaimer.

FAQ Lookup (Med Quad): First searches the curated Med Quad dataset for a matching question. If found, returns a concise answer from the dataset with disclaimer.

Generative Model (Distilled + Quantized T5): If FAQ lookup fails, the pipeline constructs a prompt with instruction + question. Tokenizes the input and generates an answer using the distilled & quantized t5-base model. Applies beam search, temperature, top-p, and repetition penalty for controlled generation. Extracts the answer portion, adds disclaimer, and returns it.

Output: Always provides the source of the answer (FAQ or AI-generated) and educational disclaimer. Ensures fast responses for routine queries (student model) and rich answers for complex/unseen queries (teacher knowledge distilled).

3)Trade-off While optimizing

Reducing Model Size (Distillation): Distilling a large model into a smaller one (e.g., FLAN-T5 → distilled T5) significantly reduces memory and computational requirements, enabling faster inference. The trade-off is a slight drop in accuracy or the richness of generated responses, especially for very complex or nuanced queries.

Quantization or Caching Frequent Answers: Quantization reduces model precision (e.g., 32-bit → 8-bit), lowering memory usage and improving inference speed. Caching frequent answers avoids re-computation for common queries. The trade-off is minor accuracy loss due to reduced precision and potential staleness if cached answers are not updated.

Limiting Response Length: Restricting output tokens ensures faster responses and prevents runaway generation, but it may truncate long, detailed answers, reducing completeness and nuance.

Overall Trade-off: These techniques prioritize efficiency, responsiveness, and scalability for real-time deployment while accepting small compromises in accuracy or answer detail.

4)Examples of Prompts and response

Prompt Engineering :-

Prompt-1 : "Instruction: Answer to the question"

Disadvantages :- It used to give answers to all non-healthcare related domain questions which is obviously not a nice thing for a General Health care QnA Chatbot.

Prompt-2 : "Instruction: Provide a concise, factual, and medical-only answer to the question. Do NOT answer non-medical questions .Include the disclaimer: "This is for educational purposes only and not a substitute for professional medical advice**"**

Example 1: Routine Medical Question (FAQ Lookup)

- **User Prompt:** "What are the symptoms of diabetes?"
- **Response:** From our curated Medical QA database: Common symptoms of diabetes include increased thirst, frequent urination, fatigue, blurred vision, and slow-healing wounds.

  This is for educational purposes only and not a substitute for professional medical advice.

Example 2: Complex Question (Model-Generated)

- **User Prompt:** “How does metformin help control blood sugar?”
- **Response:** From AI-generated medical knowledge: Metformin helps lower blood sugar by reducing glucose production in the liver and improving insulin sensitivity in body tissues.

This is for educational purposes only and not a substitute for professional medical advice.

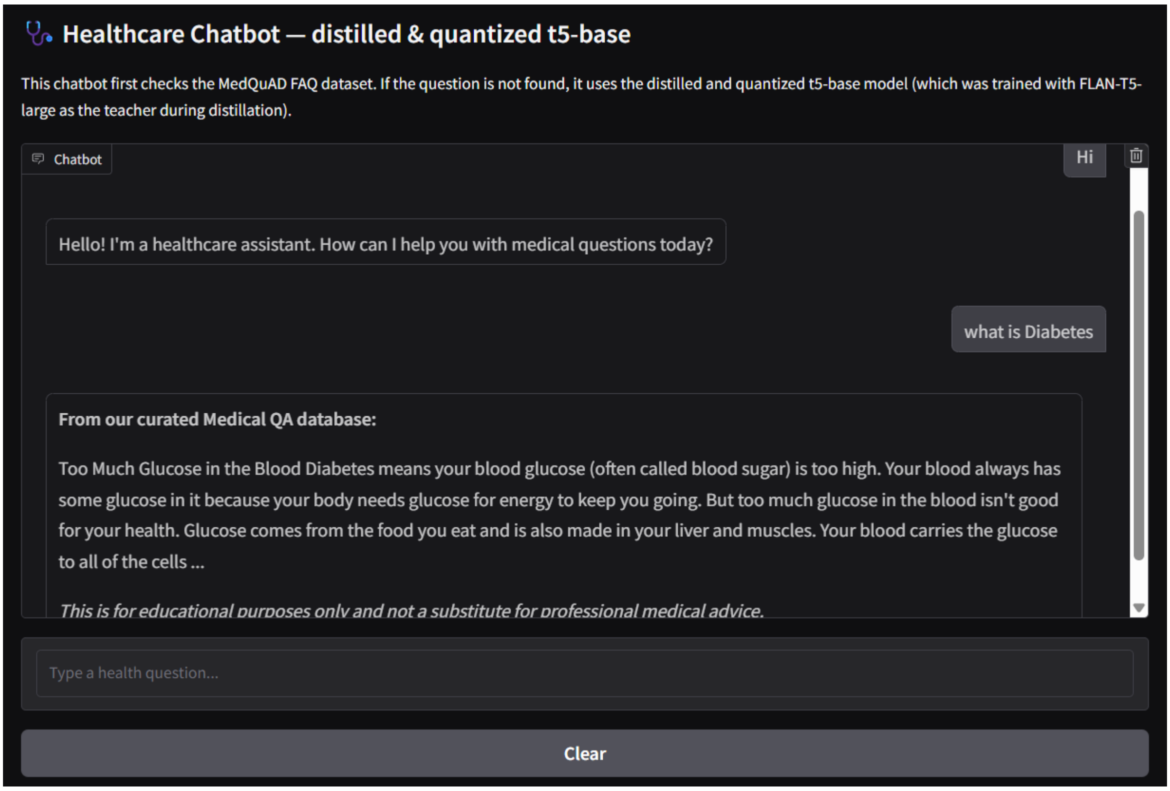
Example 3: When Non-Medical Question

- **User Prompt:** “How to buy a potato”
- **Response:** I’m sorry, I can only Answer medical or health-related Questions. Please ask a medical question.

  This is for educational purposes only and not a substitute for professional medical advice.

5)Screenshot of our Gradio App

Response from the Medical QA dataset



Healthcare Chatbot — distilled & quantized t5-base

This chatbot first checks the MedQuAD FAQ dataset. If the question is not found, it uses the distilled and quantized t5-base model (which was trained with FLAN-T5-large as the teacher during distillation).

Hello! I'm a healthcare assistant. How can I help you with medical questions today?

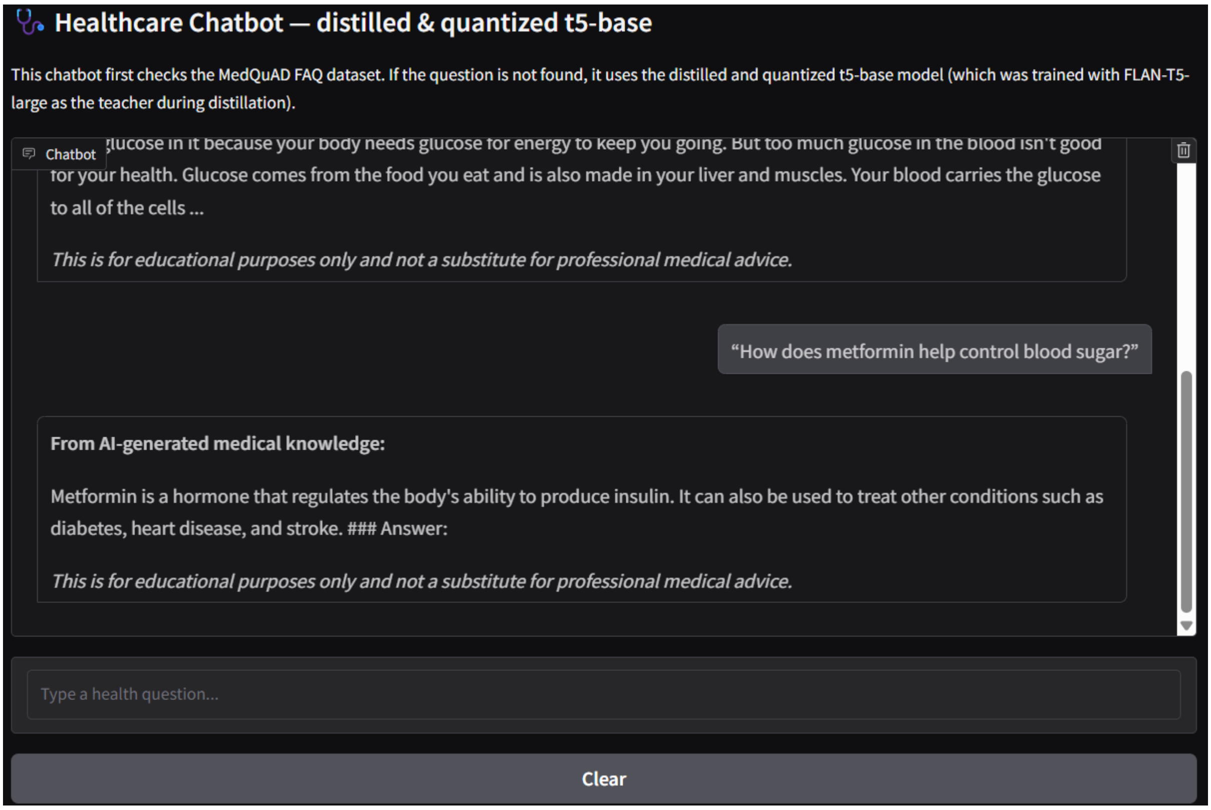
what is Diabetes

From our curated Medical QA database:

Too Much Glucose in the Blood Diabetes means your blood glucose (often called blood sugar) is too high. Your blood always has some glucose in it because your body needs glucose for energy to keep you going. But too much glucose in the blood isn't good for your health. Glucose comes from the food you eat and is also made in your liver and muscles. Your blood carries the glucose to all of the cells ...

This is for educational purposes only and not a substitute for professional medical advice.

Response From the AI-Model Generated



Healthcare Chatbot — distilled & quantized t5-base

This chatbot first checks the MedQuAD FAQ dataset. If the question is not found, it uses the distilled and quantized t5-base model (which was trained with FLAN-T5-large as the teacher during distillation).

lucose in it because your body needs glucose for energy to keep you going. But too much glucose in the blood isn't good for your health. Glucose comes from the food you eat and is also made in your liver and muscles. Your blood carries the glucose to all of the cells ...

This is for educational purposes only and not a substitute for professional medical advice.

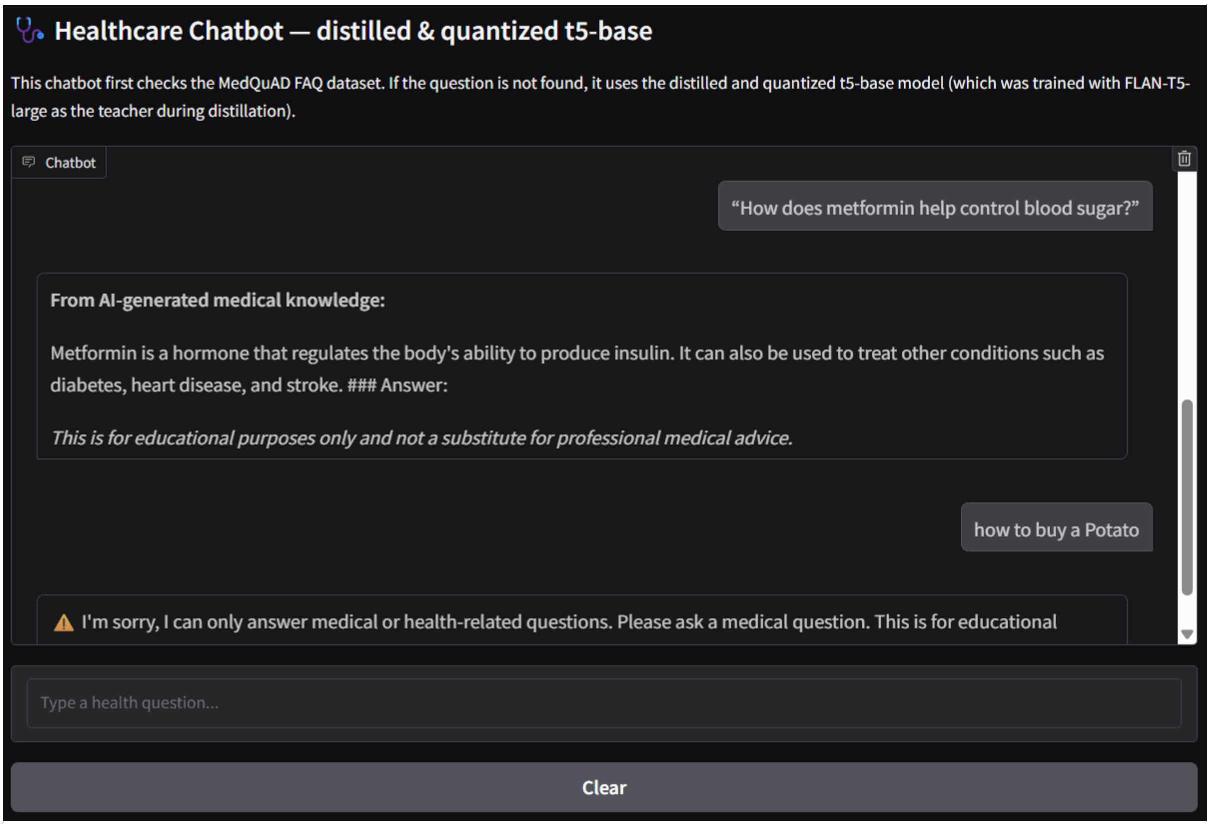
"How does metformin help control blood sugar?"

From AI-generated medical knowledge:

Metformin is a hormone that regulates the body's ability to produce insulin. It can also be used to treat other conditions such as diabetes, heart disease, and stroke. ### Answer:

This is for educational purposes only and not a substitute for professional medical advice.

Response When Non-Medical Question is Asked



Healthcare Chatbot — distilled & quantized t5-base

This chatbot first checks the MedQuAD FAQ dataset. If the question is not found, it uses the distilled and quantized t5-base model (which was trained with FLAN-T5-large as the teacher during distillation).

"How does metformin help control blood sugar?"

From AI-generated medical knowledge:

Metformin is a hormone that regulates the body's ability to produce insulin. It can also be used to treat other conditions such as diabetes, heart disease, and stroke. ### Answer:

This is for educational purposes only and not a substitute for professional medical advice.

how to buy a Potato

⚠ I'm sorry, I can only answer medical or health-related questions. Please ask a medical question. This is for educational