Chatbot M3

Sohaila Ibrahim 52-21225
Nada Elbehery 52-8973

May 18, 2025

Project Overview

- **Objective:** This project aims to evaluate and compare the performance of two large language models — Qwen2-0.5B and Flan T5 small — fine-tuned on the MedQuAD medical question-answering dataset.
- **Comparison Criteria:** BLEU, ROUGE, training and validation loss metrics on a held-out validation set, as well as qualitative assessment through sample responses and chatbot integration.

Dataset Insights: MedQuAD

MedQuAD (Medical Question Answering Dataset) is a domain-specific dataset constructed from trusted U.S. National Institutes of Health (NIH) websites. It is particularly tailored for biomedical and health-related natural language understanding and question answering.

qtype string · classes 16 values	Question string · lengths 124–142 0.1%	Answer string · lengths 6 29k
susceptibility	Who is at risk for Lymphocytic Choriomeningitis (LCM)? ?	LCMV infections can occur after exposure to fresh urine, droppings, saliva, or nesting materials from…
symptoms	What are the symptoms of Lymphocytic Choriomeningitis (LCM) ?	LCMV is most commonly recognized as causing neurological disease, as its name implies, though…
susceptibility	Who is at risk for Lymphocytic Choriomeningitis (LCM)? ?	Individuals of all ages who come into contact with urine, feces, saliva, or blood of wild mice are…
exams and tests	How to diagnose Lymphocytic Choriomeningitis (LCM) ?	During the first phase of the disease, the most common laboratory abnormalities are a low white blood cell…
treatment	What are the treatments for Lymphocytic Choriomeningitis (LCM) ?	Aseptic meningitis, encephalitis, or meningoencephalitis requires hospitalization and…
prevention	How to prevent Lymphocytic Choriomeningitis (LCM) ?	LCMV infection can be prevented by avoiding contact with wild mice and taking precautions when handling…

Figure 1: MedQuAD dataset

Composition

- **Total Samples:** ∼16,400 QA pairs
- **Question Types (qtype):**

- `cause`, `treatment`, `definition`, `symptoms`, `prevention`, etc.

- **Fields per Sample:**
  - `Question`: Natural language query
  - `Answer`: Free-form paragraph or sentence
  - `qtype`: Category label for the question

Characteristics

- **Source Credibility:** Extracted from over 12 reliable government sources (e.g., MedlinePlus, NIH, Genetics Home Reference).
- **Answer Style:** Answers are often encyclopedic, multi-sentence, and use professional clinical terminology.
- **Context-Free:** Unlike SQuAD, MedQuAD does not include context documents — only question and answer pairs.
- **QA Granularity:** High variety in answer lengths and complexity, ranging from short phrases to multi-paragraph explanations.

Challenges

- **Answer Redundancy:** Some answers repeat standard phrasing (e.g., "this condition is inherited in an autosomal recessive pattern").
- **Lack of Conversational Style:** Answers are not phrased for dialog, requiring chat-models to adapt stylistically.
- **Unbalanced qtype Distribution:** Certain categories (e.g., "definition" and "cause") dominate the dataset.

Why It's Useful for Fine-Tuning

- **Domain Adaptation:** Helps general-purpose LLMs specialize in medical QA tasks.
- **Terminology Exposure:** Introduces the model to disease names, gene/protein terms, drug classes, and clinical practices.
- **Robust Evaluation Set:** Offers a good benchmark to test factual correctness and hallucination rates in healthcare settings.

1 Experiemnt 1

Model: Qwen2-0.5B

- **Type:** Decoder-only Transformer (GPT-like)
- **Parameters:** ~500 million

- **Vocabulary Size:** 150k (byte-level BPE)
- **Max Sequence Length:** 2048 tokens
- **Architecture:**
  - Hidden size: 1024
  - Layers: 24
  - Attention heads: 16
  - Positional Encoding: Rotary (RoPE)
- **Strengths:** Optimized for chat-style interactions, efficient on low-resource GPUs

LoRA Fine-Tuning

What is LoRA?

LoRA (Low-Rank Adaptation) introduces small trainable matrices into specific layers (usually attention projections) and freezes the rest of the model, drastically reducing the number of parameters to train.

Configuration

- **Target Modules:** `[q_proj, k_proj, v_proj, o_proj]`
- **Rank (r):** 8
- **Alpha:** 32
- **Dropout:** 0.05
- **Task Type:** Causal Language Modeling
- **Quantization:** 4-bit (NF4) using bitsandbytes

Why LoRA?

- **Efficiency:** Requires fewer GPU resources
- **Speed:** Faster to train even on modest hardware (e.g., T4)
- **Modularity:** Fine-tuned adapters can be merged or reused

Sample Outputs

- **Prompt:** `What is cancer?`
- **Output:** "Cancer is a disease in which abnormal cells control and divide uncontrollably. Cancer can occur in any part of the body, but it most often occurs in the lining of the mouth, nose, throat, esophagus, stomach, colon, and rectum. Cancer can also occur in the lining of the mouth, nose, throat, esophagus, stomach, colon, and rectum. Cancer can be either benign or malignant. Benign cancer is cancer that

does not grow or spread. Malignant cancer is cancer that grows or spreads. Benign cancer is the most common type of cancer. Malignant cancer is rare. Cancer can be treated with surgery, radiation therapy, chemotherapy, and other treatments. Treatment for cancer depends on"

- **Prompt:** `How to prevent it?`
- **Output:** "Cancer prevention is a major issue. The best way to prevent cancer is to get regular checkups and screenings. This includes regular checkups with a doctor or nurse who specializes in cancer prevention. Regular checkups can help detect cancer early and prevent it from spreading. People who have a family history of cancer should also get regular checkups with a doctor who specializes in cancer prevention. Regular checkups with a doctor who specializes in cancer prevention can help detect cancer early and prevent it from spreading. People who have a family history of cancer should also get regular checkups with a doctor who specializes in cancer prevention. Regular checkups with a doctor who specializes in cancer prevention can help detect cancer early and prevent it from spreading. People who have a family history of cancer"

Chatbot Integration

- **Library:** LangChain
  We utilized LangChain for chaining prompt templates, memory management, and user interaction. It offers modular tools to structure prompt-driven conversations.
- **Prompt Template:**

  ```
  You are a helpful and concise medical assistant.
  Previous conversation: {history}
  Now answer the user's question.
  User: {input}
  Assistant:
  ```

  This template reinforces the model's role and ensures that answers remain aligned with prior context while staying concise and informative.
- **Memory:** ConversationBufferMemory
  We employed `ConversationBufferMemory` with a context window of the last two exchanges (`k=2`), allowing the model to retain relevant past turns without overloading the input. This supports context-aware responses and continuity across turns. Memory can be reset at runtime to start a new conversation thread.
- **Sampling Settings:**
  - `top_p = 0.9` — nucleus sampling to balance creativity and relevance.
  - `temperature = 0.7` — moderate randomness to avoid repetition and promote variation.
  - `repetition penalty = 1.2` — discourages repeated phrases in long answers.
  - `max_new_tokens = 150` — caps the length of generated answers to control verbosity.

- **Model Backend:** HuggingFacePipeline
  The fine-tuned model with LoRA adapters was wrapped using HuggingFace's text generation `pipeline`, then integrated into LangChain via `HuggingFacePipeline`. This enabled smooth use of the model for live inference while preserving generation settings and tokenizer alignment.

2 Experiment2

2.1 Model :FLAN-T5-small

This experiment fine-tunes FLAN-T5-small that has 80 million parameters on the MedQuad medical Q&A dataset to generate accurate answers to medical questions. The model leverages both question text and question types (e.g., diagnosis, treatment) through structured prompting. We employ selective layer freezing and evaluate using ROUGE and BLEU metrics, demonstrating effective adaptation of a general-purpose language model to the medical domain.

We present an approach for medical question answering by fine-tuning FLAN-T5, an instruction-tuned variant of Google's T5 model. The system is designed to:

- Generate free-form medical answers
- Incorporate question-type context
- Balance computational efficiency with performance

2.2 Methodology

- Medical Q&A pairs with question type labels
- Preprocessing:
  - Consolidated rare question types (< 2 samples) into "other"
  - Train/validation split (90%/10%)

2.3 Model Architecture

Table 1: FLAN-T5-small Specifications

Property	**Value**
Architecture	Encoder-decoder
Parameters	60M
Input Format	`"question type: {qtype} | question: {question}"`
Output	Free-text answer

2.4 Training Strategy

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, x; \theta) \tag{1}$$

Where:

- x: Input sequence with question type
- y: Target answer sequence
- θ: Model parameters

- **Layer Freezing**:
  - Decoder layers: All unfrozen
  - Encoder layers: First 3 + last 3 unfrozen
  - Embedding layer: Unfrozen
- **Optimization**:
  - AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
  - Learning rate: 3×10^{-5} with linear warmup
  - Gradient clipping (1.0)
  - Batch size: 4 (GPU) / 2 (CPU)

2.5 Chatbot Integration

2.5.1 Initialization

- Uses the fine tuned saved model of `google/flan-t5-small`
- Maps medical question types (e.g., symptoms, treatment) to keywords for classification.
- Runs on `GPU (CUDA)` if available, otherwise falls back to `CPU`.

2.5.2 Question Type Prediction

- Analyzes the input question to determine its type (e.g., "What are the symptoms of COVID-19?" → classified as `"symptoms"`).
- Uses a predefined mapping of question types to keywords.
- If no match is found, defaults to `"information"`.

2.5.3 Answer Generation

- Formats the input as: `"question type: [type] | question: [question]"`.
- Uses `T5 tokenizer` to encode the input.
- Generates an answer with the following settings:
  - `max_length=150` (limits answer length).
  - `num_beams=5` (beam search for better quality).
  - `temperature=0.7` (controls randomness).
  - `top_k=50` and `top_p=0.95` (nucleus sampling).
- Returns the decoded answer along with the predicted question type.

2.5.4 Conversation Memory

- Implements memory using `ConversationBufferMemory` from the `LangChain` library.
- Maintains a history of user and chatbot messages during the session.
- Stores previous inputs and generated responses to enable context-aware interaction in future enhancements.
- Each new user question and corresponding AI response are added to the memory buffer dynamically during runtime.

2.5.5 Error Handling

- Catches and reports errors during model initialization and answer generation.
- Returns a fallback message if an error occurs.

2.5.6 sample output

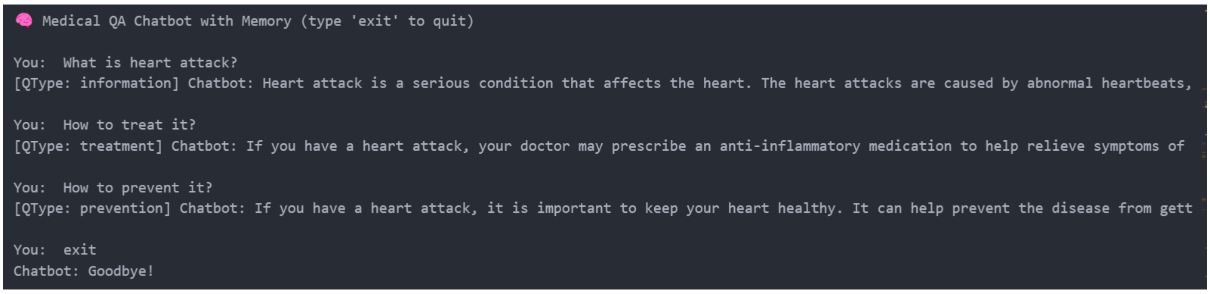

Figure 2: output

Results

To assess the performance of our fine-tuned medical language models, we computed two standard text generation metrics: **BLEU**, and **ROUGE**. BLEU measures n-gram precision, assessing how many n-grams in the generated output appear in the reference answers. It penalizes brevity using the *brevity penalty*. ROUGE focuses on recall — how much of the reference text is captured by the generated output.

Experiemnt1

BLEU

- **BLEU:** 0.0732
- **Interpretation:** This score indicates that the generated answers have limited exact n-gram overlap with the references. This is expected in open-ended question answering tasks, especially in the medical domain, where multiple valid phrasings may exist.
- **Limitation:** BLEU penalizes variation in wording and shorter outputs, making it less reliable for free-form answers.

Rogue

- **ROUGE-1:** 0.3380
  Measures the overlap of unigrams (individual words). Indicates that about 33.8% of the words in the reference answer appear in the generated response.
- **ROUGE-2:** 0.2052
  Measures the overlap of bigrams (two-word sequences). This is a stricter metric and a score above 0.2 shows the model is capturing some local context from the reference.
- **ROUGE-L:** 0.2736
  Measures the longest common subsequence (LCS) between the predicted and reference text. It reflects fluency and structure preservation, not just word matching.
- **Interpretation:** The ROUGE scores are promising, showing that key facts from references are retained. However, due to the free-form nature of answers, exact sequence overlap is often limited.

Summary: While BLEU suggests limited exact phrasing match, the ROUGE scores show the model captures meaningful and relevant parts of the reference answers. This is acceptable given the variability in natural language answers to medical questions.

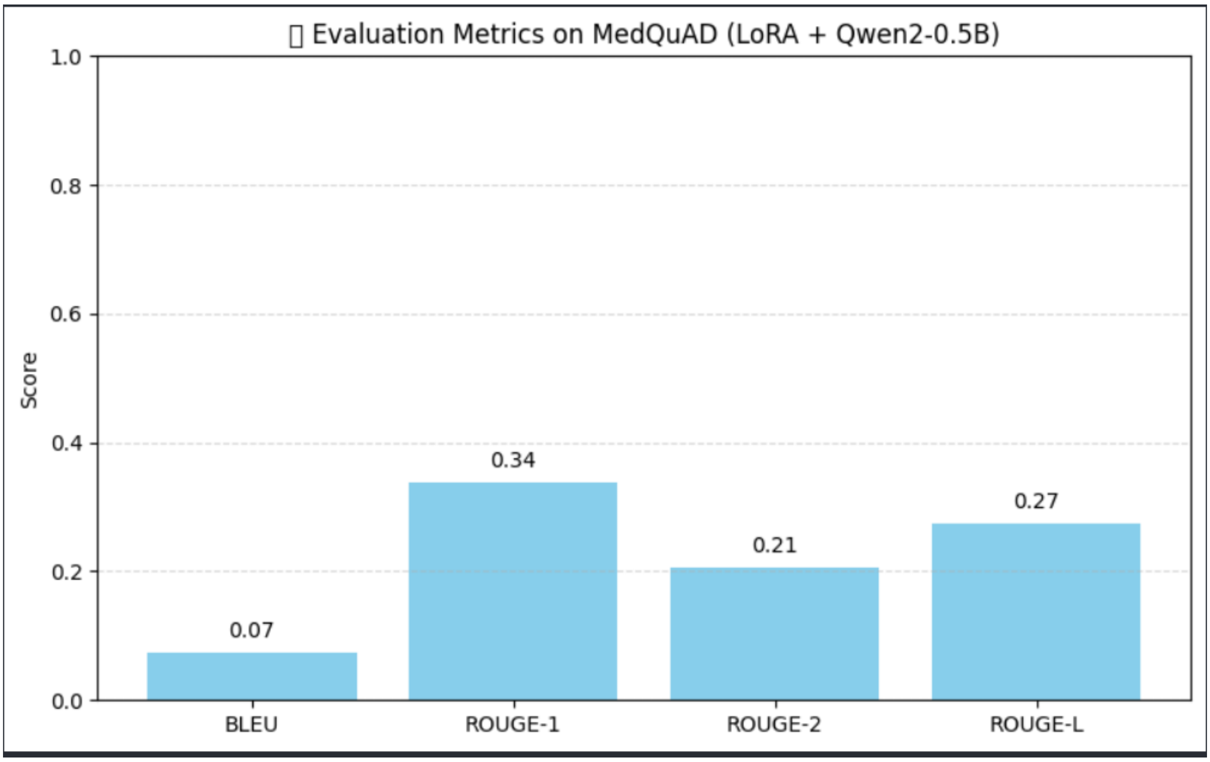

Figure 3: Exp1- Evaluation Metrics Graph

Experiemnt2

Table 2: Model Performance Metrics

Metric	Score
ROUGE-1	0.3137
ROUGE-2	0.2057
ROUGE-L	0.2802
BLEU	0.1320

- ROUGE-1 (unigram overlap) shows the highest score at 0.3137
- ROUGE-L (longest common subsequence) demonstrates reasonable answer structure preservation at 0.2802
- The BLEU score of 0.1320 indicates room for improvement in n-gram matching precision

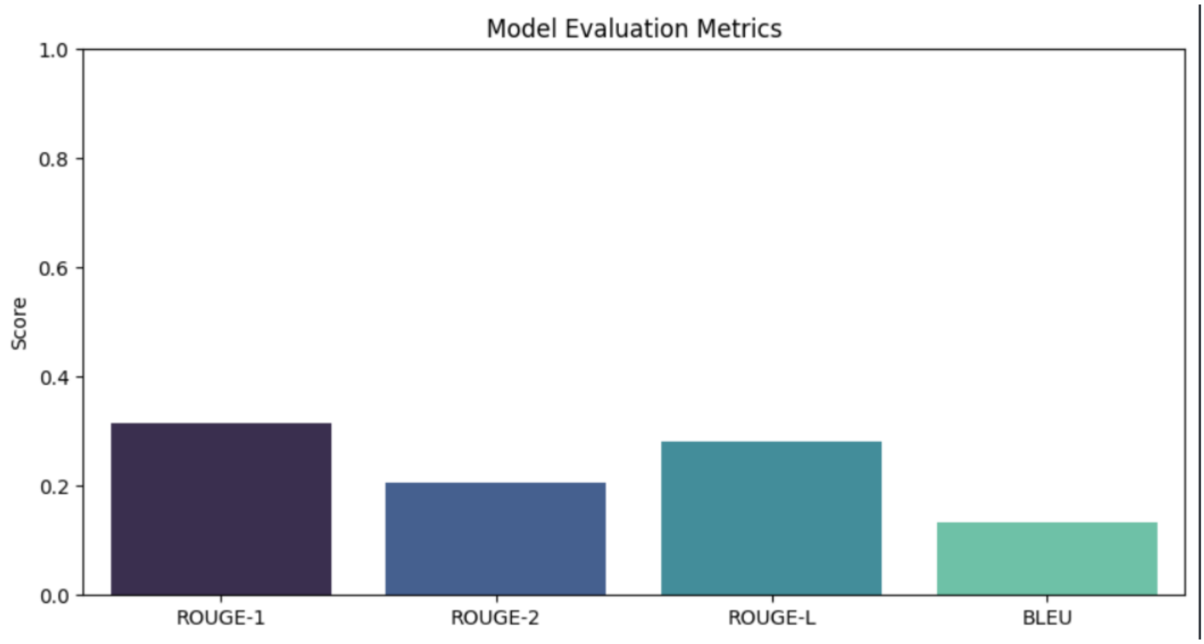

Figure 4: flan t5 model

2.6 Performance Interpretation

- **ROUGE-1 (0.3137)**: Indicates reasonable unigram overlap between generated and reference answers, suggesting the model captures basic medical terminology well.
- **ROUGE-2 (0.2057)**: Lower bigram scores reveal challenges in maintaining precise medical phraseology (e.g., drug-dose relationships).
- **ROUGE-L (0.2802)**: Demonstrates moderate coherence in answer structure, though with room for improvement in clinical reasoning flow.
- **BLEU (0.1320)**: The relatively low score highlights the difficulty of exact n-gram matching in medical domains where answer variability is high.

2.7 Model Architecture

- **FLAN-T5-small** (60M parameters) was selected for:
  - Unified text-to-text framework
  - Instruction-tuning advantage for prompt following
  - Computational efficiency

The results demonstrate that FLAN-T5-small can achieve:

- Moderate performance in medical QA (ROUGE-L > 0.28)
- Effective prompt conditioning through question types
- Computational efficiency via strategic layer freezing

Recommended Improvements:

1. Model scaling to FLAN-T5-base (250M params) with LoRA
2. Integration of retrieval mechanisms (RAG) for evidence grounding
3. Multi-task learning with related biomedical NLP tasks

Comparison

- **BLEU:** Flan-T5 outperforms Qwen2 (0.1320 vs. 0.0732), likely due to its more explicit sequence-to-sequence objective.
- **ROUGE-1 and ROUGE-L:** Both models show comparable structure and recall capabilities, with Flan-T5 slightly ahead in ROUGE-L.
- **ROUGE-2:** Surprisingly, both models tie on bigram overlap (0.205), showing equal capacity to learn short phrases.
- **Conclusion:** While Qwen2 benefits from recent architecture and LoRA tuning, Flan-T5 demonstrates stronger baseline performance in n-gram precision. Further optimization or prompt engineering could close this gap.