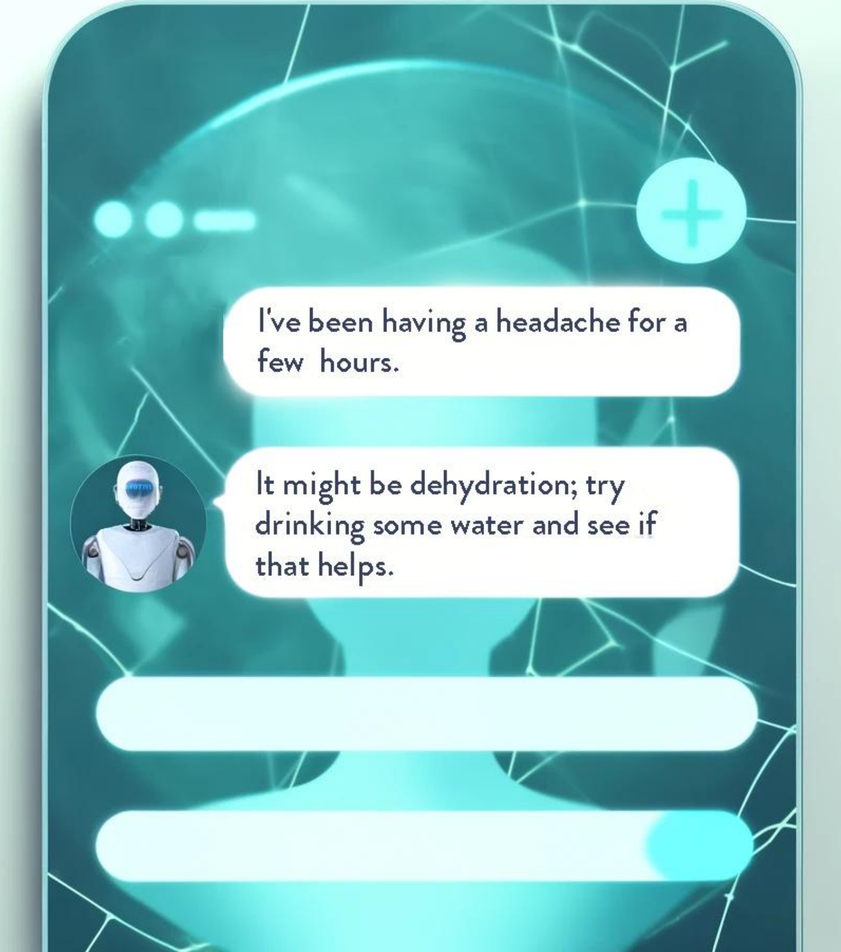

I've been having a headache for a few hours.
It might be dehydration; try drinking some water and see if that helps.

MediQA Chatbot

🤖🧠🩺

Md. Shafi Ud Doula -st124047
Sonu Adhikari-st124409
Ashmita Phuyal-st124454
Tanzil Al Sabah-st123845
Sai Haneesha Bestha-st124089

Course: Artificial Intelligence: Natural Language Understanding
Course Instructor: Dr. Chaklam Silpasuwanchai

Content

1 Introduction & Objective

2 Problem Statement Solution Requirements

3 System Architecture & Methodology

4 Data Preprocessing, Results & Evaluation

5 Demo Application

6 Limitation, Future Works & Conclusion

MediQA Chatbot 🤖🧠🩺

– Introduction & Background –

An AI-driven virtual assistant that utilizes advanced NLP techniques to provide immediate and personalized medical information, helping users make informed healthcare decisions.

– Objective & Goals–

Improve Health Literacy: Enhance users' understanding of health issues, enabling informed decisions about their medical care.

Provide Timely Access to Data: Offer 24/7 access to clinical data, facilitating timely health decisions and actions.

Demonstrate NLP Application: Showcase the practical use of NLP in improving patient satisfaction and healthcare usability in our NLP course project.

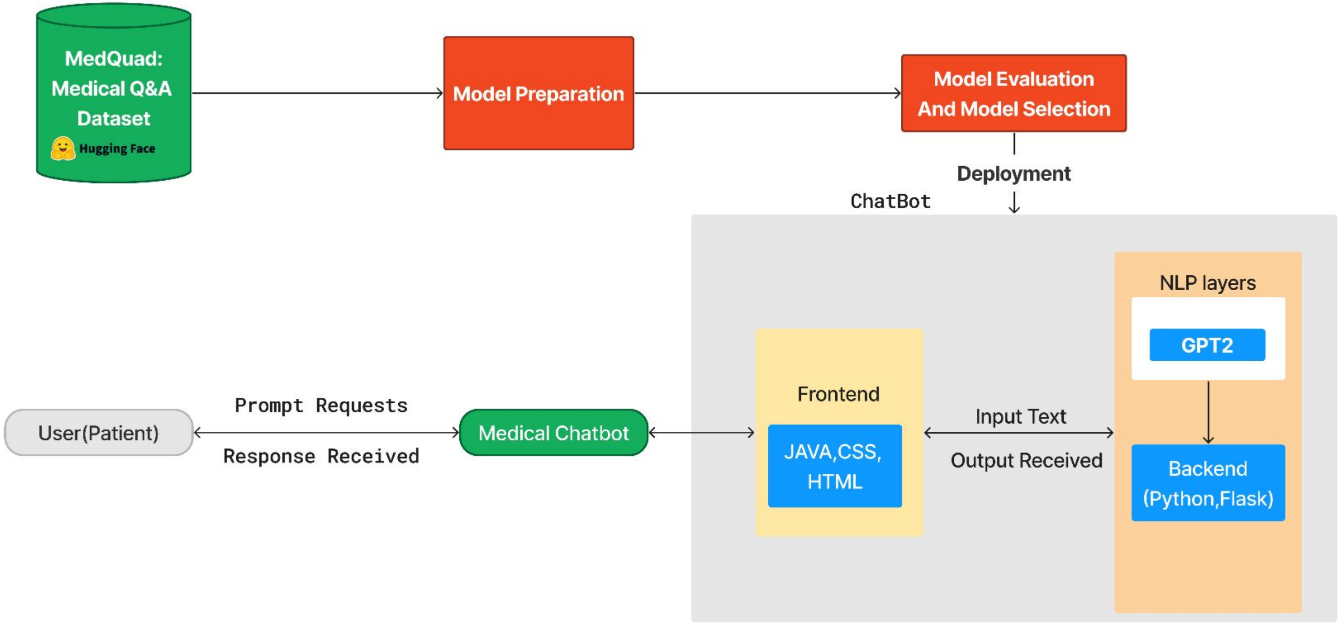
Problem Statement & Solution Requirements

– Problem –

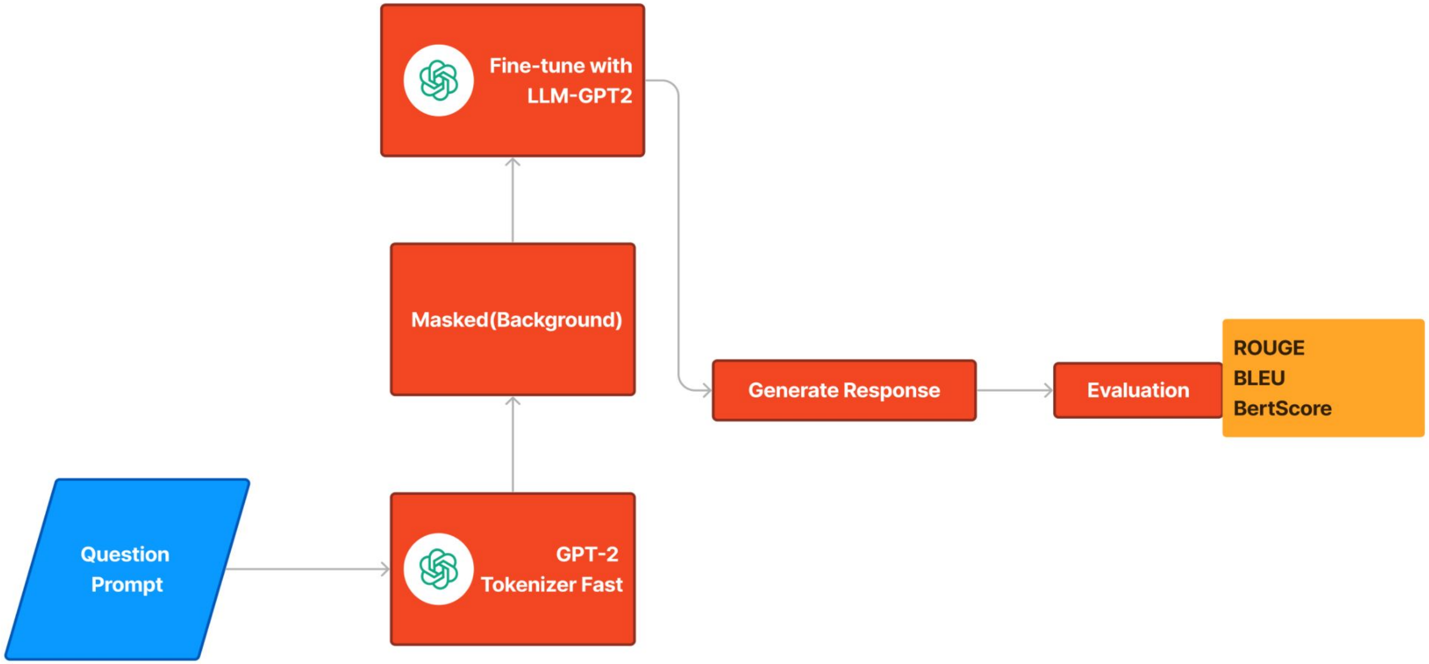
- Limited healthcare expert availability due to systemic constraints and logistical issues.
- Delays in medical attention owing to inadequate knowledge or understanding of symptom severity.
- Conventional chatbots have limitations in natural language understanding and adaptability, affecting accuracy.

– Solution Requirements –

- Utilize advanced NLP for accurate comprehension of user queries and responses.
- Integrate with a constantly updated medical knowledge base.
- Conduct trials to identify the most effective model.
- Ensure smooth integration with a web application
- Offer a user-friendly interface to increase user satisfaction.



System Diagram - Methodology



GPT-2 System Architecture

Methodology

Data Preparation:

- Employ the MedQuad dataset, enriched with data from Huggingface and GitHub.
- Conduct tokenization, numericalization, and cleansing for data integrity.

Model Development:

- Experiment with Seq2Seq, GPT-2 models.
- Implement advanced NLP techniques, including perplexity and seq2seq validation, for enhanced performance.

Implementation:

- Select the best-performing model for further development.
- Fine-tune the model on medical QA data using GPT-2.

Evaluation:

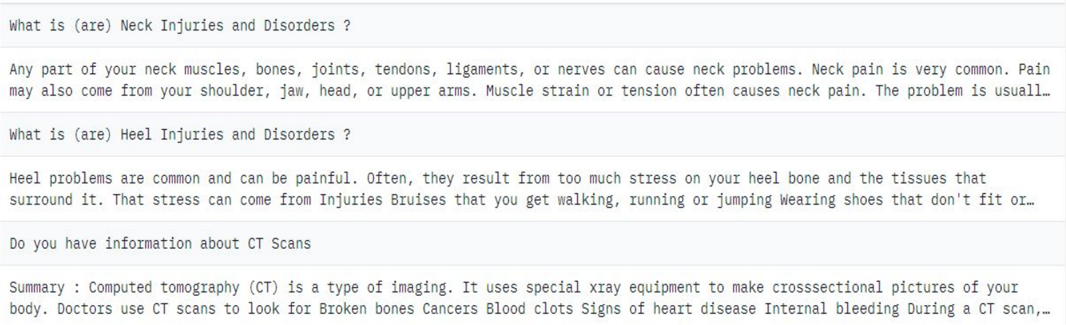
- Apply a human evaluation metrics and automatic evaluation metrics, including **BLEU, ROGUE.**
- Expand evaluation to include **precision, recall, and F1-score** to gauge the relevance and accuracy of medical advice.

Deployment:

- Integrate with a web application for user-friendly access.
- Refine the model based on user feedback and chatbot performance data.

Dataset

Dataset Content:



What is (are) Neck Injuries and Disorders ?

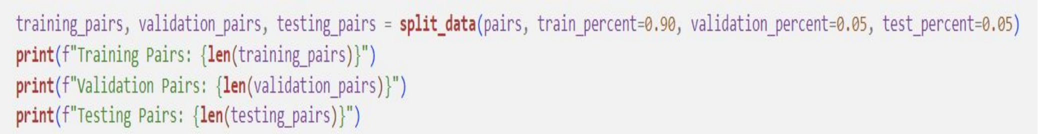
Any part of your neck muscles, bones, joints, tendons, ligaments, or nerves can cause neck problems. Neck pain is very common. Pain may also come from your shoulder, jaw, head, or upper arms. Muscle strain or tension often causes neck pain. The problem is usuall…

What is (are) Heel Injuries and Disorders ?

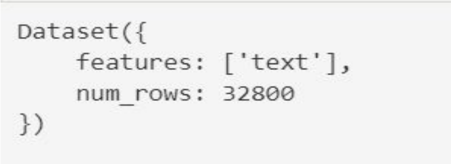
Heel problems are common and can be painful. Often, they result from too much stress on your heel bone and the tissues that surround it. That stress can come from Injuries Bruises that you get walking, running or jumping Wearing shoes that don't fit or…

Do you have information about CT Scans

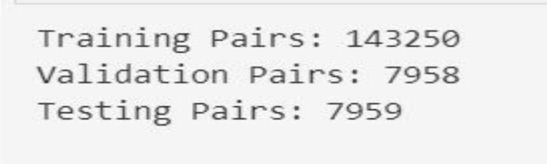
Summary : Computed tomography (CT) is a type of imaging. It uses special xray equipment to make crosssectional pictures of your body. Doctors use CT scans to look for Broken bones Cancers Blood clots Signs of heart disease Internal bleeding During a CT scan,…



```
training_pairs, validation_pairs, testing_pairs = split_data(pairs, train_percent=0.90, validation_percent=0.05, test_percent=0.05)
print(f"Training Pairs: {len(training_pairs)}")
print(f"Validation Pairs: {len(validation_pairs)}")
print(f"Testing Pairs: {len(testing_pairs)}")
```



```
Dataset({
    features: ['text'],
    num_rows: 32800
})
```



```
Training Pairs: 143250
Validation Pairs: 7958
Testing Pairs: 7959
```

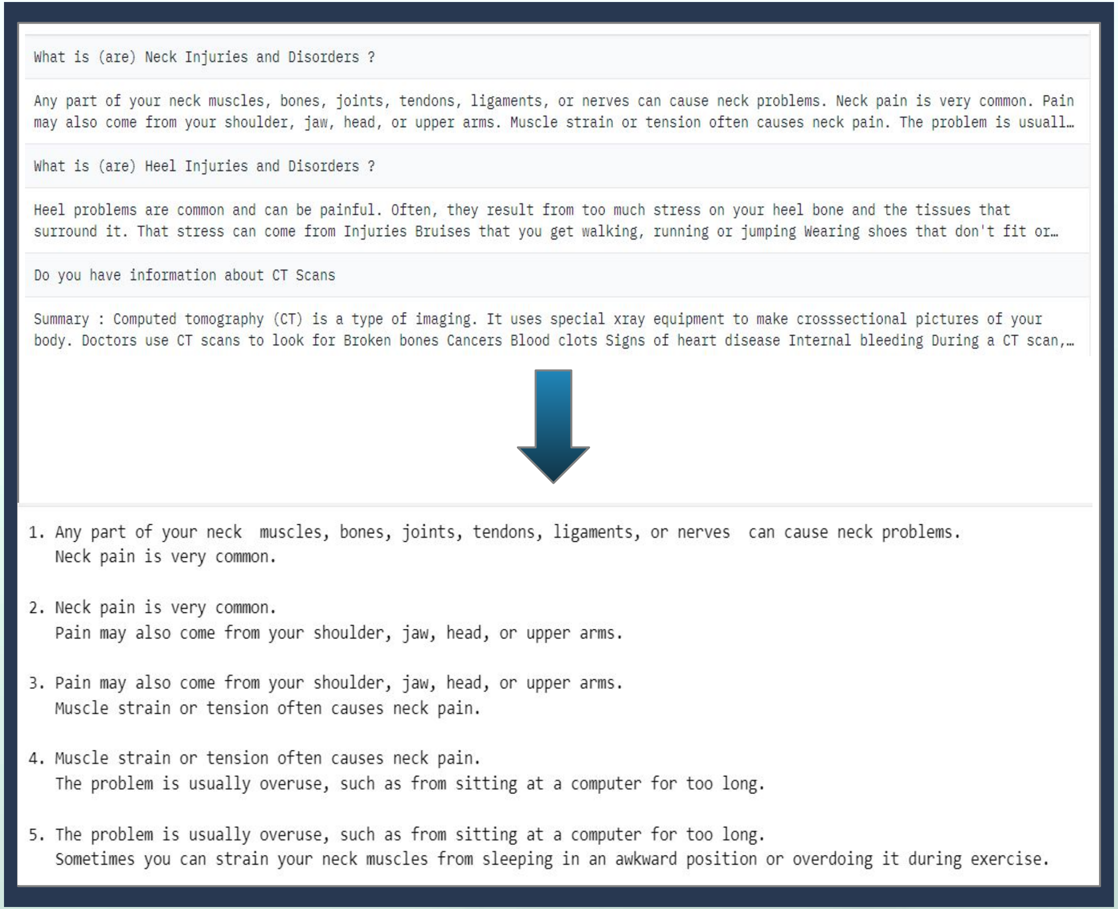
Dataset source: Dataset from Hugging Face

Total no. of Rows in Dataset: 32800

Dataset is further split into training,validation and Testing pairs.

Training_pairs count: 143250
Validation_pairs count: 7958
Testing_pairs: 7959

Dataset Processing



What is (are) Neck Injuries and Disorders ?

Any part of your neck muscles, bones, joints, tendons, ligaments, or nerves can cause neck problems. Neck pain is very common. Pain may also come from your shoulder, jaw, head, or upper arms. Muscle strain or tension often causes neck pain. The problem is usuall…

What is (are) Heel Injuries and Disorders ?

Heel problems are common and can be painful. Often, they result from too much stress on your heel bone and the tissues that surround it. That stress can come from Injuries Bruises that you get walking, running or jumping Wearing shoes that don't fit or…

Do you have information about CT Scans

Summary : Computed tomography (CT) is a type of imaging. It uses special xray equipment to make crosssectional pictures of your body. Doctors use CT scans to look for Broken bones Cancers Blood clots Signs of heart disease Internal bleeding During a CT scan,…

```
1. Any part of your neck  muscles, bones, joints, tendons, ligaments, or nerves  can cause neck problems.
   Neck pain is very common.

2. Neck pain is very common.
   Pain may also come from your shoulder, jaw, head, or upper arms.

3. Pain may also come from your shoulder, jaw, head, or upper arms.
   Muscle strain or tension often causes neck pain.

4. Muscle strain or tension often causes neck pain.
   The problem is usually overuse, such as from sitting at a computer for too long.

5. The problem is usually overuse, such as from sitting at a computer for too long.
   Sometimes you can strain your neck muscles from sleeping in an awkward position or overdoing it during exercise.
```

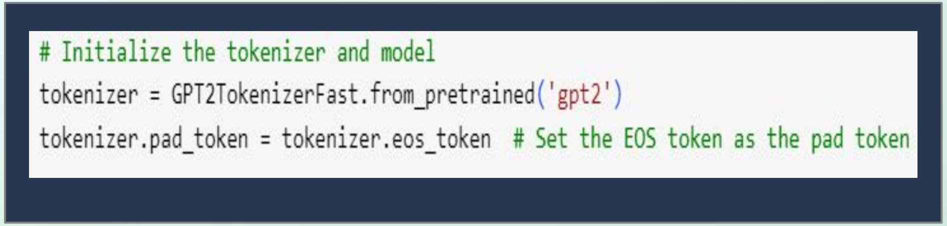
A conversational dataset is constructed by sequencing sentence pairs: each input sentence is paired with a subsequent output to train a model like Seq2Seq. This method enables the model to learn dialogue continuity and context:

Pair1: Input "What are Neck Injuries?" leads to the output "Neck muscles, bones, joints can cause problems."
Pair2: That output becomes the input for the next pair, leading to "Neck pain is common."
Pair3: Continues with "Neck pain is common," leading to "Pain can also come from shoulder, jaw, head."

This creates a cascading dialogue flow for the model to practice.

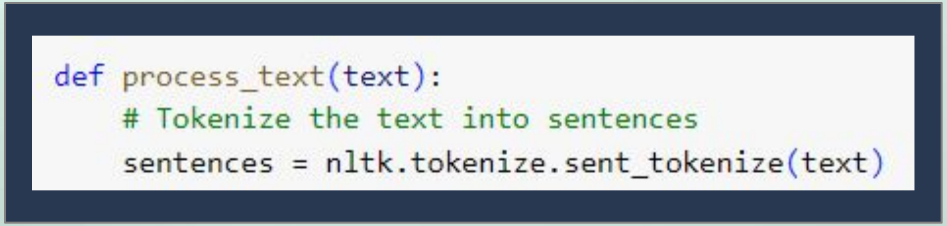
Tokenization

- We're using the GPT2TokenizerFast from the transformers library to tokenize text efficiently for processing with the GPT-2 model.



```
# Initialize the tokenizer and model
tokenizer = GPT2TokenizerFast.from_pretrained('gpt2')
tokenizer.pad_token = tokenizer.eos_token  # Set the EOS token as the pad token
```

- We’re using NLTK library for carrying out the tokenization for the seq2seq model.



```
def process_text(text):
    # Tokenize the text into sentences
    sentences = nltk.tokenize.sent_tokenize(text)
```

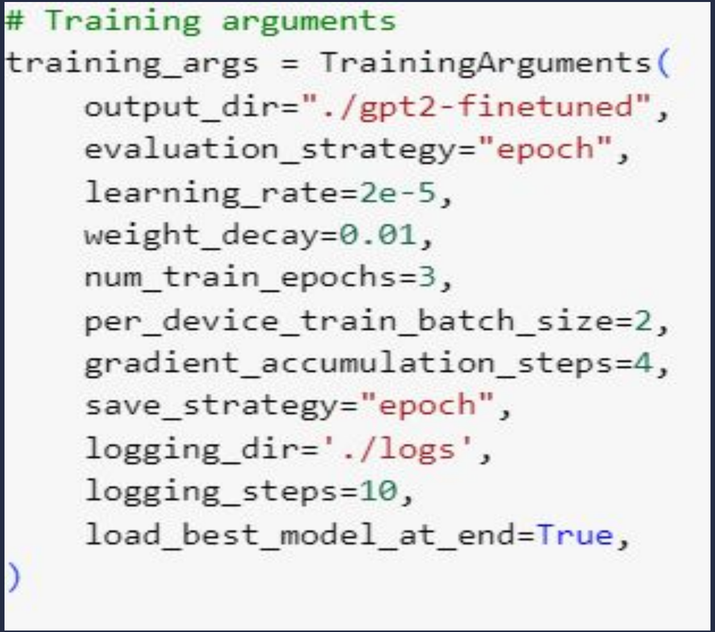
Modeling and Fine tuning

Hyperparameters for Fine Tuning GPT-2:

- Evaluation Strategy: epoch
- Learning Rate: 2e-5
- Weight Decay: 0.01
- Number of Training Epochs: 3

Hyperparameters for Seq2Seq:

- Hidden Layers = 512
- Number of Iterations = 15000
- Teacher_forcing_ratio = 0.5
- Learning Rate Encoder = 0.0001
- Learning Rate Decoder = 0.0005
- Optimizer = Adam Optimizer



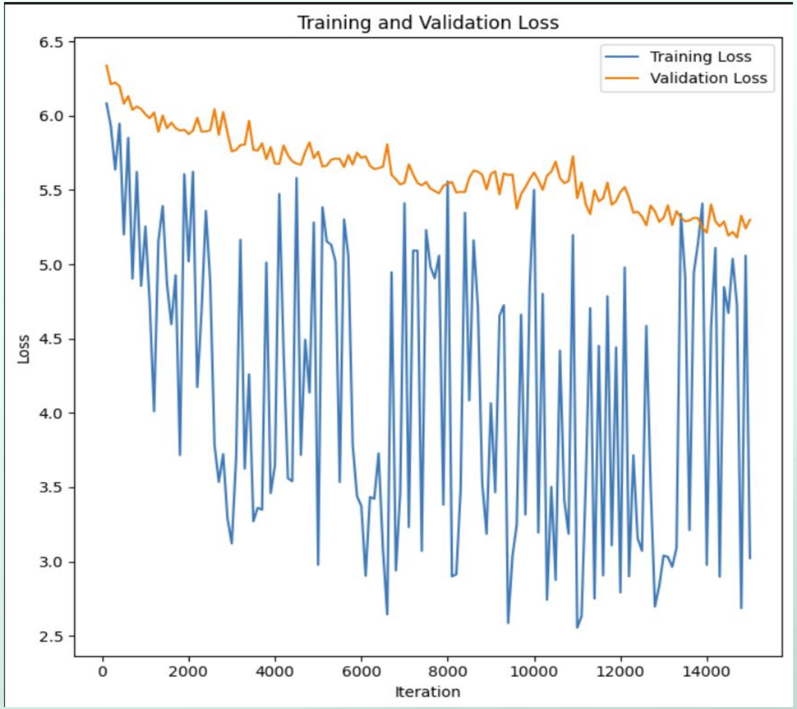
```
# Training arguments
training_args = TrainingArguments(
    output_dir="./gpt2-finetuned",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    weight_decay=0.01,
    num_train_epochs=3,
    per_device_train_batch_size=2,
    gradient_accumulation_steps=4,
    save_strategy="epoch",
    logging_dir='./logs',
    logging_steps=10,
    load_best_model_at_end=True,
)
```

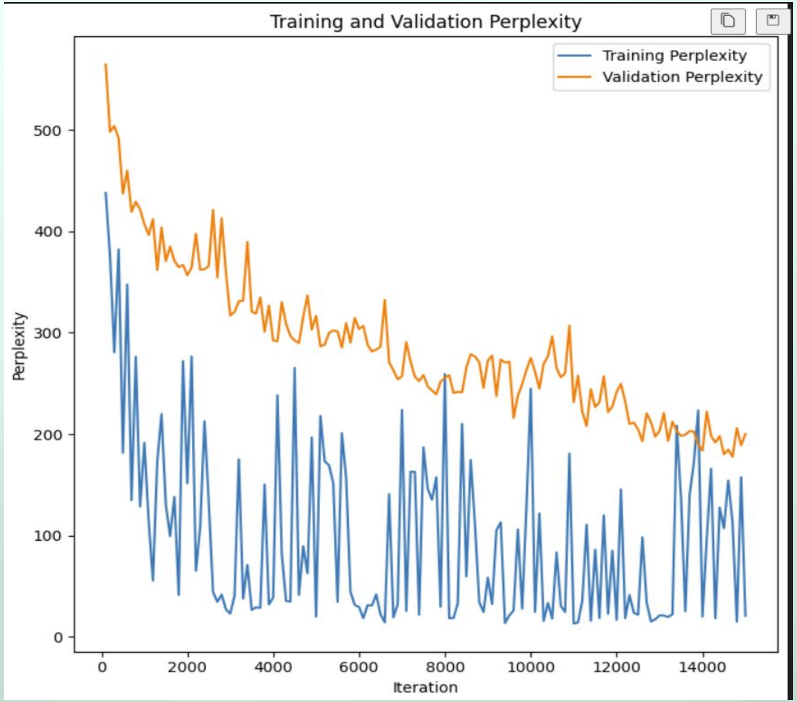
Results & Evaluation

Metric	Seq2Seq	GPT-2
BLEU Score	0.1857	0.3056
ROUGE Score	0.4170	0.3934
Precision	0.2648	0.3647
Recall	0.1485	0.2485
F1 Score	0.1723	0.2723

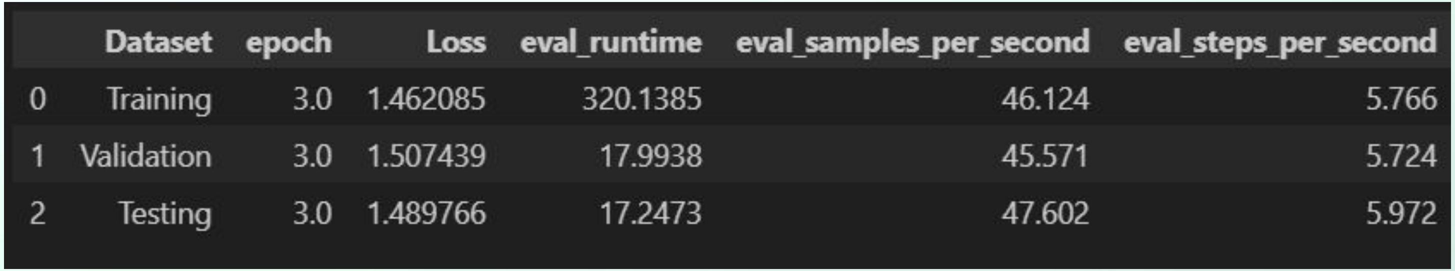
GPT-2 performs better than the Seq2Seq Model

Evaluation Seq2Seq





Evaluation GPT2



	Dataset	epoch	Loss	eval_runtime	eval_samples_per_second	eval_steps_per_second
0	Training	3.0	1.462085	320.1385	46.124	5.766
1	Validation	3.0	1.507439	17.9938	45.571	5.724
2	Testing	3.0	1.489766	17.2473	47.602	5.972

- The average training loss over 3 epochs is 1.462085, validation loss is 1.507439 and testing loss is 1.489766.
- Training loss appears to be lower than validation and testing losses
- Validation and testing losses are relatively close

- Outperformed other models in many metrics.
- Showed a gap between automated scores and human perception.

GPT2 Model ?

Which model to pick?

- Outperformed by GPT-2 in most metrics.
- Human evaluations indicated limitations in providing relevant medical advice.

Seq2Seq Model ?

Overall, the GPT-2 model outperformed the Seq2Seq model in most metrics, particularly in generating more relevant and accurate medical advice. The Seq2Seq model showed decent performance but lagged in human evaluation and some automated metrics.

Human Evaluation

- **Rating Scale: Ranges from 1 (Poor) to 5 (Excellent).**

- **Evaluator Comments: Specific feedback from medical professionals, noting both strengths and areas need improvement.**

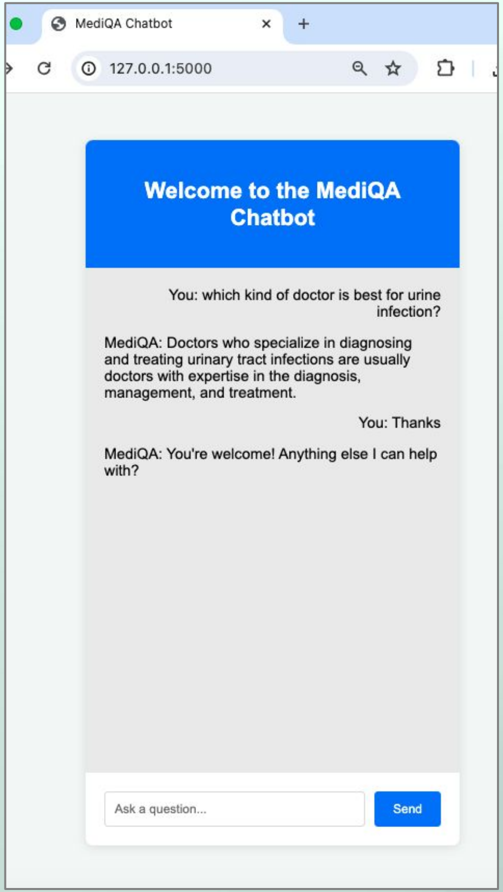
Category	Rating (1-5)	Evaluator Comments (Doctors' Feedback)
Medical Accuracy	2	"Frequently inaccurate or overly generic, lacking specific medical detail."
Guideline Adherence	2	"Often fails to follow clinical guidelines, especially in complex cases."
Clarity	3	"Responses are understandable but sometimes use jargon that could confuse patients."
Empathy	2	"Struggles to convey genuine empathy, often comes off as detached."
Response Relevance	2	"Irrelevant or off-topic responses are common, particularly in nuanced discussions."

Human Evaluation

Category	Rating (1-5)	Definitions and Criteria
Medical Accuracy	1-5	1 (Poor): Consistently incorrect, misleading. 2 (Fair): Often inaccurate, lacks detail. 3 (Average): Generally accurate, some errors.4 (Good): Mostly accurate, minor inaccuracies. 5 (Excellent): Highly accurate, detailed, fully reliable.
Guideline Adherence	1-5	1 (Poor): Ignores clinical guidelines. 2 (Fair): Struggles with guidelines, frequent errors. 3 (Average): Generally follows guidelines. 4 (Good): Consistently adheres to guidelines. 5 (Excellent): Perfect adherence, no exceptions.
Clarity	1-5	1 (Poor): Very confusing, unclear. 2 (Fair): Somewhat understandable, uses complex jargon. 3 (Average): Clear with occasional complex language. 4 (Good): Very clear, minimal jargon. 5 (Excellent): Exceptionally clear, straightforward.
Empathy	1-5	1 (Poor): Completely detached, inappropriate. 2 (Fair): Limited empathy, often seems indifferent. 3 (Average): Shows basic empathy, somewhat supportive. 4 (Good): Very empathetic and supportive. 5 (Excellent): Exceptionally empathetic, always supportive.
Response Relevance	1-5	1 (Poor): Responses mostly irrelevant or off-topic. 2 (Fair): Often irrelevant or slightly off-topic. 3 (Average): Mostly relevant, some off-topic responses. 4 (Good): Highly relevant, consistently on-topic. 5 (Excellent): Always relevant, perfectly on-topic.

Deployment

- **Purpose:** Instantly answers medical queries using AI.
- **Features:** Real-time responses, handles both simple greetings and complex medical questions.
- **Technology:** Powered by Python, Flask, and AI models via Torch and Transformers.
- **Benefits:** Accessible 24/7 medical assistance, reduces routine workload for healthcare professionals.



Welcome to the MediQA Chatbot

You: which kind of doctor is best for urine infection?

MediQA: Doctors who specialize in diagnosing and treating urinary tract infections are usually doctors with expertise in the diagnosis, management, and treatment.

You: Thanks

MediQA: You're welcome! Anything else I can help with?

Limitations/Future Enhancements

Limitations:

- Lack of Domain Knowledge

- Limited computational resources

Future Enhancements:

- Experimenting with diverse datasets
- Engage stakeholders, including healthcare professionals and patients, to gather feedback and ensure alignment with industry standards
- Integration of human Evaluation
- Implementation of sophisticated models for intent query classification

THANK YOU