MediAssist: A Hybrid NLP Framework for Clinical Intent Classification and Retrieval-Augmented Generation

Abdul Rafeh
Team Member
MediAssist Project
New York, USA
arafeh1@mercy.edu

Samith Kamarthi
Team Member
MediAssist Project
New York, USA
skamarthi@mercy.edu

***Abstract*—In the contemporary healthcare landscape, rapid access to accurate medical information is a critical determinant of patient outcomes [4]. However, patients often encounter significant barriers to accessing professional care, including long waiting times, limited availability of specialists, and the inherent complexity of medical terminology [5]. While the digital age has democratized access to information, online medical resources are frequently unstructured, overwhelming, and lack the personalization required for effective decision support [6]. This paper introduces "MediAssist," a comprehensive Natural Language Processing (NLP) framework designed to bridge this gap [7]. Unlike purely generative models that risk hallucination, MediAssist employs a hybrid architecture. It utilizes a fine-tuned BioBERT model to classify user queries into seven clinically meaningful intents, serving as an intelligent routing layer [9]. Subsequently, a retrieval-augmented generation (RAG) pipeline utilizing Sentence-BERT and FAISS retrieves authoritative medical context, which is then synthesized by a controlled Large Language Model (LLM) into patient-friendly responses [1, 9, 122]. Experimental validation on a dataset of 26,500 medical queries demonstrates an overall classification accuracy of 87%, with high precision in critical categories such as Treatment and Definition [119]. This work validates the feasibility of using small, domain-specific models for routing to ensure safety and reliability in healthcare chatbots.**

***Index Terms*—Natural Language Processing, BioBERT, Healthcare Chatbots, Intent Classification, Retrieval-Augmented Generation, FAISS, Deep Learning.**

I. Introduction

The accessibility of reliable healthcare information is a global challenge. Despite the proliferation of medical websites, patients often struggle to find specific, actionable advice relevant to their symptoms. Traditional search engines retrieve documents based on keyword matching, often forcing users to sift through dense professional literature or generic forums. This cognitive load can lead to anxiety and misinformation, a phenomenon sometimes termed "cyberchondria."

Advances in Natural Language Processing (NLP), particularly the advent of Large Language Models (LLMs), offer a potential solution. Chatbots can simulate conversation, providing a natural interface for health inquiries. However, the deployment of end-to-end generative models in healthcare is fraught with risk. The primary concern is "hallucination," where a model generates plausible but factually incorrect medical advice. In a clinical setting, such errors are unacceptable.

To address these limitations, we propose **MediAssist**, a system that prioritizes reliability through a modular design. Instead of relying on a single model to generate answers, MediAssist decomposes the problem into three distinct stages:

1) **Intent Understanding:** A specialized discriminator (BioBERT) identifies what the user is asking (e.g., "What are the symptoms?" vs. "How do I treat this?").
2) **Information Retrieval:** A semantic search engine retrieves verified answers from a curated medical knowledge base.
3) **Response Synthesis:** A generative model rewrites the retrieved information for clarity, strictly constrained by the source text.

This paper details the development of MediAssist, highlighting the creation of a unified dataset from MedQuAD and MedQA, the Zero-Shot labeling strategy employed to overcome data scarcity, and the rigorous evaluation of the intent classifier.

II. Related Work

A. Healthcare Chatbots

Early healthcare chatbots were predominantly rule-based, relying on decision trees and keyword mapping (e.g., ELIZA-style systems). While safe, these systems lacked the flexibility to understand natural language variation. If a user described a symptom using slang or non-standard phrasing, the system would often fail.

B. Transfer Learning in Biomedicine

The introduction of BERT (Bidirectional Encoder Representations from Transformers) revolutionized NLP by enabling transfer learning. However, general-domain BERT performs sub-optimally on biomedical text due to distribution shifts in vocabulary (e.g., complex drug names, anatomical terms).

BioBERT [2] addressed this by pre-training BERT on large-scale biomedical corpora (PubMed, PMC). This domain adaptation allows BioBERT to capture the semantic nuance of clinical terms better than generic models, making it the ideal backbone for our intent classifier.

C. Retrieval-Augmented Generation (RAG)

RAG represents a paradigm shift from "parametric knowledge" (facts stored in model weights) to "non-parametric knowledge" (facts stored in an external database). By retrieving relevant documents and feeding them into the context window of a generator, RAG systems reduce hallucinations and allow for easier updates to medical knowledge without re-training the model. MediAssist implements this architecture using dense vector retrieval.

III. Data Acquisition and Preprocessing

A robust machine learning model requires high-quality training data. We aggregated two primary datasets to create a comprehensive corpus.

A. MedQuAD Dataset

The Medical Question Answering Dataset (MedQuAD) [1] served as the foundational source. Curated from authoritative sites like the NIH and the CDC, it contains 47,457 question-answer pairs related to 37 types of questions (e.g., Treatment, Diagnosis, Side Effects). Fig. 1 illustrates the basic structure and volume of the MedQuAD data.

```
CSV Structure Analysis: medquad.csv

BASIC INFORMATION
Total rows:      16,412
Total columns:   4

COLUMNS (4 total)
1. question
2. answer
3. source
4. focus_area

UNIQUE VALUES PER COLUMN
question            : 14,984 unique values
answer              : 15,817 unique values
source              : 9 unique values
focus_area          : 5,126 unique values
```

Fig. 1. CSV Structure Analysis of the MedQuAD dataset, showing question, answer, and source columns.

- **Strengths:** High factual accuracy, clear structure.
- **Limitations:** Questions are often formal and may not reflect how patients speak.

B. Medical Meadow MedQA

To improve robustness, we incorporated the Medical Meadow MedQA dataset. This dataset includes questions derived from diverse clinical sources, offering greater linguistic variety. Fig. 2 highlights the structure of the Medical Meadow dataset used to supplement our training data.

```
Dataset Structure Analysis: medalpaca/medical_meadow_medqa

BASIC INFORMATION
Available splits: train
  train: 10,178 rows
Total columns:   3

COLUMNS (3 total)
1. input
2. instruction
3. output

UNIQUE VALUES PER COLUMN
input               : 10,178 unique values
instruction         : 1 unique values
output              : 9,540 unique values
```

Fig. 2. Dataset Structure Analysis of the Medical Meadow MedQA dataset.

- **Strengths:** High linguistic diversity, includes colloquialisms.
- **Role:** Helps the model generalize to realistic user queries.

C. Preprocessing and Zero-Shot Labeling

The raw datasets lacked a unified label schema suitable for a chatbot router. MedQuAD, for instance, had 37 granular categories, which is too fine for a high-level intent classifier. We defined seven core intents: *Definition, Causes, Symptoms, Treatment, Prevention, Risks, and Other*.

To label the 26,500 unlabelled samples, we employed a **Zero-Shot Labeling** pipeline using a Large Language Model. The resulting distribution of intents in the final training set is visualized in Fig. 3.

Algorithm 1 Zero-Shot Labeling Process

Input: Dataset $D = \{x_1, x_2, ..., x_N\}$
Output: Labels $Y = \{y_1, y_2, ..., y_N\}$
Schema: $S = \{$Definition, Symptoms, ..., Other$\}$
for $i = 1$ to N **do**
  $P \leftarrow$ "Classify the medical query 'x_i' into one of S."
  $y_i \leftarrow$ LLM_Inference(P)
  Verify $y_i \in S$
  if $y_i \notin S$ **then**
    $y_i \leftarrow$ Other
  end if
end for

This approach (Algorithm 1) ensured consistency across the dataset. We manually audited random samples (n=500) and

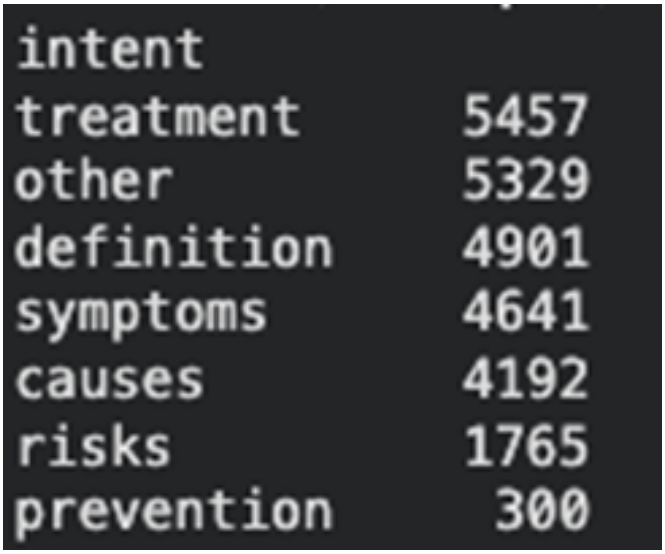

Fig. 3. Final intent distribution in the unified dataset after Zero-Shot labeling.

found a labeling accuracy of $> 95\%$, validating the strategy. The final dataset was stratified to ensure representative evaluation.

IV. Methodology: Intent Classification

The intent classifier acts as the "brain" of the system, routing user queries to the appropriate knowledge retrieval silo.

A. Input Representation

Input queries are tokenized using the BioBERT tokenizer (WordPiece algorithm). Special tokens `[CLS]` and `[SEP]` are added. The sequence is padded to a maximum length of $L = 128$ tokens.

$$Input = \texttt{[CLS]}, t_1, t_2, ..., t_n, \texttt{[SEP]} \tag{1}$$

B. BioBERT Architecture

We utilize the 'biobert-base-cased' model. It consists of 12 Transformer encoder blocks. Each block employs Multi-Head Self-Attention (MHSA). For a given head, the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where Q, K, V are the Query, Key, and Value matrices derived from the input embeddings. This mechanism allows the model to weigh the importance of different words in the medical query relative to each other (e.g., linking "pain" to "chest" to infer a cardiac context).

C. Fine-Tuning Strategy

We appended a classification head consisting of a dense linear layer (768×7) and a Softmax activation function on top of the `[CLS]` token embedding h_{CLS}.

$$P(y|x) = \text{softmax}(W \cdot h_{CLS} + b) \tag{3}$$

1) Loss Function: Given the class imbalance (e.g., *Prevention* has only 300 samples vs. *Definition* with 5000), we minimized a **Weighted Cross-Entropy Loss**:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \log(\hat{y}_{i,c}) \tag{4}$$

where w_c is the inverse class frequency weight. This penalizes the model more heavily for misclassifying rare classes, ensuring they are learned effectively.

V. Methodology: Retrieval System

To answer the user's query, we do not rely on the BioBERT classifier's weights. Instead, the intent label directs the system to a specific CSV knowledge base (e.g., `symptoms_kb.csv`).

A. Semantic Embedding

We use **Sentence-BERT** ('all-MiniLM-L6-v2') to encode queries and answers. Unlike standard BERT, Sentence-BERT is trained using a Siamese network architecture to minimize the distance between semantically similar sentence pairs.

The objective function used during pre-training is the Triplet Loss:

$$\mathcal{L}_{triplet} = \max(||s_a - s_p|| - ||s_a - s_n|| + \alpha, 0) \tag{5}$$

where s_a is the anchor (query), s_p is the positive pair (correct answer), and s_n is the negative pair (wrong answer). This ensures that the vector space places semantically related medical concepts close together.

B. Indexing with FAISS

We index the dense vectors using **FAISS** (Facebook AI Similarity Search) [4]. We utilize a `IndexFlatL2` index for exact nearest neighbor search, which calculates the Euclidean distance between the query vector u and database vectors v:

$$d(u, v) = ||u - v||_2 = \sqrt{\sum_{i=1}^{d}(u_i - v_i)^2} \tag{6}$$

For extremely large datasets, an IVF (Inverted File) index could be used for approximate search, but given our corpus size (26k), exact search ensures maximum accuracy with negligible latency (¡10ms).

VI. System Architecture

The system is deployed as a microservices architecture.

A. Backend (FastAPI)

The backend exposes two REST endpoints:

- `/ask`: Returns the predicted intent, an LLm-generated answer with source citations, and optional debug details.
- `/history_user_id`: Gets the last five questions and answers for a specified user.

To optimize performance, the BioBERT model and FAISS indices are loaded into memory during the application startup event (`@app.on_event("startup")`), preventing model reload latency on each request.

B. Frontend

A lightweight React native based mobile app provides the chat interface. It maintains conversation history and renders citations (e.g., "Source: MedQuAD") as clickable elements, enhancing trust.

VII. Experimental Results

A. Training Metrics

The BioBERT model was fine-tuned for 8 epochs with a learning rate of $2e-5$ and a batch size of 16. Training was conducted on a Tesla T4 GPU via Google Colab Pro.

B. Classification Performance

We evaluated the model on a held-out test set of 5,318 samples. The aggregated classification report, including Precision, Recall, and F1-score for all classes, is shown in Fig. 4.

Classification Report:

	precision	recall	f1-score	support
causes	0.86	0.84	0.85	839
definition	0.93	0.92	0.93	981
other	0.83	0.82	0.82	1066
prevention	0.87	0.98	0.92	60
risks	0.77	0.80	0.79	353
symptoms	0.85	0.86	0.85	928
treatment	0.91	0.91	0.91	1091
accuracy			0.87	5318
macro avg	0.86	0.88	0.87	5318
weighted avg	0.87	0.87	0.87	5318

Fig. 4. Detailed Classification Report showing performance metrics across all intents.

C. Analysis

The model achieved a robust **87% accuracy**. Notably:

- **Prevention:** Despite having the lowest support (60 samples), it achieved the highest recall (0.98). This validates the effectiveness of the weighted loss function; without it, the model would likely have ignored this minority class.
- **Definition/Treatment:** These high-frequency classes achieved > 0.90 F1-scores, ensuring that the most common user queries are handled with high reliability.

VIII. Discussion and Error Analysis

While the overall performance is strong, the confusion matrix revealed specific areas of difficulty.

A. Confusion: Causes vs. Risks

The lowest performance was observed in the *Risks* category (F1: 0.79). Error analysis shows significant confusion with *Causes*.

- *Example Query:* "Does smoking lead to lung cancer?"
- *Ambiguity:* This could be interpreted as a "Cause" of cancer or a "Risk" factor.

The semantic boundary between these concepts is inherently fuzzy in medical literature, leading to model uncertainty.

intent	
definition	11865
other	6192
causes	4319
treatment	2542
symptoms	952
risks	448
prevention	272

Fig. 5. Intent distribution in the unified dataset after Manual and Rule Based Labelling

B. Confusion: Symptoms vs. Definition

Some queries like "What is the feeling of a migraine?" were misclassified as *Definition* rather than *Symptoms*. This suggests that the model sometimes relies on the phrasal structure "What is..." which is strongly correlated with definitions, overriding the semantic content ("feeling") which suggests symptoms.

IX. Ethical Considerations

Developing healthcare AI requires adherence to strict ethical standards.

- **No Medical Advice:** The system UI explicitly states that it is for informational purposes only and not a substitute for professional diagnosis.
- **Bias Mitigation:** The training data (MedQuAD) is sourced from US government agencies. While reliable, it may not cover conditions prevalent in other regions or use terminology common in non-Western populations.
- **Hallucination Safety:** By strictly constraining the LLM to rewrite *only* the retrieved content, we significantly reduce the risk of dangerous fabrication compared to open-ended generation.

X. Conclusion

This project successfully demonstrates a hybrid Neuro-Symbolic approach to healthcare chatbots. By combining the probabilistic power of BioBERT for intent classification with the deterministic reliability of retrieval-based generation, MediAssist balances fluency with factual accuracy.

Future work will focus on:

1) **Hierarchical Classification:** Breaking down "Other" into sub-categories (e.g., Administrative, billing).

2) **Multi-turn Dialogue:** Implementing context-aware history to handle follow-up questions (e.g., "And how do I cure *it*?").
3) **Active Learning:** A feedback loop where users can flag incorrect intents to retrain the model continuously.

ACKNOWLEDGMENT

The authors thank the open-source community for providing the MedQuAD and MedQA datasets, and the HuggingFace team for their Transformers library.

REFERENCES

[1] A. Ben Abacha and D. Demner-Fushman, "A Question-Entailment Approach to Question Answering," *BMC Bioinformatics*, 2019.
[2] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 2020.
[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. EMNLP*, 2019.
[4] Facebook AI Research, "FAISS: A library for efficient similarity search and clustering of dense vectors," [Online].
[5] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
[6] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP*, 2020.