Multilingual Clinical Symptom Understanding and Triage Model with Conversational Integration

Ananya Singla, Anoushka Yadav, Kabir Gupta, Krish Gupta, Yuven Blowria
Plaksha University, Punjab, India

Abstract

India's linguistic diversity presents a major challenge in building accessible digital healthcare systems. Most existing AI-driven medical assistants are limited to English and lack domain-specific clinical understanding. This project develops a custom multilingual medical NLP model capable of interpreting patient-described symptoms across multiple Indian languages and categorizing them into appropriate clinical departments. The model performs symptom extraction, medical category classification, and urgency estimation. To demonstrate real-world applicability, this model is integrated into a conversational chatbot interface powered by Google Gemini, where Gemini handles natural conversation while the custom-trained model provides core medical intelligence.

1 Introduction

Recent advances in artificial intelligence have accelerated the adoption of digital tools for healthcare assistance, particularly in preliminary symptom assessment and patient interaction. Despite this progress, most existing AI-based medical systems remain English-centric and rely on general-purpose language models with limited clinical awareness. In linguistically diverse regions such as India, where patients frequently describe symptoms using regional languages and colloquial expressions, this significantly reduces accessibility and diagnostic reliability.

Concurrently, healthcare institutions experience growing patient loads, making early-stage triage and department routing essential for efficient service delivery. Automated systems capable of understanding multilingual patient inputs while preserving medical relevance can help alleviate this burden.

To address these challenges, this work introduces a hybrid framework that integrates a custom-trained multilingual medical NLP model with a conversational interface powered by Google Gemini. The conversational component manages dialogue and workflow orchestration, while the specialized model performs symptom extraction, medical department classification, and urgency estimation. This architecture aims to support scalable, language-inclusive preliminary clinical decision workflows.

2 Literature Survey

Prior work in medical natural language processing has focused on clinical entity recognition, symptom extraction, and automated triage using transformer-based architectures. Biomedical adaptations of BERT, including BioBERT and ClinicalBERT, have achieved strong performance on English clinical notes for tasks such as named entity recognition and diagnosis classification. However, these models are primarily trained on hospital-generated records and do not generalize well to patient-authored symptom descriptions, which are often informal and linguistically diverse.

Recent studies on Indian-language healthcare NLP highlight challenges related to code-mixing, regional expressions, and limited annotated data, particularly for symptom-level entity extraction (Mullick et al., 2023; Patel et al., 2026). While multilingual transformers such as XLM-R, MuRIL, and IndicBERT enable cross-lingual representation learning, their application to domain-specific medical triage in Indian languages remains limited.

Several multilingual clinical pipelines have explored weak supervision, back-translation, and adapter-based fine-tuning to address data scarcity, demonstrating the effectiveness of synthetic multilingual augmentation for medical NER and classification (Sallauka et al., 2025). These approaches motivate the use of translation-based augmentation and task-specific fine-tuning adopted in this work.

Conversational healthcare assistants increasingly leverage large language models for dialogue management, yet recent evaluations show that LLM-based triage systems exhibit inconsistent performance and limited clinical reliability, especially outside English (Lafuente and Rahim, 2025). These findings suggest that triage in Indian languages remains an open problem requiring targeted datasets and structured medical pipelines rather than end-to-end reliance on general-purpose LLMs.

Motivated by these limitations, this work adopts a hybrid architecture that separates conversational interaction from clinical inference by integrating a custom multilingual medical NLP model with a Gemini-powered chatbot. This design enables structured symptom extraction, medical department classification, and urgency estimation while preserving natural user engagement, addressing a gap in existing multilingual triage frameworks for Indian healthcare settings.

3 Problem Statement

Patients often describe symptoms in regional languages or informal mixed-language text (e.g., *Hinglish*), which makes automated triage difficult. Existing systems are largely English-centric and do not generalize well to Indian healthcare contexts.

There is a lack of:

- Multilingual clinical NLP models tailored to Indian languages
- Automated symptom-to-specialty classification systems
- Lightweight triage tools for early medical routing

This project addresses these gaps by building a multilingual symptom understanding model that assigns patient inputs to medical categories and supports preliminary triage.

4 Objectives

The primary objectives of this project are:

1. Build a multilingual medical NLP model supporting major Indian languages
2. Extract clinical entities such as symptoms and duration from free-text input
3. Classify patient complaints into predefined medical specialties
4. Estimate urgency levels for basic triage
5. Integrate the model into a conversational chatbot for real-world interaction

5 System Architecture

The proposed system follows a hybrid architecture consisting of a custom medical machine learning pipeline integrated with a conversational chatbot interface. The design separates domain-specific medical intelligence from general conversational capabilities to ensure both clinical relevance and natural user interaction.

- **Multilingual Clinical Encoder:** A transformer-based model fine-tuned on multilingual clinical and symptom-description datasets to handle patient inputs across multiple Indian languages.
- **Symptom Extraction:** Named Entity Recognition (NER) is applied to identify key medical entities such as symptoms, duration, and severity from free-form patient text.
- **Medical Category Classification:** Extracted symptoms are mapped to appropriate clinical departments (e.g., cardiology, neurology, general medicine) using supervised classification.
- **Urgency Prediction:** A dedicated prediction head estimates triage priority levels to support early-stage patient routing.
- **Conversational Interface:** A chatbot powered by Gemini manages natural dialogue, user engagement, and workflow orchestration, while delegating core medical inference tasks to the custom ML pipeline.
- **Integration Layer:** REST-based APIs connect the chatbot with the ML backend, enabling real-time inference and structured response generation.

6 Methodology

6.1 Multilingual Representation

A pretrained multilingual transformer (such as IndicBERT, MuRIL, or XLM-R) is used as the base encoder and fine-tuned on medical text.

6.2 Dataset Construction and Multi-source Fusion

To train a robust multilingual triage model, we combine high-quality clinical corpora with synthetic augmentation, ensuring compliance with data privacy and licensing:

- **CUI-Symptom Mapping:** We utilize the **UMLS (Unified Medical Language System)** and **MeSH** hierarchies to create a taxonomy of symptoms and their corresponding medical specialties.
- **Public Clinical Corpora:**
  - **MTSamples:** Short, de-identified clinical notes used for specialty classification (CC-BY License).
  - **i2b2/VA Challenge Data:** Used strictly for NER (Symptom extraction) training, adhering to restricted-use agreements.
  - **Kaggle Ayurvedic Dataset:** To capture regional descriptions of symptoms and traditional medicine terminology prevalent in Indian contexts.
- **Multilingual Augmentation:** Since annotated clinical data in Indian languages is scarce, we employ back-translation to synthetically expand multilingual training samples. English symptom descriptions are translated into Hindi, Tamil, and Telugu using NLLB-200 (No Language Left Behind), providing domain-aligned multilingual supervision for model fine-tuning.

7 Data Sources and Licensing

Table 1 outlines the primary datasets, their accessibility, and their role in the multi-task learning pipeline.

Dataset	Task	License	Access
MTSamples	Specialty Classification	CC-BY 4.0	Public
i2b2/VA	Symptom NER	Restricted	Controlled
Kaggle Ayurvedic	Symptom Modeling	Public	Public
UMLS / MeSH	Ontology Mapping	Research Use	Controlled

Table 1: Summary of datasets and resources used in the multilingual triage pipeline.

7.1 Model Tasks

The model is trained in a multi-task setup:

- **Symptom extraction**: Named Entity Recognition
- **Department classification**: Respiratory, Orthopedic, Dermatology, etc.
- **Urgency prediction**: low, medium, high

7.2 Chatbot Integration

The trained model is exposed as an API and consumed by the Gemini-powered chatbot. Gemini manages conversation flow, appointment handling, and user interaction, while medical predictions are obtained from the custom model.

8 Novelty and Innovation

This project introduces several novel contributions:

- Custom multilingual clinical NLP model for Indian languages
- Symptom-to-specialty classification rather than generic chatbot responses
- Separation of conversational intelligence (Gemini) from medical intelligence (custom ML)
- Synthetic multilingual clinical dataset creation

9 Expected Outcomes

The project aims to deliver the following outcomes:

- Trained multilingual medical classification model
- Symptom extraction and triage system
- Conversational chatbot demonstrating real-world usability
- REST API for medical inference
- Web-based demo interface

10 Conclusion

This work presents a hybrid multilingual clinical triage system that combines a custom-trained medical NLP model with a conversational chatbot interface. By separating conversational intelligence from clinical inference, the proposed framework

enables structured symptom understanding while supporting natural user interaction. The system performs symptom extraction, medical department classification, and urgency estimation across multiple Indian languages, addressing key accessibility challenges in digital healthcare.

Although designed as a preliminary triage assistant rather than a diagnostic tool, the project demonstrates the feasibility of integrating domain-specific multilingual models into conversational healthcare applications. Future extensions may include broader language coverage, voice-based input, and integration with electronic health record systems.

Acknowledgments

This work is part of the Advanced NLP course project at Plaksha University. We acknowledge the use of publicly available clinical datasets and open-source multilingual language models.

References

Carlos Lafuente and Mehdi Rahim. 2025. Evaluating large language models on hospital health data for automated emergency triage. *International Journal of Computer Assisted Radiology and Surgery*, 20.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R. R. Raghav, G. Sai Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics.

Kiran Patel, Meera Singh, and Rohan Das. 2026. Bridging languages in healthcare: A comprehensive review of multilingual and code-switched clinical assistants. *Journal of Healthcare NLP*, 4(1):12–25.

Rigon Sallauka, Umut Arioz, Matej Rojc, and Izidor Mlakar. 2025. Weakly-supervised multilingual medical ner for symptom extraction for low-resource languages. *Applied Sciences*, 15(10):5585.