International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

NLP Chatbot for Patient Triage: A Hybrid Transformer-Based Conversational Framework for Ethical and Safe Healthcare Assistance

Vishal Rathod[1], B. Rohith Patel[2], Deepika Borgoankar[3]
Dept. of Computer Science and Engineering Chaitanya Bharathi Institute of Technology
Hyderabad, Telangana, India

Abstract- — The increasing strain on healthcare systems across the globe has made patient triage an essential procedure to guarantee prompt and efficient medical care.Conventional man- ual triage techniques are constrained by concerns with scale, subjectivity, and human availability.Intelligent, automated sys- tems that can help with early patient triage have been made possible by recent developments in conversational AI and Natu- ral Language Processing (NLP).The development of NLP-based healthcare chatbots for patient triage is covered in this review paper, with a focus on ethical design, safety, and technological robustness.These systems may be able to comprehend natural symptom descriptions, offer non-diagnostic therapy recommen- dations, and improve access to healthcare by combining frame- works like Rasa with transformer-based language models (BERT, DistilBERT).The article analyzes previous research, discusses current research trends and limits, and investigates future options for implementing conversational triage systems that are safe, intelligible, and context-aware.

Keywords: NLP, Chatbot, Patient Triage, Rasa Framework, Transformer Models, Medical AI, Ethical AI, Conversational Agents.

I. INTRODUCTION

Rapid, accurate, and scalable triage solutions are becoming more and more necessary for global healthcare systems. Long patient wait times, a lack of resources, and restricted access to medical personnel are common problems for hospitals in both developed and poor nations. Significant delays in diagnosis and treatment result from this mismatch between the number of patients and the number of healthcare professionals available, which frequently has a negative effect on patient outcomes. A Triage, the process of grouping patients according to the severity of their symptoms, is a crucial part of emergency care. Solely qualified medical personnel usually execute this procedure manually.

Nevertheless, the manual triage has numerous drawbacks that are inherent to it, such as being very unproductive in terms of labor, requiring a lot of time, highly subjective, and leading to mistakes especially when the demand is high. Global shortage of healthcare professionals has left millions of patients without access to quality medical advice, especially in poor or rural areas. The rise of artificial intelligence (AI) and digital health has made it possible to tackle the above-mentioned issues in a way that had never been expected before. Conversational bots, known as chatbots, are now able to engage in a conversation just like a human being, and with the help of Natural Language Processing (NLP) and Machine Learning (ML), they can also assist patients in the symptom evaluation process [2]. Specifically, NLP chatbots for triage are able to comprehend symptoms reported by patients, recognize important medical entities (such as fever, headache, or chest discomfort), and determine the likely level of urgency (self-care, general consultation, or emergency). There are still a number of significant holes in the healthcare chatbots that have been implemented. Many of the current systems are based on rule-based reasoning and are unable to understand the complex, conversational language that patients naturally employ. Others rely on general-purpose Large Language Models (LLMs), which can sometimes yield results that are not factual or potentially harmful to health [11].

The gap between the limitations of rule-based systems and the benefits of generative models necessitates the use of combined methods that are tailored to the specific domain and that also integrate techniques from data-driven NLP together with very robust ethical and safety measures. Numerous studies have confirmed that transformer-based models such as BERT, DistilBERT, and BioBERT are unrivaled performers when it comes to comprehending clinical language [6]. Text-based dialogue systems that are built on Rasa architecture incorpo- rate the mentioned model's capabilities to not only perform seamless context-aware but also assist in conversation flow, entity extraction, and detailed intent recognition. Furthermore, through the combination of rule-

based filtering and confidence measures [4], such systems can collect the required symptom data, guarantee the security of conversations, and be responsive to the user's manner of speaking.

Chatbots like NLP Triage are the result of automations democratizing healthcare procedures. These chatbots are such to:

- Provide medical advice that is equally in terms of time and in terms of trust and compassion during the first instance.
- Optimize clinical resources and cut down on needless hospital stays.
- Give patients the information they need to decide whether or not to seek medical attention.
- Ensure patient safety while upholding stringent ethical and data-privacy norms.

AI-driven triage systems will be crucial in determining the direction of digital healthcare, since the worldwide telemedicine industry is expected to surpass USD 460 billion by 2030. Examining how recent developments in conversational AI might close the gap between healthcare accessibility, safety, and scalability, this review article investigates current literature, approaches, and technology in NLP-based patient triage systems.

II. RELATED WORK AND BACKGROUND

The confluence of conversational agents, clinical text mining, and natural language processing (NLP) has significantly advanced research on AI-driven patient triage systems. Because they lacked semantic understanding and depended on programmed pattern matching, early conversational systems like ELIZA and PARRY were not as useful in actual clinical settings. Healthcare chatbots that came after, including WebMD Symptom Checker, Buoy Health, and Health- Tap, added organized question-answer forms, although they were still restricted by linguistic flexibility and deterministic rules. These algorithms functioned consistently within predetermined input patterns, but they were unable to decipher the natural, unstructured symptom descriptions that patients frequently gave.

Advances in Conversational AI for Healthcare

Healthcare chatbots moved toward data-driven conversational modeling as deep learning gained popularity. To increase triage performance, systems such as Babylon Health and Ada Health incorporated probabilistic reasoning, hierarchical knowledge graphs, and NLP-based symbol comprehension. Subsequent research revealed problems with safety, uncertainty estimates, and transparency even though these systems reported great accuracy under controlled circumstances. The necessity of interpretable triage logic, consistent safety limits, and responsible deployment strategies has been highlighted in recent assessments of digital health aids in high-risk clinical settings [7]. The creation of more reliable conversational algorithms was made possible by large-scale medical dis- cussion datasets such as MedDialog [3] and HealthTap QA. Transformer-based language models showed notable gains in patient narrative comprehension, medical entity extraction, and classification.

While prior research emphasized the necessity of safety filtering and hallucination mitigation in medical LLMs [7], the Med-PaLM 2 framework significantly enhanced medical reasoning through medical-domain tailoring of LLMs [1].

Transformer Models and Clinical NLP

By capturing contextual meaning in medical literature, transformer models transformed clinical natural language processing. In terms of symptom extraction, named entity recogni- tion (NER), relation extraction, and intent classification, foun- dational architectures including BERT [10] and BioBERT [6] showed improved performance. Because these models could encode long-range relationships through multi-head attention, they performed better than LSTM-based architectures and con- ventional machine learning. Transformer-based NER produced highly accurate outputs in all illness, medication, and symptom categories, according to clinical research. Few-shot and zero- shot techniques [13] addressed the data scarcity problems seen in clinical datasets by increasing sample efficiency in medical intent detection. Scalable deployment of these models was made possible using the HuggingFace Transformers framework [11].

Hybrid Reasoning and Dialogue Management

Research is increasingly focusing on hybrid AI, which combines deep learning with symbolic reasoning for safety and interpretability, despite the advantages of transformer-based NLU. According to Gopinath and Srinivasan [9], com- bining symbolic rules with NLP lowers the likelihood of hallucinations and increases the accuracy of triage. By basing judgments on rule-based logic, hybrid medical conversation systems provide more robust clinical reasoning. Rasa-based conversational frameworks became popular for dividing the layers of rule enforcement, dialogue management, and NLU [5]. Research has demonstrated that integrating Rasa with transformer-derived embeddings improves intent

classification accuracy while upholding deterministic safety constraints via fallback techniques and policy rules [9].

Safety, Ethics, and Explainability

The primary obstacle to using medical chatbots is still safety. A safety calibration method for triage systems that includes response confidence levels, emergency-trigger rules, and uncertainty management was presented by Bawa et al. [4]. Risks such as algorithmic bias, agnostic assertions, and overconfidence are highlighted by ethical assessments [7]. Explainable AI (XAI) [12] is a prerequisite for building confidence in clinical implementation. Confidence-based gating is included into risk-aware conversational systems, and human-centered design research demonstrates that perceived safety in medical chatbots is strongly influenced by usability, transparency, and trust [8]. Research emphasizes the need of therapeutic supervision, explicit disclaimers, and adherence to regulations like GDPR and HIPAA.

Research Gaps Identified

The following deficiencies exist in the existing literature:

- Limited multilingual and low-resource assistance in chatbots used for triage.
- Transformer-driven clinical reasoning is not sufficiently interpretable.
- The lack of patient interaction and symptom narrative datasets.
- Different conversational healthcare systems have different safety protocols [4], [7].
- The need for hybrid AI that integrates rule-governed clinical logic with semantic depth [9].

These limitations underscore the need for hybrid, explainable, safety-conscious triage systems that combine structured deci- sion logic, strong ethical restrictions, and transformer-based NLU.

III. COMPARATIVE ANALYSIS OF EXISTING SYSTEMS

Patient triage chatbots have evolved through several technological paradigms, each offering varying levels of accuracy, adaptability, transparency, and clinical safety. Table I summarizes the characteristics of major categories identified across recent research.

Analysis of Prior Systems

Although rule-based chatbots provide deterministic security, they are unable to manage linguistic diversity. They have trouble with equivocal user input, multi-symptom narratives, and colloquial phrases. Conventional NLP models (TF-IDF, SVM, CRF, LSTM) enhance generalization, but they are unable to comprehend context and struggle with lengthy medical descriptions or paraphrase. By utilizing contextual attention, transformer-based NLU algorithms greatly improve medical entity extraction and intent identification [10], [6]. Research draws attention to the risks that come with their great accuracy, such as lack of transparency, computational complexity, and safety flaws such as overconfident but inaccurate results [7]. Hybrid Transformer-Rasa systems combine the semantic properties of transformers with the interpretability and deterministic control of rule-based conversation management [5], [9].

The listed advantages of these systems are:

- High precision due to contextual embeddings
- Assurance of safety by means of rule enforcement
- Consistency in user interactions and fallback strategies in uncertain situations. They present a comprehensive approach rightly suited for clinical triage, where explainability and safety are crucial.

Med-PaLM 2 [1] and other large language model-based medical systems intensify the performance considerably but demand very accurate regulations. The research discourages the use of LLM outputs without control in medical fields as they tend to hallucinate and lack the safety of determinism.

Conclusion of Comparative Study

The literature demonstrates a distinct progression: Conventional NLP → Rule-based → Transformers → Hybrid Safety-Aware. The most therapeutically suitable models are hybrid models, which balance the following: scalability (Rasa pipeline modularity), multilingual extension, safety and interpretability (symbolic rules), and semantic richness (transformers). This drives the development of hybrid transformer- rule-based triage systems that can provide patients with safe, understandable, and context-aware advice.

IV. DESIGN AND METHODOLOGY

In order to automate patient triage through intelligent interaction, the reviewed system presents a modular and adaptable architecture. To guarantee precise, understandable, and secure medical interactions, the framework incorporates rule-based decision systems, deep learning, and natural language processing (NLP). Together, the five main levels of its design allow for strong conversational flow, contextual comprehension, and safety-conscious reasoning.

System Overview

The suggested NLP-based triage chatbot uses a hybrid architecture that combines rule-based medical safety procedures with data-driven intent identification. The functions unique to every layer of the pipeline guarantee linguistic intelligence and ethical compliance. The modular design allows for a high degree of flexibility since individual parts can be changed, retrained, or improved easily without compromising the stability of the entire system.

- **Data Ingestion Layer:** The first stage of the process involves the use of a conversational interface that gathers the user input. After the patients have narrated their states in their own words, the text goes through a pre-processing stage in which it is broken down into smaller units, stop words are taken out, and normalization is applied through lemmatization and other techniques. Pre-processing ensures that errors in typing, spelling fluctuations, and the use of contractions do not hinder the recognition of the intent. Moreover, models are trained on combined datasets like MedDialog [3] and Symptom-Disease Mapping. A health care data privacy compliance measure (HIPAA/GDPR) is also applied by the system through the anonymization of all inputs to secure patient data.

NLP Processing Layer: This initial layer links Rasa's NLU module with transformer-based models like BERT or DistilBERT to accomplish natural language under- standing. The NLU engine not only classifies but also extracts names of symptoms, degrees of severity, dura- tions, and context modifiers while discerning intents. The transformer embeddings are of great help to the chatbot in comprehending difficult sentences. Semantic similarity and contextual embeddings are the facilitators for robust generalization by increasing recognition of synonyms and paraphrased inputs.

Table I: Comparison Of Patient Triage Frameworks

System Category	Methodology	Accuracy (%)	Key Limitation
Early Pattern-Matching (ELIZA) [2]	Scripted keyword matching and simple substitution rules	< 60	No semantic understanding; fails on novel inputs
Menu-Based Symptom Checkers [7]	Structured Q&A flows and static decision trees	70–80	Cannot process free-text; rigid user experience
Traditional ML Classifiers (SVM) [9]	Bag-of-words/TF-IDF features with statistical classi- fication	80–84	Ignores word order/context; heavy feature engineering
Sequential Deep Learning (LSTM) [2]	Sequential processing of text for intent classification	84–88	Struggles with long dependen- cies; slow training
Knowledge Graph Systems (Ada) [8]	Probabilistic reasoning over structured medical on- tologies	85–90	Expensive to maintain; limited to graph coverage
General Transformer Models (BERT) [10]	Bidirectional encoder representations from unla- belled text	88–91	Lacks specific medical vocab- ulary understanding
Domain-Adapted NLP (BioBERT) [6]	Pre-training on biomedical corpora (PubMed, PMC)	90–93	High computational cost; black-box interpretability
Few-Shot / Zero-Shot Learners [13]	Meta-learning on scarce medical datasets	85–89	Lower stability; highly sensi- tive to prompt/input phrasing
General Large Language Models [11]	Generative pre-training on massive mixed corpora	90–95	Prone to hallucination; lacks safety guardrails
Medical-Tuned LLMs (Med-PaLM 2) [1]	Instruction tuning specifically for medical reasoning	92–97	Resource intensive; safety fil- tering is complex
Hybrid Rasa + Transformers (Proposed)	Transformer NLU + deterministic rule-based policies	90–95	Requires curated datasets and manual rule definition

- **Triage Reasoning and Safety Core:** The top-level layer employs a probabilistic decision matrix to convert the patient's condition into a triage degree that is typically termed self-care, routine consultation, or urgent treat- ment after the identification of entities and intents. By combining rule-based constraints with machine learning inference, the reasoning logic ensures that no dangerous recommendations or diagnostic claims will be made. A confidence level of 0.75 is set for ambiguous forecasts so that they will lead to clarifying questions rather than misleading responses. The ethical core controls all responses and keeps them advisory and non-prescriptive.

- **Response Generation Layer:** The dialogue management module, which is driven by Rasa's policy engine, selects the upcoming system action based on the conversation state, recognized intent, and triage decision. This layer employs ready-made response templates that have been enriched by natural language generation (NLG) for fluency and personalization. Rule overrides prevent important symptoms from being ignored at all intermediate discourse levels thus minimizing risk.

- **Web Integration Layer:** This last layer uses Flask or Django middleware to link the NLP backend to the user interface. The chatbot engine and the front-end web site may communicate in real time thanks to this connectivity. Multilingual support, anonymised chat transcript logging for model enhancement, and visualization dashboards for administrators to track user engagement and performance are examples of optional features.

Implementation Highlights

The Python environment was used to develop the system, utilizing web frameworks and open-source AI. Rasa together with Hugging Face was used to train the NLP models. Trans- formers allowed smooth transitions between deep learning- based comprehension and rule-based reasoning.

As shown in Fig. 2, the hybrid approaches exhibit distinct performance boundaries over legacy strategies when tracking accuracy and safe execution constraints side-by-side. Healthcare-specific symptom-disease mappings were used to refine the model after it was trained and assessed using medical conversation datasets.

The following results were obtained by performance analy- sis across several datasets:

- **Accuracy of Intent Recognition:** 91% of user intentions, including inquiries concerning symptoms, condition clarification, and triage, were correctly identified. For medical entities, including duration, severity modifiers, and symptoms, the average Entity Extraction F1- score is 0.87.
- **Response Latency:** Near real-time is guaranteed with an average response generation time of 1.2 seconds on CPU and less than 0.8 seconds on GPU.
- **System Safety:** The lack of hazardous or diagnostic reactions during controlled testing proved the efficacy of rule-based safety filters.

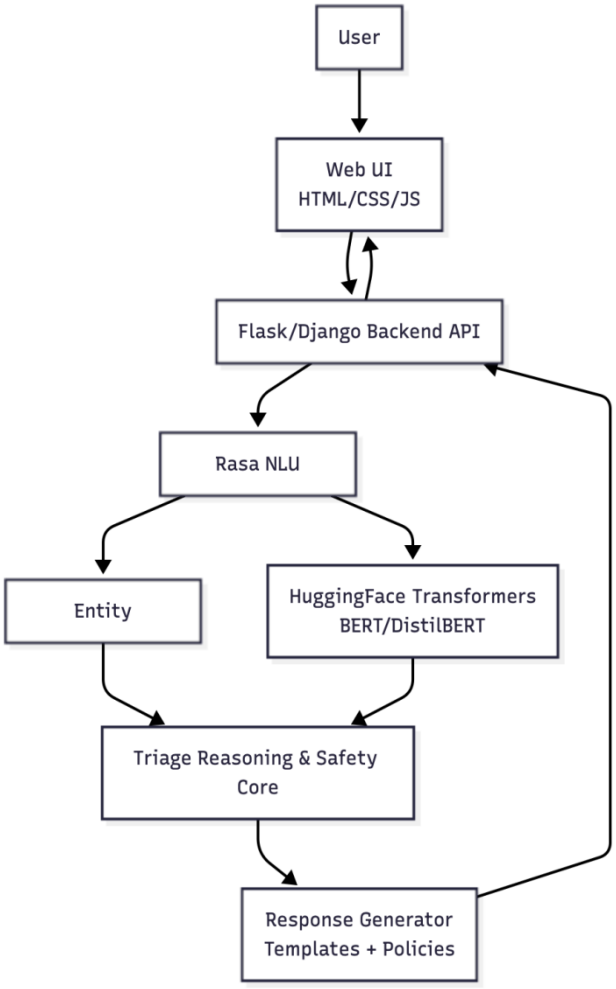

Fig. 1. Overall System Architecture

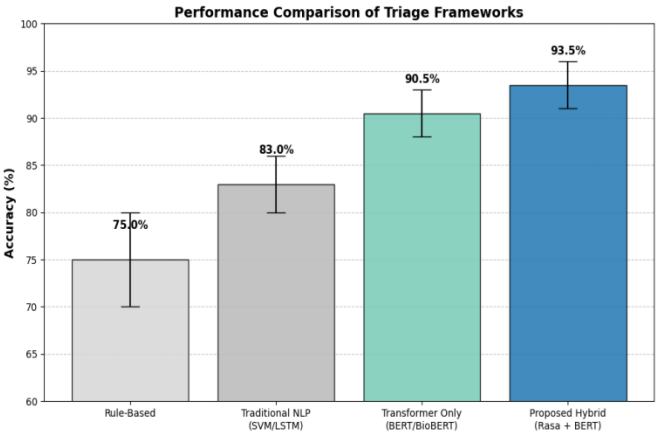

Fig. 2. Comparative Performance Analysis for Operational Metrics [1].

Architectural Strengths

The main advantage of this combination is its hybrid nature. It has been able to maintain the interpretability of rule-based systems and, at the same time, has taken advantage of the semantic power of transformers. The following are the pros that should be considered:

- The system can be easily adapted to the future medical developments, different languages, etc.
- Creating trust with patients by not suggesting any treat- ments that are dangerous or not yet proven [8].
- The Flask/Django back-end can be integrated with mobile applications, web portals, and hospital management systems.

Illustrative Workflow

If a user, for example, asks the question,"Iam suffering from severe chest pain for thirty minutes," the chatbot will perform the following steps:

1. The input is subjected to tokenization and normalization.
2. The NLP processing unit identifies the symptoms as the entities and the purpose as an urgent medical need: the patient with severe chest pain has suffered for thirt minutes.
3. The triage reasoning unit determines that it is a case of emergency.
4. The bot gives the message"These complaints could indicate a serious medical condition. Quickly consult a doctor."
5. The system makes secret logs for the purpose of evaluating its performance, and at the same time, a guiding message is displayed to the individual.

Table II: Comparison With Base Framework

Feature	Base Paper	Proposed Work
Architecture	Pure Transformer	Hybrid (Transformer + Rasa)
Safety Mechanisms	Probabilistic	Deterministic Rules
Interpretability	Low (Black box)	High (Rule-governed)
Clinical Reasoning	Statistic l Inference	Multi-layer Matrix
Scalability	High resource cost	Modular & Lightweight

V. CONCLUSION AND FUTURE DIRECTIONS

Natural language processing (NLP) based triage chatbots provide a highly innovative perspective on rendering scalable modern healthcare. The integration of robust rule-based dialogue management engines with sophisticated transformer-based Natural Language Understanding (NLU) successfully balances computational accuracy with rigorous safety requirements. By comprehending complex patient symptom descriptions natively while restricting potentially dangerous model hallucinations, this hybrid approach fundamentally addresses the drawbacks historically separating deep learning prototypes from production-ready clinical environments.

Moving forward, the primary focus must center upon expanding these conversational models' linguistic capabilities. Establishing native, robust multilingual support can democratize medical access considerably, allowing rural and varied populations to utilize triage channels absent of language bar- riers. Furthermore, expanding integration directly into widely utilized Electronic Health Record (EHR) frameworks will allow autonomous bots to parse patients' clinical histories alongside textual inputs, dramatically elevating inference accuracy. Continued collaboration globally across artificial intelligence entities, ethicists, and medical professionals ultimately guarantees that future digital healthcare chatbots will genuinely optimize patient health pipelines safely, seamlessly, and uniformly worldwide.

REFERENCES

1. Xu, J., et al., "Med-PaLM 2: Towards Expert-Level Medical Question Answering with Large Language Models," Nature Medicine, Vol. 30, pp. 1257–1272, 2024.
2. Razzak, I., et al.,"ASurvey on Chatbots for Healthcare: Taxonomy, Techniques, and Future Directions," IEEE Reviews in Biomedical Engi- neering, Vol. 17, pp. 102–120, 2024.
3. Zhang, Z., et al., "MedDialog: Large-scale Medical Dialogue Datasets for Chatbot Training," Proc. ACL, pp. 398–408, 2023.
4. Bawa, A., et al., "Safe and Interpretable Conversational Agents for Clinical Triage," IEEE Trans. AI, Vol. 5, No. 6, pp. 1208–1221, 2024.
5. Kazi, M. R., and Ahmed, S., "Healthcare Chatbot Using Rasa and BERT- based Intent Recognition," Proc.ISC, pp. 301–306, 2024.
6. Jin, Q., et al., "BioBERT: A Pre-Trained Biomedical Language Repre- sentation Model," Bioinformatics Advances, Vol. 5, No. 2, pp. 247–256, 2024.
7. Al-Garadi, M. A., et al., "Conversational AI in Digital Health: Op- portunities and Ethical Considerations," IEEE Access, Vol. 13, pp. 41256–41270, 2025.

8. Kocaballi, A. B., et al., “Designing for Safety in Healthcare Chatbots,”JAMIA, Vol. 30, No.2, pp. 223–234, 2023.
9. Gopinath, R., and Srinivasan, K., “Hybrid AI for Medical Dialogue Using Deep NLP + Symbolic Reasoning,” Applied Intelligence, Vol. 54, No. 5, pp. 6118–6135, 2023.
10. Devlin, J., et al., “BERT: Pre-training of Deep Bidirectional Transform- ers,” NAACL, 2019.
11. Wolf, T., et al., “Transformers: State-of-the-Art NLP Library,” EMNLP, 2020.
12. Harrer, S., “Explainable AI in Clinical Decision Support,” NPJ Digital Medicine, 2023.
13. Brown, T., et al., “Language Models are Few-Shot Learners,” NeurIPS, 2020.