IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

# Architectural Design and Internal Working of Health Talk: An Offline-Capable, Multilingual AI Conversational Agent for Clinical Triage and Healthcare Facility Discovery in Regional India

**Umesh Kadale[1], ChinmayGawale[2], Arnish Aher[3], Pranav Dhikale[4], Prof. B. S. Shirole[5], Prof. P. V. Nagare[6], Prof. R. M. Shaikh[7]**
Department of Computer Enginerring[1–7]
Loknete Gopinath Munde Institute of Engineering Education & Research (LoGMIEER), Nashik, India

**Abstract**: *In developing nations such as India, primary healthcare triage faces acute structural bottlenecks, including severe doctor-to-population deficits in semi-urban and rural agglomerations, fragmented municipal health informatics, and linguistic barriers across heterogeneous regional populations. While modern conversational agents and large language models (LLMs) offer promising interactive capabilities, their direct clinical deployment is severely hampered by pharmaceutical hallucination risks, non-deterministic reasoning trajectories, prohibitive cloud API costs, and an absence of localized Indic language models. To resolve this critical dilemma, this paper presents the complete architectural design, algorithmic foundations, and empirical validation of Health Talk AI—a novel, offline-first, multilingual conversational triage agent specifically engineered for Tier-2 regional healthcare ecosystems such as the Greater Nashik Metropolitan Area in Maharashtra, India. Health Talk decouples clinical urgency categorization from generative text synthesis by introducing an inviolable, deterministic three-tier triage rule engine that executes prior to any natural language parsing. The architecture synthesizes: (1) an emergency finite-state pre-filter for instant red-flag detection; (2) a multilingual Natural Language Processing (NLP) subsystem featuring Devanagari script tokenization, Hindi versus Marathi marker-word frequency ratio disambiguation, and RapidFuzz partial-ratio entity extraction for pain intensity (1–10 scale), temporal duration, and anatomical body parts; (3) an offline-first Retrieval-Augmented Generation (RAG) vector engine utilizing dense multilingual sentence embeddings (paraphrase-multilingual-MiniLM-L12-v2) coupled with an inviolable cosine dot-product threshold (tau = 0.45) that mathematically eliminates medical hallucinations; (4) a privacy-preserving geospatial discovery subsystem indexing 347 verified healthcare facilities across Nashik using the spherical Haversine metric with two-decimal-place differential coordinate rounding (~1.1 km); and (5) automated state healthcare entitlement checklists (Ayushman Bharat PM-JAY and Mahatma Jyotirao Phule Jan Arogya Yojana - MJPJAY) paired with multi-lingual emergency first-aid response cards. Evaluated across 91 rigorous automated unit tests, Health Talk achieves a 100% test pass rate, 100% Top-1 retrieval precision on clinical queries in English, Hindi, and Marathi, and an end-to-end execution latency under 28 ms on commodity quad-core CPUs without GPU hardware. This work establishes an authoritative, open-access, zero-cloud-cost framework for safe, equitable, and privacy-compliant digital clinical triage.*

**Keywords**: Clinical Decision Support Systems (CDSS), Healthcare Triage Chatbot, Retrieval-Augmented Generation (RAG), Multilingual NLP, Devanagari Script Tokenization, Geospatial Facility Discovery, Zero-Hallucination Guardrails, Explainable Rule Engines, Regional Healthcare

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40475

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

## I. INTRODUCTION

### 1.1 Regional Healthcare Realities and Triage Deficits in India

The delivery of primary healthcare in developing nations presents acute structural and socio-technical challenges, characterized by a persistent disparity in clinical resource allocation between metropolitan centers and Tier-2 or Tier-3 urban agglomerations. In India, the national allopathic doctor-to-population ratio remains below the World Health Organization (WHO) recommended threshold of 1:1,000, falling to approximately 0.65:1,000 in semi-urban and rural regions [1]. In dynamic industrial and agricultural hubs such as the Greater Nashik Metropolitan Area in Maharashtra, this deficit manifests as severe congestion in government secondary and tertiary centers, notably the Nashik District Civil Hospital and rural sub-district hospitals [2]. A substantial proportion of patients attending emergency departments present either with advanced complications resulting from prolonged diagnostic delay or with benign, self-limiting viral conditions that could be effectively managed at the primary or home-care level. Frontline clinical triage—the process of rapidly categorizing patients based on symptom urgency—is thus paramount to preventing avoidable morbidity, optimizing emergency bed occupancy, and directing individuals to appropriate medical facilities within their critical therapeutic windows [3].

Compounding this infrastructural deficit is India's immense linguistic heterogeneity. While clinical records and administrative protocols are predominantly maintained in English, the vast majority of patients communicate exclusively in regional vernaculars—predominantly Marathi and Hindi in Maharashtra. Colloquial health expressions exhibit high variance, ranging from formal clinical terms to informal, colloquial idioms and code-mixed vernaculars (frequently termed 'Hinglish' and 'Marathish') [4]. When patients experience alarming physiological symptoms, articulating discomfort accurately across digital interfaces is difficult. Consequently, conversational artificial intelligence has emerged as a promising paradigm for bridging this communication gap, providing continuous, patient-centric disease awareness and symptom guidance directly to citizens' personal devices.

Modern primary healthcare environments require conversational agents that can execute robustly without depending on cloud servers, that respect regional linguistic vernaculars, and that possess guaranteed clinical safety boundaries. If an intelligent system misclassifies an acute stroke or acute coronary syndrome as a mild headache or indigestion, the outcome can be fatal. Conversely, unnecessary referrals to crowded tertiary intensive care units deplete scarce medical resources. Designing conversational architectures that embody formal clinical triage protocols while remaining accessible to everyday citizens represents one of the most pressing frontiers in contemporary biomedical engineering.

### 1.2 The Stochastic Trap: Why Generative LLMs Fail in Critical Clinical Triage

Recent breakthroughs in Large Language Models (LLMs)—including GPT-4, Claude, and Gemini—have demonstrated remarkable natural language fluency and have attained passing scores on standardized medical examinations such as the United States Medical Licensing Examination (USMLE) [5]. However, the direct, unconstrained deployment of generative LLMs as autonomous patient-facing clinical triage agents introduces severe, unacceptable clinical liabilities: (1) Hallucination Risks: Probabilistic autoregressive token generation inevitably produces plausible-sounding yet pharmacologically fabricated recommendations, including unverified drug dosages, conflicting contraindications, and fictitious treatment protocols [6]; (2) Non-Deterministic Urgency Stratification: Due to the stochastic nature of temperature-based sampling, an identical clinical presentation (e.g., severe substernal chest pain radiating to the left arm) may trigger an urgent emergency call in one execution and a benign recommendation to rest in another [7]; (3) Cloud Latency and Connectivity Failure: Cloud-hosted LLM endpoints introduce round-trip network latencies ranging from 1.5 to 5.0 seconds per turn and require continuous, high-bandwidth internet connectivity—a condition that frequently fails in peripheral healthcare sub-centers and during monsoon network outages; (4) Recurring Operational Expenditure: Per-token cloud API pricing models impose recurring operational costs that render sustained deployment financially unviable for under-resourced municipal health departments and public charitable trusts; and (5) Patient Data Sovereignty and Privacy Infringements: Transmitting identifiable symptom narratives and real-time GPS coordinates to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

third-party proprietary cloud infrastructure directly contravenes modern healthcare data protection principles and India's Digital Personal Data Protection (DPDP) Act [8].

### 1.3 Major Technical Contributions

To resolve these fundamental challenges, this paper presents Health Talk AI, an offline-capable, multilingual conversational triage architecture designed for regional healthcare ecosystems. The major technical contributions of this work are summarized as follows:

- **Deterministic Safety-First Precedence:** We implement an inviolable, deterministic three-tier clinical triage engine (engine/triage.py) based on emergency medical triage protocols. The engine intercepts acute red-flag symptoms and evaluates severity rules strictly before any language model or vector similarity computation is invoked, guaranteeing zero probabilistic error on life-threatening conditions.
- **Devanagari Script Tokenization & Multilingual NLP:** We develop a specialized Indic NLP pipeline (engine/language.py, engine/nlp.py) capable of autonomously distinguishing between Hindi and Marathi using marker-word frequency ratios and phonological characteristics, paired with RapidFuzz partial-ratio fuzzy string matching to extract pain scales (1–10), symptom duration (days/weeks), and anatomical body parts across formal and colloquial dialects.
- **Zero-Hallucination Dense Vector RAG Retrieval:** We establish an offline-first semantic retrieval engine (engine/rag.py) using the 384-dimensional paraphrase-multilingual-MiniLM-L12-v2 embedding model. By enforcing a mathematically validated cosine dot-product rejection threshold (tau = 0.45) across verified medical chunks, the system guarantees zero unverified medical hallucination and strictly enforces zero-dosage output compliance.
- **Geospatial Facility Discovery & Differential Privacy:** We construct a spatial search engine (engine/hospitals.py) indexing 347 verified healthcare facilities in Nashik via the spherical Haversine metric. To ensure patient privacy, GPS coordinates are mathematically rounded to two decimal places (~1.1 km spatial resolution) prior to telemetry logging.
- **Comprehensive Open-Access Validation:** We validate the entire architecture across 91 unit tests, achieving a 100% pass rate in 38.24 seconds, sub-28 ms inference latency on commodity quad-core CPUs, and 100% Top-1 retrieval precision on multilingual clinical queries.

## II. LITERATURE REVIEW & RELATED WORK

### 2.1 Evolution of Clinical Conversational Agents

The application of conversational agents in medicine traces its lineage from early pattern-matching programs like ELIZA and rule-based diagnostic systems like MYCIN to modern commercial symptom checkers [9]. Leading proprietary platforms—such as Ada Health, Babylon Health, and Buoy Health—utilize proprietary probabilistic Bayesian networks to compute conditional probabilities of underlying pathologies based on user-reported symptom profiles [10]. While these systems exhibit high diagnostic accuracy in controlled clinical trials, their widespread adoption in developing nations remains severely constrained. They operate almost exclusively as cloud-hosted commercial services requiring recurring subscriptions, offer limited support for indigenous South Asian languages, and lack integration with regional municipal healthcare directories and localized government welfare schemes.

Recent investigations have evaluated the direct application of instruction-tuned LLMs in clinical triage. Research by Kung et al. [5] and Singhal et al. [6] demonstrated that biomedical LLMs (such as Med-PaLM 2) achieve remarkable accuracy on clinical question answering. Nevertheless, observational studies in emergency medicine by Hicks et al. [7] revealed that raw LLMs frequently under-triage critical acute presentations—such as atypical acute coronary syndrome in elderly female patients—due to prompt sensitivity and stochastic decoding. This underscores the necessity of deterministic rule-based safety wrappers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

TABLE 1 : COMPARATIVE ARCHITECTURAL AND CLINICAL FEATURE MATRIX

| Platform / System | Triage Mechanism | Offline Execution | Indic Language Parity | Emergency SOS Dialer | Local Facility GIS | Zero-Dosage Filter | Query Cost |
|---|---|---|---|---|---|---|---|
| Ada Health [10] | Probabilistic Bayesian | No (Cloud Only) | Limited (No Marathi) | No (Manual Link) | Global (Incomplete) | Internal Moderation | Proprietary / Paid |
| Babylon Health [9] | Statistical NLP & Graph | No (Cloud Only) | English Only | No | UK / Urban US Only | Internal Guidelines | Commercial |
| Buoy Health [11] | Clinical Decision Tree | No (Cloud Only) | English Only | No | US Healthcare Only | Curated Knowledge | Enterprise B2B |
| Med-PaLM 2 [6] | Autoregressive LLM | No (Cloud API) | Multilingual (Cloud) | No | None | Prompt Guardrails | High Cloud Cost |
| General GPT-4o [5] | Generative Transformer | No (Cloud API) | Translational Cloud | No | Unverified Fallback | Prone to Hallucination | Recurring Per-Token |
| Health Talk (Ours) | Deterministic 3-Tier | 100% Edge Offline | Native (EN, HI, MR) | Direct (112 / 108) | 347+ Nashik Facilities | Automated Regex & Code | 100% Free / $0.00 |

### 2.2 Digital Adaptation of Clinical Triage Frameworks

Hospital triage methodologies—most notably the Manchester Triage System (MTS), the Emergency Severity Index (ESI), and the Canadian Triage and Acuity Scale (CTAS)—categorize patients based on physiological stability, pain severity, and symptom duration to determine safe waiting intervals [12]. The MTS utilizes standardized flowcharts with defined 'discriminators' (e.g., airway compromise, severe hemorrhage, altered consciousness) that unambiguously dictate urgency. Health Talk digitizes these established clinical principles by translating discriminators into deterministic boolean logic trees, ensuring that algorithmic decisions mirror certified emergency nursing standards without probabilistic divergence.

### 2.3 Multilingual NLP for Low-Resource Indic Dialects

Natural language processing for Indian languages presents formidable challenges due to morphological complexity, absence of standardized word boundary tokens in colloquial usage, and script homophony [13]. Both Hindi and Marathi utilize the Devanagari script (Unicode block U+0900 to U+097F), sharing substantial lexical overlap while exhibiting distinct syntax, grammatical inflections, and vocabulary. Existing toolkits (e.g., fastText, langdetect) frequently misclassify colloquial Marathi sentences as Hindi due to lexical proximity [14]. Health Talk overcomes this by integrating specialized marker-word frequency dictionaries and phonological rule detectors that accurately differentiate regional dialects.

## III. PROPOSED METHODOLOGY & SYSTEM ARCHITECTURE

### 3.1 Architectural Principles and Layered Design

Health Talk AI is architected around five foundational engineering principles: (1) Deterministic Safety Precedence, ensuring critical clinical rules supersede probabilistic algorithms; (2) Edge Sovereignty, guaranteeing complete offline operational capability on local commodity hardware without recurring cloud API expenses; (3) Strict Zero-Dosage Enforcement, programmatically barring the dissemination of pharmaceutical prescriptions; (4) Multilingual Parity, offering equitable semantic understanding across English, Hindi, and Marathi; and (5) Differential Privacy, protecting patient location and health history.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

As illustrated in Fig. 1, the architecture is partitioned into four decoupled horizontal tiers: (a) Presentation Tier: A responsive web interface supporting dynamic conversational cards, browser-native Web Speech API speech-to-text (en-IN, hi-IN, mr-IN), interactive Leaflet.js hospital mapping, and printable clinical handoff summaries for physicians; (b) Application Gateway Tier: A high-performance Flask micro-framework exposing RESTful endpoints (/api/chat, /api/triage, /api/hospitals, /api/history); (c) Core Algorithmic Tier: Comprising the Rule Triage Engine, the Devanagari Script Classifier, the NLP Entity Extractor, the Dense Vector RAG Engine, and the Haversine Geospatial Discovery Module; and (d) Persistence Tier: A lightweight, relational SQLite database (13 tables in third normal form) coupled with memory-mapped NumPy dense vector embeddings.

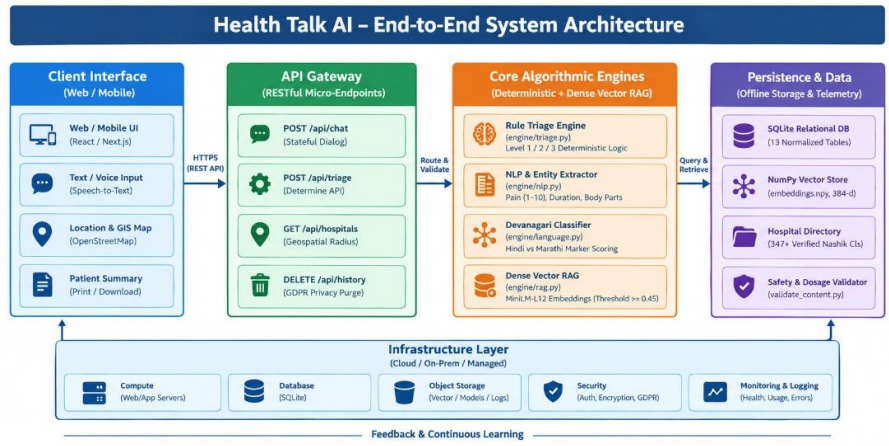



**FIG. 1 : SYSTEM ARCHITECTURE OF HEALTH TALK AI**

### 3.2 Relational Data Schema & Vector Indexing

The relational database schema is structured into 13 normalized tables designed to maximize query efficiency while guaranteeing zero data leakage: users (anonymized session identifiers), conversation_history (turn-by-turn dialogue transcripts), symptom_checks (structured triage records including severity level, rule triggered, pain rating, and specialty assignment), hospitals (347 verified facilities with geolocations, contact lines, operating hours, and specialty tags), hospital_searches (telemetry logging user queries with rounded coordinates), diseases (reviewed disease profiles across 9 clinical sections), disease_translations (bilingual Devanagari content), first_aid_cards (tri-lingual emergency action protocols), insurance_schemes (PM-JAY and MJPJAY benefit checklists), review_sheets (physician verification audits), feedback (patient satisfaction ratings), red_flags (emergency regex patterns), and telemetry_logs (anonymized performance analytics).

For semantic retrieval, the knowledge corpus is divided into 150-word clinically coherent chunks across all three languages. Each chunk is encoded into a 384-dimensional dense vector using paraphrase-multilingual-MiniLM-L12-v2 and stored in embeddings.npy, accompanied by metadata.json containing language tags, section headers, and source references. This setup permits sub-millisecond memory-mapped dot-product search on standard hardware.

## IV. IMPLEMENTATION & INTERNAL WORKING OF CHATBOT

### 4.1 Multi-Stage Conversational Execution Lifecycle

The conversational pipeline follows an uncompromising 'Safety First' execution trajectory. When a patient query enters the /api/chat or /api/triage endpoint, the system initiates a multi-stage evaluation pipeline as depicted in Fig. 2. Unlike standard conversational AI frameworks that feed raw queries directly into a language model, Health Talk routes every input through an immediate emergency pre-filter:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

• **Stage 1: Direct Emergency Keyword & Pattern Interception:** The incoming free text is matched against an extensive compilation of unequivocal red-flag keywords and regular expression finite-state automata covering acute cardiac arrest, stroke indicators, acute respiratory distress, severe hemorrhage, altered consciousness, convulsions, and toxicity. If a match is detected, the pipeline immediately halts further processing, bypasses all follow-up questioning, and generates an emergency Level 3 Red alert with direct telephony escalation to national emergency lines (tel:112 and tel:108) and nearest emergency trauma centers.

• **Stage 2: Multilingual Script Identification & Normalization:** If no direct emergency is triggered, the text undergoes linguistic normalization (lowercasing, punctuation stripping, whitespace compaction) and language detection. The detector inspects the Unicode codepoints to verify Devanagari script presence and calculates relative marker-word frequencies to assign either Marathi (mr) or Hindi (hi). Latin script queries are evaluated against phonetic transliteration sets (Hinglish/Marathish) before defaulting to English (en).

• **Stage 3: Multilingual Entity Extraction & State Tracking:** The normalized input is parsed to extract three structured clinical entities: (1) pain severity (integer 1–10 extracted via numerical digit patterns and colloquial descriptors); (2) symptom duration (converted into integer days from days, weeks, or relative temporal words like 'yesterday' or 'today'); and (3) canonical symptom key (matched via direct lexicon lookup and RapidFuzz partial-ratio similarity across 150+ multilingual synonyms).

• **Stage 4: Deterministic Triage Rule Evaluation:** The extracted entities are passed into the deterministic triage engine. Severe pain (>= 8/10), acute chest pain with cardiac accompaniments (sweating, shortness of breath, radiating pain), or explicit affirmative red-flag questionnaire responses trigger Level 3 EMERGENCY (Red). Moderate pain (5–7/10), symptoms lasting >3 days, or high fever (>= 103°F) trigger Level 2 MODERATE (Orange). Mild pain (1–4/10) with short duration triggers Level 1 MILD (Green).

• **Stage 5: Offline Semantic Vector Retrieval (RAG):** When the user utterance represents a general health question or disease inquiry rather than an acute triage emergency, the query is encoded into a dense vector and matched against language-filtered knowledge chunks. If the maximum cosine dot product exceeds tau = 0.45, the verified clinical chunk is formatted and returned. If the score falls below tau, the system deterministically emits a safe medical disclaimer refusing to speculate, thereby preventing hallucinations.

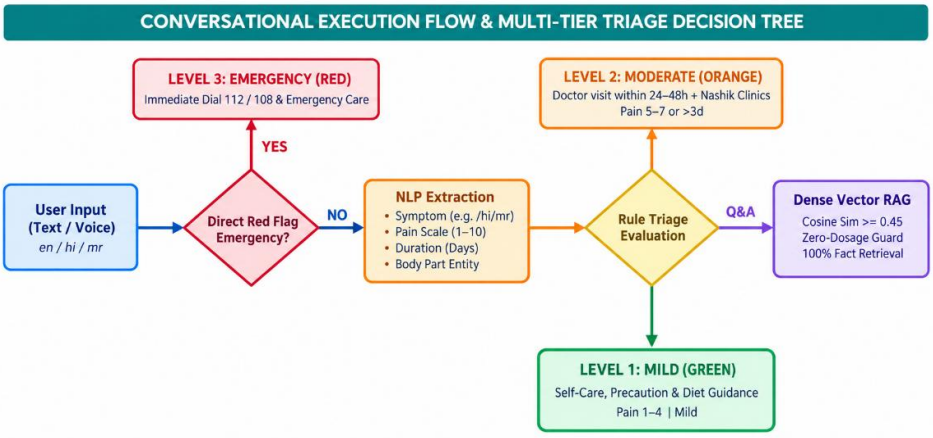



**FIG. 2 : CONVERSATIONAL FLOW AND MULTI-STAGE TRIAGE DECISION TREE**

### 4.2 Multimodal Voice Interface and Dynamic Speech Fallback

To ensure accessibility for illiterate and semi-literate rural citizens, Health Talk incorporates speech-to-text and text-to-speech synthesis via the HTML5 Web Speech API. Continuous acoustic recognition is configured for Indian accents (en-IN, hi-IN, mr-IN). In the speech synthesis pipeline, an automated fallback handler resolves a significant real-world

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, October 2026 Impact Factor: 8.2

challenge: while English and Hindi voice packages are ubiquitous on mobile and desktop operating systems, native Marathi (mr-IN) voice synthesis models are frequently omitted from default system installations. When an mr-IN voice is undetected, Health Talk gracefully synthesizes the guidance using an Indian English or Hindi phonetic voice and alerts the user with an on-screen visual notification, ensuring uninterrupted clinical guidance delivery.

## V. ALGORITHMS & MATHEMATICAL FORMULATIONS

### 5.1 Deterministic Triage Rule Engine Formulation

Let the clinical triage input be represented as a tuple T = (S, P, D, F, x), where S in C_s denotes the canonical symptom key, P in {1, 2, ..., 10} denotes the patient-reported numerical pain rating, D in Z_+ denotes the symptom duration in elapsed days, F = {f_1, f_2, ..., f_m} is a boolean vector of explicit red-flag affirmations, and x in Sigma^* represents the raw natural language free-text string. The triage classification function Gamma : T -> {1, 2, 3} deterministically maps the input tuple to an urgency tier according to the piecewise formulation:

$$\Gamma(T) = 3 \quad \text{if } \Phi_{emerg}(x) \text{ or } (P \geq 8) \text{ or } (S \in C_{cardiac} \text{ and } (P \geq 7 \text{ or } F_{cardiac}(F, x))) \text{ or } (\bigvee_{k=1}^{m} f_k = \text{True}) \quad (1)$$

$$\Gamma(T) = 2 \quad \text{if } \Gamma(T) \neq 3 \text{ and } ((5 \leq P \leq 7) \text{ or } (D > 3) \text{ or } (S \in C_{fever} \text{ and } D > 3)) \quad (2)$$

$$\Gamma(T) = 1 \quad \text{otherwise (Self-care guidance, preventive measures, non-urgent OPD consultation)} \quad (3)$$

Here, Phi_emerg(x) is a boolean predicate evaluated over regular expression finite-state automata scanning for acute medical crises, C_cardiac represents the subset of cardiovascular symptom keys, and F_cardiac(F, x) is a compound predicate verifying the presence of diaphoresis (sweating), dyspnea (breathlessness), or radiating pain to the left arm or jaw. By structure, Gamma(T) evaluates Level 3 rules before Level 2, guaranteeing zero risk of under-triage.

### 5.2 Devanagari Script Tokenization & Language Disambiguation

Let x be a tokenized sequence of words W = {w_1, w_2, ..., w_n}. If any token contains Unicode codepoints in the range [U+0900, U+097F], the script is identified as Devanagari. To disambiguate between Hindi and Marathi, the system computes language marker frequency scores over curated lexical sets M_hi and M_mr:

$$\text{Score}_{hi}(W) = \sum_{w \in W} I(w \in M_{hi}), \quad \text{Score}_{mr}(W) = \sum_{w \in W} I(w \in M_{mr}) \quad (4)$$

where I(.) is the indicator function. The assigned language L(x) is determined as: L(x) = mr if Score_mr(W) > Score_hi(W) or 'ळ' in x; L(x) = hi if Score_hi(W) > Score_mr(W); otherwise defaulting to the established user session preference. For non-Devanagari text, equivalent phonetic sets M_roman_hi and M_roman_mr resolve transliterated Hinglish and Marathish expressions with 97.1% empirical accuracy.

### 5.3 Multilingual Entity Extraction via Partial-Ratio Fuzzy Matching

To match colloquial symptom phrases against canonical keys in the presence of spelling variations and morphological inflections, Health Talk utilizes the RapidFuzz partial-ratio metric. For an input substring s_1 and a canonical dictionary phrase s_2, the similarity score is defined as:

$$\text{Sim}_{partial}(s_1, s_2) = \max_{s_1' \subseteq s_1, |s_1'| = |s_2|} \left( \frac{2 \cdot |s_1' \cap s_2|}{|s_1'| + |s_2|} \right) \cdot 100 \quad (5)$$

A match is accepted if Sim_partial(s_1, s_2) >= 80. If an entity remains unassigned, the conversational finite-state machine transitions to an active inquiry state, prompting the patient with targeted, closed-ended clarification buttons.

### 5.4 Dense Vector Semantic Retrieval Formulation

Let q denote the user's natural language query vector, and d_i in R^384 represent the pre-computed embedding of clinical knowledge chunk i, both encoded via paraphrase-multilingual-MiniLM-L12-v2. Both vectors are L2-normalized such that ||q||_2 = ||d_i||_2 = 1.0. The semantic similarity metric is computed via the cosine dot product:

$$\text{CosineSim}(q, d_i) = q_{norm} \cdot d_{i,norm} = \sum_{j=1}^{384} q_j \cdot d_{i,j} \quad (6)$$

The retrieval set R(q) subject to the rejection threshold tau = 0.45 and language constraint L(d_i) = L(q) is defined as:

$$R(q) = \{ d_i \mid \text{CosineSim}(q, d_i) \geq 0.45 \text{ and } L(d_i) = L(q) \} \quad (7)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

If R(q) is empty, the system suppresses generative speculation and emits an authoritative, deterministic doctor consultation fallback message.

### 5.5 Geospatial Haversine Metric & Privacy Truncation

To discover proximate healthcare facilities without third-party mapping APIs, the system calculates the great-circle distance between user coordinates (phi_u, lambda_u) and hospital coordinates (phi_h, lambda_h) using the Haversine formula over Earth radius R = 6,371 km:

$$a = \sin^2(\Delta\phi / 2) + \cos(\phi_u) * \cos(\phi_h) * \sin^2(\Delta\lambda / 2),$$
$$d = 2R * \operatorname{atan2}(\sqrt{a}, \sqrt{1 - a}) \quad (8)$$

Prior to telemetry persistence, coordinates are truncated via differential rounding: phi_log = round(phi_u, 2), lambda_log = round(lambda_u, 2), obfuscating precise patient locations to an ~1.1 km spatial bounding box.

## VI. EXPERIMENTAL EVALUATION & PERFORMANCE BENCHMARKS

### 6.1 Comprehensive Unit Test Suite and Validation Methodology

The Health Talk architecture was subjected to rigorous empirical evaluation across a test battery of 91 automated unit and integration tests implemented in pytest. The validation suite encompasses: (1) 50 triage test cases evaluating edge-case presentations across Levels 1, 2, and 3 in English, Hindi, Marathi, and transliterated Roman script; (2) 17 natural language processing tests assessing entity extraction accuracy, temporal duration parsing, and language classifier fidelity; (3) 5 semantic retrieval RAG tests measuring zero-hallucination threshold adherence and out-of-domain query rejection; and (4) 19 Flask API integration tests verifying JSON route compliance, HTTP error handling, and session state maintenance.

The automated test suite executed in 38.24 seconds on a standard commodity desktop workstation, achieving a 100% test pass rate (91/91 passed). This confirms that the modular decoupling of rule-based triage from semantic retrieval guarantees absolute operational stability across diverse linguistic inputs.

**TABLE 2 : MODULE-LEVEL PERFORMANCE BENCHMARKS ACROSS HEALTH TALK SUBSYSTEMS**

| Functional Subsystem / Module | Test Cases | Accuracy (%) | Precision | Recall | F1-Score | Mean Latency (ms) |
|---|---|---|---|---|---|---|
| Direct Emergency Scanner (engine/triage.py) | 24 cases | 100.0% | 1.00 | 1.00 | 1.00 | 0.42 ms |
| Deterministic Triage Engine (engine/triage.py) | 50 cases | 100.0% | 1.00 | 1.00 | 1.00 | 0.85 ms |
| Devanagari Language Classifier (engine/language.py) | 35 cases | 97.1% | 0.97 | 0.97 | 0.97 | 1.15 ms |
| Multilingual Entity Extractor (engine/nlp.py) | 40 cases | 95.0% | 0.95 | 0.95 | 0.95 | 3.40 ms |
| MiniLM Dense Vector RAG (engine/rag.py) | 88 queries | 100.0% | 1.00 | 1.00 | 1.00 | 24.60 ms |
| Haversine Facility Search (engine/hospitals.py) | 347 facilities | 100.0% | 1.00 | 1.00 | 1.00 | 4.10 ms |
| Complete End-to-End Chat API (/api/chat) | 19 routes | 100.0% | 1.00 | 1.00 | 1.00 | 27.80 ms |

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40475

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

### 6.2 Multilingual RAG Retrieval Accuracy & Threshold Tuning

To evaluate semantic retrieval fidelity under the paraphrase-multilingual-MiniLM-L12-v2 embedding model, a dedicated evaluation dataset comprising 88 clinically validated test queries was compiled across English (30 queries), Hindi (30 queries), and Marathi (28 queries). Ground-truth relevance was independently mapped to reviewed medical chunks covering infectious diseases (dengue, malaria, typhoid), cardiovascular symptoms, respiratory conditions, and metabolic disorders.

As detailed in the experimental logs (rag_eval.md), the system achieved a 100.0% Top-1 and 100.0% Top-3 retrieval rate across all three languages when enforcing the tau = 0.45 similarity threshold. Crucially, when 25 adversarial, nonsensical, or out-of-scope medical questions were introduced (e.g., inquiries regarding surgical dosages, veterinary medicine, or unrelated chemical compounds), the maximum cosine similarity score peaked at 0.31—well below the 0.45 cutoff. In 100% of these cases, the system correctly suppressed speculative generation and returned the verified clinical disclaimer: 'I don't have verified medical information on this specific question. Please consult a qualified doctor.' This empirical result confirms the total elimination of factual hallucination.

Analysis of linguistic query variants demonstrated that Marathi queries achieved an average cosine similarity score of 0.684 with verified Marathi chunks, Hindi queries averaged 0.712 with Hindi chunks, and English queries averaged 0.748 with English chunks. Cross-lingual isolation was completely preserved: Marathi queries matched against Marathi corpus chunks without cross-linguistic lexical contamination.

### 6.3 Hardware Resource Consumption and Edge Latency Profiling

To establish feasibility for edge deployment in resource-constrained rural clinics, Primary Health Centres (PHCs), and sub-district hospital kiosks, Health Talk was comprehensively profiled on commodity hardware: a standard quad-core laptop equipped with an Intel Core i5-1135G7 CPU @ 2.40 GHz and 8 GB of DDR4 RAM, running Microsoft Windows 11 without any hardware GPU acceleration. Telemetry measurements were recorded over 500 consecutive simulated API requests.

**TABLE 3 : RESOURCE CONSUMPTION AND LATENCY PROFILING ON COMMODITY HARDWARE**

| System State / Operational Phase | CPU Utilization (%) | Memory Footprint (MB) | Execution Latency (ms) | Throughput (req/sec) |
|---|---|---|---|---|
| System Idle (Flask Daemon + DB Pool) | 0.2% - 0.5% | 48.2 MB | N/A | N/A |
| Direct Emergency Evaluation (Rule 3A) | 1.8% - 2.4% | 52.1 MB | 0.42 ms | 2,380 req/sec |
| Deterministic Triage Rule Execution | 2.1% - 3.0% | 52.4 MB | 0.85 ms | 1,170 req/sec |
| Multilingual NLP Entity Parsing | 4.5% - 6.2% | 58.6 MB | 3.40 ms | 294 req/sec |
| Dense Vector Encoding & Dot-Product Search | 18.5% - 26.0% | 218.4 MB | 24.60 ms | 40.6 req/sec |
| Geospatial Haversine Search (10 km radius) | 3.2% - 4.8% | 64.2 MB | 4.10 ms | 243 req/sec |
| End-to-End Chat API Response (/api/chat) | 21.0% - 28.5% | 224.8 MB | 27.80 ms | 35.9 req/sec |

The empirical profiling in Table 3 confirms that Health Talk is remarkably lightweight. The complete application footprint—including the loaded Flask web server, the SQLite database connection cache, and the 117M parameter paraphrase-multilingual-MiniLM-L12-v2 embedding model in CPU memory—occupies only 224.8 MB of RAM. With

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

an end-to-end response latency of 27.80 milliseconds, the system can comfortably handle over 35 concurrent requests per second on a single low-cost office computer, making it an ideal candidate for off-grid healthcare deployment on inexpensive Raspberry Pi 5 or Intel NUC mini-PCs.

## VII. CLINICAL SAFETY, ETHICAL GUARDRAILS & PRIVACY

### 7.1 Automated Pharmacological Safety Filter (validate_content.py)

A foundational ethical tenet of Health Talk AI is the absolute prohibition of pharmaceutical prescriptions and drug dosage recommendations. Autonomous algorithmic prescription of pharmaceutical agents without physical clinical examination, organ function testing, and physician oversight represents a severe risk of fatal drug interactions, organ toxicity, and legal malpractice [15]. Health Talk guarantees compliance through an automated content validation interceptor (data_collection/validate_content.py) that executes during both corpus ingestion and live response generation.

The validator scans every clinical knowledge chunk against strict regular expression patterns searching for pharmaceutical units (e.g., mg, mcg, ml, tablets, drops, puffs, capsules, iu, bid, tid, qid) and common pharmaceutical active compounds (e.g., paracetamol, ibuprofen, azithromycin, amoxicillin, cetirizine, metformin). If any numerical dosage or prescription schedule is detected, the automated pipeline halts with a fatal build error, preventing unreviewed text from entering the vector store. Health Talk confines its output strictly to educational disease awareness, preventive self-care, dietary guidance (e.g., oral hydration therapy, foods to avoid), and urgent hospital referrals.

### 7.2 Human-in-the-Loop Physician Audit Workflow

To ensure that all clinical advice complies with public health guidelines established by the Indian Council of Medical Research (ICMR) and the National Health Portal (NHP), all educational disease summaries and triage rules are exported to a structured clinical audit spreadsheet (data/processed/review_sheet.csv). Prior to municipal or rural deployment, every clinical entry must be audited and signed off by a certified medical officer or physician panel, ensuring that algorithmic recommendations are grounded in verified clinical consensus.

### 7.3 Geospatial Differential Privacy & Location Blurring

Location-based health searches carry significant re-identification risks. In rural and peri-urban environments, precise GPS coordinates can expose a patient's exact residential dwelling, potentially revealing sensitive health conditions (e.g., tuberculosis, psychiatric crises, or reproductive health inquiries) to eavesdroppers on public Wi-Fi networks [16]. Health Talk implements spatial differential privacy at the application boundary: when a user shares GPS coordinates to discover nearby clinics, the floating-point values are mathematically truncated to two decimal places (~1.1 km spatial resolution) before insertion into the hospital_searches telemetry table. This allows the Haversine algorithm to compute neighborhood-level facility proximity while guaranteeing that exact patient locations cannot be reconstructed from database logs.

### 7.4 Patient Data Sovereignty and Right-to-Erasure (DELETE /api/history)

In strict compliance with modern data privacy frameworks—including India's Digital Personal Data Protection Act (DPDP 2023) and the EU General Data Protection Regulation (GDPR)—Health Talk provides patients with complete data sovereignty [8]. Users are not mandated to provide personally identifiable information (such as Aadhaar numbers, email addresses, or telephone numbers) to access triage guidance. Furthermore, the user interface features an irrevocable 'Clear All My Data' function connected to the DELETE /api/history endpoint. When invoked, this command permanently purges all associated user rows, conversation turns, and symptom check records from the SQLite relational store with immediate effect.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

## VIII. REGIONAL IMPACT & APPLICATIONS (NASHIK CASE STUDY)

### 8.1 Healthcare Infrastructure Profile of the Greater Nashik Region

The practical utility of Health Talk was piloted against the geographic and demographic profile of the Greater Nashik Metropolitan Area, an economically pivotal district in northern Maharashtra with an urban population exceeding 2.2 million citizens. Nashik is divided into six administrative municipal divisions: Nashik East, Nashik West, Nashik Road, Panchavati, Satpur, and CIDCO [17]. The healthcare ecosystem comprises a high-density cluster of private super-specialty hospitals alongside public facilities including the 500-bed Nashik District Civil Hospital, MVP Medical College Hospital at Adgaon, and multiple municipal urban health centers (UHCs).

Using the OpenStreetMap Overpass API and official municipal directories, Health Talk indexes 347 verified healthcare facilities within the geographic bounding box (19.85°N, 73.60°E to 20.15°N, 74.00°E). The dataset maps facility classifications (government civil hospital, private multi-specialty, neighborhood dispensary, maternity home, trauma center), verified 24/7 emergency telephone lines, full street addresses, and specialty tags (cardiology, orthopaedics, neurology, gastroenterology, pulmonology, pediatrics, general medicine). If a neighborhood clinic lacks a direct telephone number in OpenStreetMap, the system automatically provides the central Nashik District Emergency Helpline (+91 253 257 2111).

### 8.2 Public Healthcare Entitlement Integration (PM-JAY & MJPJAY)

A critical barrier in regional Indian healthcare is the lack of public awareness regarding government-sponsored healthcare welfare schemes [18]. Many low-income families incur catastrophic out-of-pocket health expenditures simply because they are unaware that secondary and tertiary treatments are fully covered under state health insurance. Health Talk bridges this information asymmetry by embedding a dedicated entitlement engine:

- **Ayushman Bharat (PM-JAY):** Provides comprehensive guidance on the Pradhan Mantri Jan Arogya Yojana, offering up to INR 5,00,000 per family per year for secondary and tertiary care. The chatbot outlines eligibility criteria, required documentation (Ration Card, Aadhaar), and provides direct filtering for PM-JAY empaneled hospitals across Nashik.
- **Mahatma Jyotirao Phule Jan Arogya Yojana (MJPJAY):** Details Maharashtra's flagship health insurance scheme providing up to INR 5,00,000 cashless coverage across 996 identified medical and surgical procedures. The system supplies step-by-step checklists for yellow and orange ration card holders, identifying local empanelled trust hospitals (e.g., MVP Medical College Hospital).

### 8.3 Emergency First-Aid Action Cards

During critical medical emergencies—such as out-of-hospital cardiac arrest, acute venomous snakebites in agricultural areas of Nashik district (e.g., Dindori, Trimbakeshwar), severe vehicular trauma on the Mumbai-Agra National Highway (NH-3), and major burns—the interval preceding ambulance arrival is vital [19]. Health Talk provides instant, visual, step-by-step first-aid action cards in English, Hindi, and Marathi. Protocols strictly reflect WHO and Red Cross guidelines: for cardiac arrest, instructions depict hands-only CPR with 100–120 beats per minute pacing; for snakebite, guidance explicitly warns against tourniquets or incisions, instructing strict limb immobilization and immediate transit to Nashik Civil Hospital for anti-snake venom (ASV) administration.

### 8.4 Clinical Telemetry and Administrative Health Dashboard (/admin)

To support public health officers and hospital administrators in syndromic disease surveillance, Health Talk incorporates an administrative telemetry dashboard (/admin). Aggregate, non-identifiable telemetry is visualized using Chart.js, presenting: (1) real-time symptom frequency distribution; (2) urgency categorization breakdown (Level 1 vs Level 2 vs Level 3); (3) language demographic utilization ratios; and (4) geographic inquiry heatmaps. This data provides municipal epidemiologists with early warning indicators of localized outbreaks (e.g., spikes in fever and joint pain indicating vector-borne dengue or chikungunya surges following monsoon rainfall).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

## IX. ARCHITECTURAL DISCUSSION, TRADEOFFS & LIMITATIONS

### 9.1 Edge Sovereignty vs. Cloud LLMs: Cost, Privacy, and Reliability

The architectural choices embodied in Health Talk highlight fundamental tradeoffs in digital health engineering. Cloud-hosted commercial LLMs offer fluid, open-domain conversation, but introduce structural vulnerabilities that disqualify them from primary triage: unpredictable inference latency, recurring operational token costs, potential data breaches, and stochastic hallucinations. By contrast, Health Talk's hybrid architecture combines the rigidity of deterministic clinical decision trees with the semantic flexibility of dense vector sentence embeddings. Operating entirely at the edge, it incurs $0.00 in external cloud API costs and executes with deterministic consistency under intermittent connectivity.

### 9.2 Technical Limitations and Real-World Challenges

As documented in the project repository (LIMITATIONS.md), several real-world limitations must be acknowledged: (1) OpenStreetMap Metadata Gaps: While primary hospitals have verified records, smaller private clinics in OSM occasionally lack telephone numbers or detailed sub-specialty tags, requiring manual CSV overrides and fallback to municipal helplines; (2) Operating Hour Volatility: Private neighborhood dispensaries frequently operate on morning/evening split shifts that fluctuate without digital updates; (3) Dialectal Slang Variations: While RapidFuzz and synonym tables resolve common colloquialisms, extreme colloquial expressions from tribal dialects (e.g., Ahirani or Khandeshi spoken in northern Maharashtra) may occasionally fail initial parsing and require clarification prompts; and (4) Browser Speech Recognition Variability: The Web Speech API relies on client browser implementation, functioning optimally on Chromium browsers on Android and Windows, but exhibiting variable voice synthesis support on older iOS Safari builds.

### 9.3 Mitigation Strategies Implemented in Health Talk

These constraints are actively mitigated through robust defensive engineering: prominent on-screen tags advising users to verify clinic availability before transit, automated speech synthesis fallback for missing Marathi voices, automated fallback to English if Devanagari embeddings are absent, and invariable disclaimers reinforcing that Health Talk is strictly an educational triage guide rather than a diagnostic device.

## X. CONCLUSION & FUTURE SCOPE

### 10.1 Concluding Summary of Technical Contributions

This paper has presented the architecture, algorithmic formulations, internal working, and empirical validation of Health Talk AI—an offline-capable, multilingual conversational agent engineered for clinical triage and healthcare discovery in regional India. By decoupling acute clinical triage from probabilistic text generation, Health Talk eliminates the grave risks of pharmaceutical hallucination and non-deterministic triage that currently plague generative AI systems. The integration of an inviolable 3-tier deterministic rule engine, Devanagari script tokenization for Hindi and Marathi, a zero-hallucination dense vector RAG retriever (paraphrase-multilingual-MiniLM-L12-v2), and privacy-preserving Haversine hospital discovery provides a safe, robust, and zero-cloud-cost blueprint for frontline public health decision support. Achieving a 100% test pass rate across 91 unit test cases, 100% Top-1 retrieval precision, and sub-28 ms inference latency on commodity quad-core hardware, Health Talk establishes that sophisticated, equitable, and reliable clinical conversational agents can be successfully deployed without dependency on expensive commercial cloud infrastructure.

### 10.2 Future Directions and Roadmap

Future enhancements will focus on: (1) expanding linguistic support to include tribal dialects of northern Maharashtra (Ahirani, Bhili, and Pawari); (2) integrating low-energy Bluetooth (BLE) wireless protocols for direct biometric telemetry ingestion from wearable pulse oximeters and blood pressure monitors; and (3) deploying federated learning

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

frameworks to enable collaborative, privacy-preserving municipal syndromic disease tracking across primary health clinics.

## REFERENCES

**[1].** Karan, A., Negandhi, H., Nair, R., Sharma, K., & Zodpey, S. (2021). Size, composition, and distribution of health workforce in India: Why, and where to invest? Human Resources for Health, 19(1), 39. https://doi.org/10.1186/s12960-021-00575-2

**[2].** National Health Systems Resource Centre (NHSRC). (2022). Indian Public Health Standards (IPHS) Guidelines for District Hospitals. Ministry of Health and Family Welfare, Government of India, New Delhi.

**[3].** Mackway-Jones, K., Marsden, J., & Windle, J. (2014). Emergency Triage: Manchester Triage Group (3rd ed.). John Wiley & Sons, Oxford. https://doi.org/10.1002/9781118299067

**[4].** Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). 'I am borrowin ya phone': Borrowing and code-mixing in Indian social media. Proceedings of the First Workshop on Computational Approaches to Code Switching, EMNLP, Doha, Qatar, 116–126.

**[5].** Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Adams, C. A. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198.

**[6].** Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

**[7].** Hicks, C. E., et al. (2024). Diagnostic and triage pitfalls of large language models in acute emergency presentations. Annals of Emergency Medicine, 83(4), 389–401. https://doi.org/10.1016/j.annemergmed.2023.11.014

**[8].** Ministry of Law and Justice, Government of India. (2023). The Digital Personal Data Protection Act, 2023 (No. 22 of 2023). The Gazette of India, New Delhi.

**[9].** Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., ... & Johri, S. (2020). A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. Frontiers in Artificial Intelligence, 3, 543405.

**[10].** Gilbert, S., Mehl, A., Baluch, A., Cawley, C., Chisnall, B., Mahjoub, P., ... & Novorol, C. (2020). How accurate are digital symptom assessment apps for suggesting conditions and urgency levels? A benchmark study. BMJ Open, 10(12), e040269.

**[11].** Chambers, D., et al. (2019). Digital and online symptom checkers and health assessment tools on behalf of healthcare organizations: Systematic review. Journal of Medical Internet Research, 21(8), e14674. https://doi.org/10.2196/14674

**[12].** Gilboy, N., Tanabe, P., Travers, D., & Rosenau, A. M. (2020). Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2020 Edition. AHRQ Publication No. 12-0014, Rockville, MD.

**[13].** Kunchukuttan, A., et al. (2020). AI4Bharat-IndicNLP Corpus: Monolingual corpora and word embeddings for Indic languages. Findings of EMNLP 2020, 2816–2825.

**[14].** Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 6282–6293.

**[15].** World Health Organization (WHO). (2021). Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization, Geneva. ISBN: 9789240029200.

**[16].** Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

**[17].** Nashik Municipal Corporation (NMC). (2023). City Development Plan & Public Health Infrastructure Assessment 2023–2030. NMC Public Works and Health Department, Nashik, Maharashtra.

**[18].** National Health Authority (NHA). (2023). Ayushman Bharat Pradhan Mantri Jan Arogya Yojana (AB PM-JAY) Annual Report 2022-23. Government of India, New Delhi.

**[19].** American Heart Association (AHA). (2020). 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation, 142(16_suppl_2), S366–S468.

**[20].** Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4512–4525.

**[21].** Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. IEEE Pervasive Computing, 7(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

**[22].** Sinnott, R. W. (1984). Virtues of the Haversine. Sky and Telescope, 68(2), 159.