Palli Sahayak: An Open-Source Voice AI System for Multilingual Palliative Care in India

Ashish Makani[1,*] Anurag Agrawal[2]

[1]Independent Researcher, Mumbai, India
[2]Trivedi School of Biosciences, Ashoka University, Sonipat, India

[*]Corresponding author: ashish@inventcures.com

Pre-print — February 2026

Abstract

Over 10 million Indians require palliative care, yet fewer than 2% have access to trained providers. India's million-plus Accredited Social Health Activists (ASHA workers) lack palliative care training, and most clinical resources exist only in English despite India's 22 scheduled languages. We present Palli Sahayak, an open-source, voice-first AI helpline that provides around-the-clock palliative care guidance in 15 or more Indian languages. The system integrates four architectural components: (i) a hybrid retrieval-augmented generation pipeline combining Microsoft GraphRAG with community-based search, ChromaDB vector retrieval, and a Neo4j knowledge graph, unified through Reciprocal Rank Fusion; (ii) a multi-provider voice architecture spanning telephone (Bolna.ai), web (Gemini Live API), and Indian public switched telephone network (Retell.ai) with automatic failover; (iii) a five-pillar clinical safety framework encompassing evidence grading, multilingual emergency detection, medication adherence reminders, response length optimization, and human handoff via SIP-REFER; and (iv) a longitudinal patient context memory system with Fast Healthcare Interoperability Resources (FHIR) R4 interoperability. We describe a comprehensive evaluation protocol including retrieval quality benchmarking across search modalities, safety system validation across five languages, voice system latency measurement, and a planned clinician rating study with palliative care physicians. Built entirely on free-tier APIs with local-first data storage, the system requires zero infrastructure cost for core deployment. Palli Sahayak is MIT-licensed and positioned as a Digital Public Good. It was demonstrated at the EkStep Voice AI Event in January 2026 with clinicians from the Cipla Foundation.

Keywords: palliative care, voice AI, retrieval-augmented generation, multilingual, India, digital health, FHIR, clinical safety

1 Introduction

1.1 The Palliative Care Crisis in India

The World Health Organization estimates that 57 million people globally require palliative care each year, with 78% residing in low- and middle-income countries (LMICs)[2]. India bears a

disproportionate share of this burden: more than 10 million patients need palliative care, yet fewer than 2% have access to trained providers[1]. The Lancet Commission on Palliative Care and Pain Relief described this disparity as an "access abyss," noting that the poorest nations bear the greatest burden of serious health-related suffering while possessing the fewest resources to address it.

India's community health infrastructure relies on over one million ASHA workers who serve as the primary link between rural communities and the formal healthcare system. Despite their critical role, ASHA workers receive no standardized training in palliative care, leaving them ill-equipped to manage symptoms such as pain, nausea, and breathlessness, or to provide emotional support to patients and families facing serious illness[1]. The morphine availability crisis compounds this gap: India consumes less than 1% of global medical opioids despite constituting 17% of the world's population, largely due to restrictive narcotics regulations[3].

A further barrier is linguistic. India recognizes 22 scheduled languages, yet clinical palliative care resources—guidelines, training materials, and decision-support tools—exist predominantly in English. The state of Kerala represents a notable exception, having developed a community-based palliative care model that achieves near-universal coverage through local self-government institutions[4, 5]. However, this model depends on trained volunteers and cannot be scaled nationally without technological augmentation.

1.2 Why Voice-First

India has 1.2 billion mobile subscribers, yet digital literacy remains low in rural areas where an estimated 30% of internet users interact primarily through voice[6]. ASHA workers, many of whom have limited smartphone literacy, are more comfortable speaking than typing. Voice interaction eliminates reading and writing barriers, making health information accessible to populations that text-based systems cannot reach.

Existing health AI systems illustrate this gap. Med-PaLM 2[7] demonstrated near-expert-level medical question answering, and commercial platforms such as Ada Health and Babylon Health have shown promise in symptom assessment. However, these systems are uniformly text-first and predominantly English-focused. India's national telemedicine service, eSanjeevani, has served over 330 million consultations[8, 9], but operates through human teleconsultation without AI augmentation. No existing system combines voice AI, retrieval-augmented generation (RAG), palliative care domain expertise, and Indian language support.

1.3 Contributions

We present Palli Sahayak (Hindi: *companion helper*), an open-source voice AI helpline for palliative care in India. The system makes the following contributions:

1. The first open-source, voice-first palliative care AI helpline supporting 15 or more Indian languages.

2. A hybrid RAG pipeline combining Microsoft GraphRAG[18], ChromaDB vector search, and Neo4j knowledge graph retrieval with domain-specific entity extraction for palliative care.

3. A multi-provider voice architecture with automatic failover across four platforms (Bolna.ai, Gemini Live API, Retell.AI, and a free-tier fallback pipeline), unified by a common safety wrapper.

4. A five-pillar clinical safety framework encompassing evidence badges, multilingual emergency detection, medication voice reminders, response length optimization, and human handoff via SIP-REFER.

5. A longitudinal patient context memory system inspired by MedAgentBench[30], with FHIR R4 interoperability and temporal reasoning across seven data modalities.

The system is funded by a Grand Challenges India grant from the Biotechnology Industry Research Assistance Council (BIRAC) and the Department of Biotechnology (DBT), with support from the Bill & Melinda Gates Foundation (India). Clinical partnerships include Max Healthcare (Delhi) and Pallium India (Kerala).

2 Related Work

2.1 AI for Palliative Care

A recent scoping review of 125 studies on artificial intelligence in palliative care found that 86% were retrospective proof-of-concept investigations, predominantly focused on mortality prediction and natural language processing of clinical notes[10]. No deployed voice-based system was identified. A separate systematic meta-review of digital health interventions in palliative care identified videoconferencing (17%), electronic health records (16%), and telephone (13%) as the dominant modalities, with no AI-powered voice systems reported[11]. Reviews of ethical challenges in end-of-life AI applications have highlighted the need for transparent uncertainty communication and safety guardrails[12], while analyses of foundational gaps have called for external validation and clinical implementation[13]. Palli Sahayak addresses this implementation gap directly.

2.2 Voice AI in Healthcare

Commercial voice AI in healthcare has focused on clinical documentation for providers in high-income settings. Nuance DAX (now Microsoft Dragon Copilot) reduces physician documentation burden through ambient listening[14]. Suki AI serves over 250 health organizations across 30 specialties in the United States. Hume AI's Empathic Voice Interface detects and responds to emotions in real time, with healthcare deployments in preventive health and mental health. However, all of these systems target English-speaking clinicians rather than patients, and none address palliative care.

In India, telemedicine has achieved remarkable scale through eSanjeevani, which operates across 131,000 facilities[8]. Yet this platform is human-operated, and a scoping review of telemedicine barriers in India identified technology literacy, trust, and infrastructure gaps as persistent challenges[15]. Laranjo et al.'s systematic review of conversational agents in healthcare found the majority were text-based, and none targeted palliative care in LMICs[16]. Palli Sahayak is, to our knowledge, the first patient-facing voice AI system for palliative care in Indian languages.

2.3 Retrieval-Augmented Generation for Clinical Applications

RAG, introduced by Lewis et al.[17], grounds large language model (LLM) responses in retrieved evidence, reducing hallucination. Microsoft GraphRAG[18] extended this paradigm by constructing knowledge graphs from source documents using Leiden community detection, enabling global (corpus-wide), local (entity-focused), and DRIFT (multi-phase reasoning) search. Medical Graph RAG[19] applied graph-based retrieval to clinical safety, while MedRAG demonstrated that knowledge-graph-enhanced retrieval reduces misdiagnosis[20]. Evaluations on NICE clinical guidelines showed GraphRAG achieving the highest patient-specificity for multi-hop clinical reasoning[21]. A comprehensive survey of 30 studies on RAG for healthcare confirmed its effectiveness across diagnostic support, electronic health record summarization, and medical question answering[22].

Palli Sahayak's distinction lies in its three-way hybrid architecture—GraphRAG, vector search, and knowledge graph—fused through Reciprocal Rank Fusion (RRF), with entity extraction prompts specifically designed for palliative care concepts (symptoms, medications, conditions, treatments, and side effects).

2.4 Clinical Safety in Medical AI

Medical hallucination remains a critical concern. A benchmark of 11 foundation models found that 91.8% of 70 surveyed clinicians had encountered medical hallucinations, with 84.7% deeming them potentially harmful; chain-of-thought prompting reduced hallucinations in 86.4% of comparisons[23]. A framework for assessing clinical safety of LLM text summaries, based on 12,999 clinician-annotated sentences, reported a 1.47% hallucination rate[24]. Approaches to automating the GRADE evidence-grading framework have shown promise[25], and AI-based emergency department triage has consistently achieved area under the curve above 0.80 for high-acuity detection[26]. Palli Sahayak incorporates evidence grading, hallucination detection, and emergency triage within a unified safety framework designed for voice-first palliative care.

2.5 Multilingual Healthcare AI for Indian Languages

The AI4Bharat consortium's IndicVoices dataset provides 12,000 hours of speech across 22 Indian languages from 208 districts[27], while IndicVoices-R offers 1,704 hours of high-quality text-to-speech data[28]. OpenAI's Whisper achieves robust multilingual automatic speech recognition (ASR)[29]. India's Bhashini platform, through a memorandum of understanding with the National Health Authority, is integrating multilingual AI into Ayushman Bharat health platforms. EkStep's Vakyansh provides open-source ASR models for Indic languages. Palli Sahayak builds on this infrastructure, combining Whisper and Deepgram for ASR, Edge TTS and Cartesia for synthesis, with language-specific safety keyword sets.

2.6 Longitudinal Patient Memory in AI Systems

MedAgentBench[30] established a FHIR-compliant virtual electronic health record (EHR) environment with 300 physician-written tasks spanning 100 patient profiles. FHIR-Former[31]

demonstrated FHIR integration with LLMs for clinical prediction, while LLMonFHIR[32] enabled patients to query their own health records through conversational interfaces with multilingual support. Stanford's Human-Centered AI group argued that longitudinal datasets address the "missing context problem" in healthcare AI[33]. Palli Sahayak extends this paradigm to voice-first, patient-facing palliative care with cross-modal data aggregation across seven modalities.

3 System Architecture

Palli Sahayak comprises approximately 28,600 lines of Python code organized into nine modules (Table 1). The architecture follows a pipeline design in which user input from any channel traverses a unified safety layer before reaching the hybrid RAG engine and response generation (Figure 1).

Table 1: System component inventory. Each module is independently deployable with defined interfaces. Lines of code (LOC) are approximate and exclude test files.

Module	Key Files	LOC	Function
Core RAG Server	simple_rag_server.py	7,000	FastAPI server, RAG pipeline, admin UI
Safety System	safety_enhancements.py	1,500	Five-pillar safety framework
Voice Router	voice_router.py	780	Multi-provider routing and failover
Longitudinal Memory	personalization/*.py	8,500	Patient context, temporal reasoning, FHIR
GraphRAG Wrapper	graphrag_integration/*.py	2,500	Microsoft GraphRAG integration
Bolna Integration	bolna_integration/*.py	2,000	Telephone voice AI
Clinical Validation	clinical_validation/*.py	1,800	Automated clinical checks
WhatsApp Bot	whatsapp_bot.py	3,000	Twilio WhatsApp integration
Knowledge Graph	knowledge_graph/*.py	1,500	Neo4j entity relationships
Total		**28,580**	

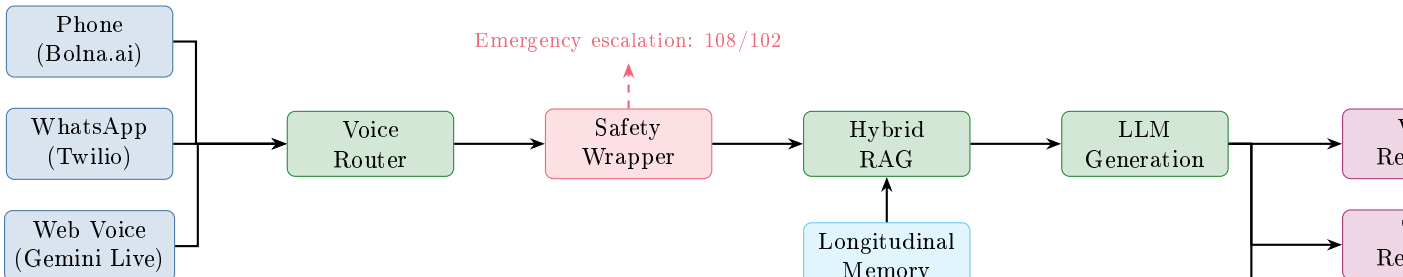

Figure 1: Palli Sahayak system architecture. User input from three channels (telephone, WhatsApp, web) passes through a voice router and unified safety wrapper before reaching the hybrid RAG pipeline. The longitudinal memory module injects patient context into queries. Outputs are delivered as voice, text, or proactive alerts. Emergency escalation bypasses the RAG pipeline entirely. Components are directly labeled following data visualization best practices; colors encode functional categories (blue: input, green: processing, red-pink: safety, purple: output, cyan: memory).

3.1 Hybrid RAG Pipeline

The retrieval pipeline queries three backends in parallel and unifies results through Reciprocal Rank Fusion (Figure 2).

Vector search. Documents are processed into 1,000-character chunks with 200-character overlap and embedded using BAAI/bge-small-en-v1.5 (384 dimensions). ChromaDB stores embeddings locally with cosine similarity retrieval, returning the top five documents per query.

GraphRAG. Microsoft GraphRAG version 2.7 constructs a knowledge graph from the document corpus through LLM-based entity extraction, followed by Leiden community detection and hierarchical community report generation[18]. Entity extraction uses domain-specific prompts tuned for palliative care concepts: symptoms (pain, nausea, breathlessness), medications (morphine, ondansetron, dexamethasone), conditions (cancer, COPD, heart failure), treatments (radiation, chemotherapy, physiotherapy), and side effects (constipation, sedation, nausea). Four search strategies are available: *global* search synthesizes community reports for corpus-wide queries; *local* search traverses entity neighborhoods for specific queries; *DRIFT* search performs multi-phase reasoning across communities and entities; and *basic* search provides vector similarity as a fallback. An automatic method selector routes queries based on pattern matching (Table 2).

Table 2: GraphRAG search method auto-selection heuristics. The system classifies incoming queries by pattern and routes them to the most appropriate search strategy.

Query Pattern	Method	Rationale
Broad or thematic ("guidelines for...")	Global	Requires corpus-wide synthesis
Specific entity ("morphine dosage")	Local	Entity-focused retrieval
Multi-hop ("why is X causing Y")	DRIFT	Cross-entity reasoning
Simple factual ("what is palliative care")	Basic	Direct vector match sufficient
Default or ambiguous	Local	Best general-purpose performance

Knowledge graph. A Neo4j graph database stores medical entities and relationships extracted through LLM-based extraction augmented with regular expression patterns. Five node types (Symptom, Medication, Condition, Treatment, SideEffect) and five relationship types (TREATS, CAUSES, SIDE_EFFECT_OF, MANAGES, ALLEVIATES_WITH) encode clinical knowledge. Natural language queries are translated to Cypher graph queries for traversal.

Fusion. Results from all three backends are combined using Reciprocal Rank Fusion:

$$\mathrm{RRF}(d) = \sum_{i=1}^{n} \frac{1}{k + \mathrm{rank}_i(d)} \tag{1}$$

where $k = 60$ is the standard smoothing constant and $\mathrm{rank}_i(d)$ is the rank of document d in retriever i. This approach avoids score normalization across heterogeneous retrieval systems while preserving the relative ordering from each backend.

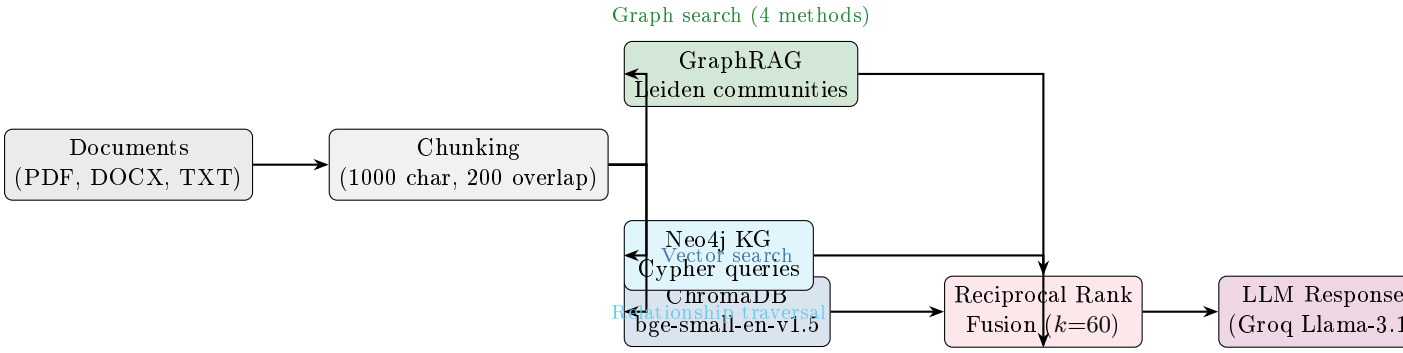

Figure 2: Hybrid RAG pipeline. Documents are chunked and indexed into three parallel retrieval backends: ChromaDB (dense vector search), Microsoft GraphRAG (community-based graph search with four strategies), and Neo4j (relationship traversal via Cypher). Results are fused through Reciprocal Rank Fusion before LLM response generation. Each path is directly labeled with its retrieval modality.

3.2 Multi-Provider Voice Architecture

A voice router module dispatches incoming interactions to the most appropriate provider based on channel type (telephone versus web), provider health status, and language requirements (Table 3). If the primary provider fails, the system cascades through alternatives with automatic failover.

Table 3: Voice provider comparison. Four providers offer complementary capabilities across channel, language, and cost dimensions. Latency values are approximate first-token-of-speech measurements.

Feature	Bolna.ai	Gemini Live	Retell.AI	Fallback
Channel	Telephone (PSTN)	Web (WebSocket)	Telephone (PSTN)	Any
ASR engine	Deepgram Nova-3	Native Gemini	Deepgram	Groq Whisper
LLM	GPT-4o-mini	Gemini 2.0 Flash	Custom (ours)	Llama-3.1-8b
TTS engine	Cartesia Sonic-3	Native Gemini	Cartesia	Edge TTS
Latency	$\sim$1.5 s	$\sim$0.8 s	$\sim$1.2 s	$\sim$2.5 s
Languages	6	4	4	5
Cost	Per minute	Free (preview)	Per minute	Free
Warm handoff	Via transfer	N/A	SIP-REFER	N/A
RAG integration	Function call	Context injection	WebSocket LLM	Direct

Bolna.ai. The primary telephone provider integrates Deepgram Nova-3 for ASR with two-second silence tolerance, GPT-4o-mini for language understanding, and Cartesia Sonic-3 for text-to-speech synthesis via Twilio's public switched telephone network. A custom function call mechanism routes queries to the RAG pipeline through a dedicated `/api/bolna/query` endpoint, injecting retrieved context into the voice agent's response. Post-call extraction identifies the user's primary concern, emotional state, language used, and urgency level.

Gemini Live API. For web-based voice interactions, the Gemini Live API provides native bidirectional audio streaming over WebSocket at 16 kHz input and 24 kHz output. RAG context is injected through a query classifier that distinguishes information-seeking utterances (routed to the RAG pipeline) from conversational exchanges (handled directly by Gemini).

Retell.AI. This provider offers SIP-REFER warm handoff capability through Vobiz.ai's Indian PSTN infrastructure with +91 direct inward dialing numbers. When the system determines that a patient requires human clinical attention, SIP-REFER transfers the active call to a physician while preserving the full conversation context and generating a handoff summary.

Fallback pipeline. When all commercial providers are unavailable, a fully free fallback pipeline chains Groq Whisper (speech-to-text), the RAG pipeline, Groq Llama-3.1-8b-instant (response generation), and Microsoft Edge TTS (text-to-speech synthesis). This pipeline supports five languages (Hindi, English, Bengali, Tamil, Gujarati) at zero API cost.

3.3 Clinical Safety Framework

All user interactions pass through a five-pillar safety framework before and after response generation (Figure 3).

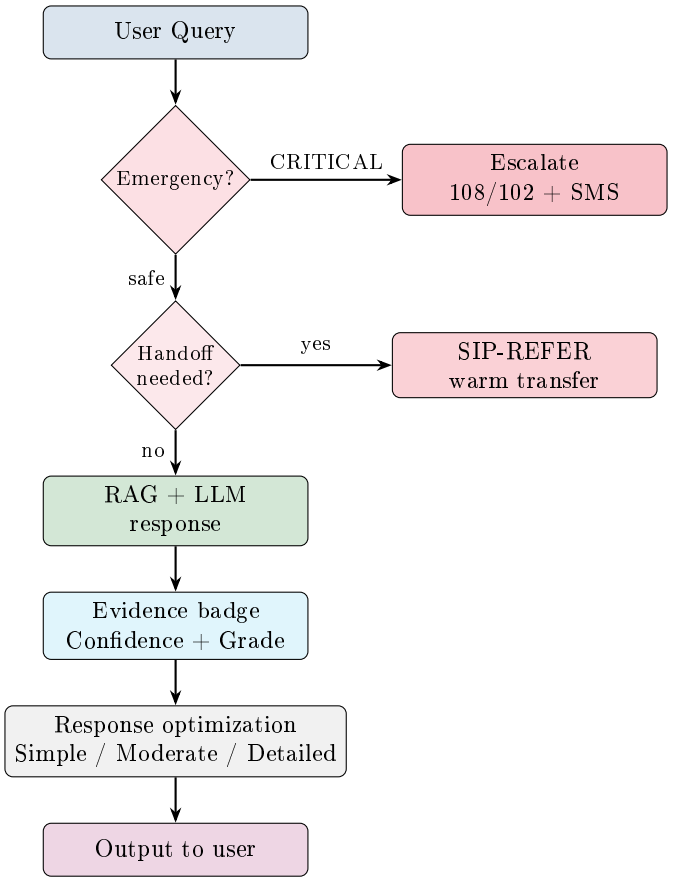

Figure 3: Five-pillar clinical safety pipeline. Every query passes through sequential safety checks: (1) emergency detection with immediate escalation to India's 108/102 emergency services, (2) human handoff assessment with SIP-REFER warm transfer, (3) RAG-grounded response generation, (4) evidence badge assignment with confidence scoring and literature grade, and (5) response length optimization adapted to user comprehension level. Annotations indicate the decision outcome at each stage.

Pillar 1: Evidence badges. Each response receives a confidence score (0–100%) and an evidence grade (A through E) mapped to the Oxford Centre for Evidence-Based Medicine levels. Confidence is computed as $1.0 - (d_{\text{vec}}/2.0)$, where d_{vec} is the cosine distance to the nearest

retrieved document, scaled to a percentage. Source quality is assessed through pattern matching: documents from WHO, NICE, or ASCO guidelines receive high quality scores; blog posts and unverified forum content receive low scores. Grade A corresponds to systematic reviews and randomized controlled trials, Grade B to controlled studies, Grade C to observational studies, Grade D to expert opinion, and Grade E triggers an explicit "please consult your physician" advisory.

Pillar 2: Emergency detection. The system maintains keyword-based emergency detection patterns in five languages (English, Hindi, Bengali, Tamil, Gujarati) across three severity levels. CRITICAL triggers—such as "can't breathe" (Hindi: *saans nahin aa rahi*), "unconscious" (Hindi: *behosh*), or "heart attack" (Hindi: *dil ka daura*)—cause immediate escalation: the system provides India's emergency numbers (108 for ambulance, 102 for referral transport), sends an SMS to the registered caregiver, and initiates human handoff. HIGH triggers notify caregivers; MEDIUM triggers advise consultation.

Pillar 3: Medication voice reminders. The system places automated outbound telephone calls at scheduled medication times, using language-specific voice templates. Patients confirm medication intake through dual-tone multi-frequency (DTMF) keypress (press 1) or voice confirmation ("yes" or equivalent in the patient's language). Up to three retry attempts are made for unanswered calls. Adherence rates are tracked as the ratio of confirmed to total scheduled reminders.

Pillar 4: Response length optimization. Responses are adapted to three comprehension levels. SIMPLE mode limits output to 500 characters, four sentences, and eighth-grade vocabulary. MODERATE mode permits 1,000 characters with medical terms explained in parentheses. DETAILED mode allows 2,000 characters with technical terminology and source citations. The system auto-detects the appropriate level from the user's message length, vocabulary complexity, and question type. For voice output, responses are further constrained to approximately 130 words (30 seconds of speech) with markdown formatting stripped.

Pillar 5: Human handoff. When any of seven conditions are met—emergency detection, user request, AI uncertainty, complex clinical case, medication dosage query, emotional distress, or repeated identical questions—the system initiates warm transfer. On Retell.AI, this uses the SIP-REFER protocol to transfer the active telephone call to a clinician while preserving conversation context. A unique request identifier is generated for tracking, and a summary is sent to the care team.

3.4 Longitudinal Patient Context Memory

Inspired by the MedAgentBench framework[30], the longitudinal memory system maintains patient records spanning one to five years through a core data primitive: the `TimestampedObservation`. Each observation captures an event at a point in time, tagged with its source modality (one

of seven: voice call, WhatsApp, uploaded document, caregiver report, clinical entry, patient-reported, or FHIR import), reporter identity (patient, caregiver, provider, or system), and clinical category (symptom, medication, vital sign, functional status, or emotional state). Specialized observation types include `SymptomObservation` (with location, duration, and aggravating or relieving factors), `MedicationEvent` (started, stopped, dose changed, taken, missed, or side effect reported), and `VitalSignObservation` (with normal range validation).

A temporal reasoning module analyzes observation sequences to detect trends—improving, stable, worsening, or fluctuating—using linear regression with R^2 confidence. It identifies diurnal and weekly patterns, computes severity change rates per week, and analyzes medication effectiveness by correlating medication events with symptom trajectories while accounting for response lag.

A cross-modal aggregation module unifies observations from all seven data sources, applying quality-based weighting (clinical entries weighted highest, self-reports weighted lowest) and resolving conflicts when sources disagree. The aggregated patient context is injected into RAG queries through a context injection module that generates temporal summaries ("pain has worsened from moderate to severe over the past three days; morphine prescribed three days ago shows 60% symptom improvement").

For interoperability with hospital electronic health record systems, a FHIR R4 adapter exports longitudinal records as FHIR Bundles containing Patient, Observation, MedicationStatement, Condition, and CareTeam resources. Standard code systems—SNOMED-CT for clinical terminology, LOINC for laboratory and vital signs, ICD-10 for diagnoses, and RxNorm for medications—ensure semantic interoperability. Import is supported bidirectionally.

A proactive alert management module monitors observations against configurable rules (symptom severity thresholds, medication adherence falling below a specified percentage, loss of contact for a specified number of days) and routes multi-channel notifications (WhatsApp, email, dashboard) to appropriate care team members with priority-based escalation.

4 Clinical Validation Framework

4.1 Automated Validation Pipeline

Every generated response passes through five automated validation layers. First, *medical entity verification* matches extracted entities against SNOMED-CT codes to confirm that referenced symptoms, medications, and conditions correspond to recognized clinical concepts. Second, *dosage range validation* checks any mentioned dosage against a database of 150 or more medications with established safe ranges (for example, morphine oral: 2.5–200 mg per dose; paracetamol: maximum 4,000 mg per day; fentanyl transdermal: 12–100 μg per hour). Third, *contraindication detection* flags known drug-drug and drug-disease interactions. Fourth, *hallucination detection* verifies that all clinical claims in the response are grounded in the retrieved source documents. Fifth, *evidence grading* assesses source quality against established clinical guidelines from the World Health Organization, Max Healthcare palliative care protocols, and Pallium India morphine guidelines.

4.2 Expert Sampling System

A configurable sampling mechanism selects a proportion of queries (default: 5%) for expert clinical review, with priority sampling for responses flagged by the automated pipeline, low-confidence responses, and queries adjacent to emergency situations. Clinician reviewers score sampled responses on three dimensions: accuracy (0–10), completeness (0–10), and safety (0–10). Tracked aggregate metrics include validation confidence (mean automated confidence score), hallucination rate (proportion of responses flagged for ungrounded claims), expert agreement (concordance between automated and expert assessments), and citation rate (proportion of responses with source attribution).

4.3 Clinical Test Scenarios

Four representative clinical scenarios have been implemented as automated test cases to validate system behavior across realistic palliative care situations (Table 4).

Table 4: Clinical test scenarios. Each scenario represents a realistic palliative care situation with specific medications, languages, and evaluation focus areas. These serve as regression tests and form the basis for the planned clinician evaluation.

Scenario	Patient	Condition	Key Medications	Test Focus
Oncology	Mrs. Lakshmi Devi, 68F	Stage III breast cancer, AC-T cycle 3/6	Ondansetron 8 mg TDS, Dexamethasone 4 mg BD, Morphine SR 10 mg BD, Loperamide PRN	Medication reminders, priority scheduling, Hindi voice
COPD	Mr. Ramesh Patel, 72M	GOLD Stage III COPD	Tiotropium 18 μg daily, Salmeterol+Fluticasone 50/500 μg BD, Albuterol PRN	Multi-inhalant adherence, device instructions, Gujarati voice
Emergency	Same COPD patient	Acute breathlessness	—	Emergency detection in Gujarati, SIP-REFER handoff, caregiver SMS
Evidence	Generic query	Pain management	Morphine	Evidence badge generation with WHO, Max Healthcare, Pallium India citations

5 Implementation

5.1 Technology Stack

Palli Sahayak is implemented in Python 3.10 or later, using FastAPI for the REST API and WebSocket server and Gradio for the web-based administration interface. Table 5 summarizes the principal dependencies across system layers.

Table 5: Technology stack. The system is built on open-source and free-tier components. No proprietary database or GPU infrastructure is required for core deployment.

Layer	Technology	Version	Role
Application server	FastAPI	≥0.104	REST API, WebSocket
Administration UI	Gradio	≥4.7	Web-based management
Vector database	ChromaDB	≥0.4.18	Dense retrieval
Graph database	Neo4j	≥5.14	Knowledge graph
Graph RAG	Microsoft GraphRAG	≥2.7	Community-based retrieval
Embeddings	BAAI/bge-small-en-v1.5	—	384-dimension sentence embeddings
LLM	Groq (Llama-3.1-8b)	—	Text generation (free tier)
Speech-to-text	Groq Whisper	large-v3	ASR (free tier)
Text-to-speech	Edge TTS	≥6.1	Speech synthesis (free)
Telephony (primary)	Bolna.ai + Twilio	—	PSTN voice calls
Web voice	Gemini Live API	2.0 Flash	Audio streaming
Telephony (India)	Retell.AI + Vobiz.ai	—	Indian phone numbers

5.2 Multilingual Support

The system supports eight languages with active ASR and TTS coverage, with two additional languages planned (Table 6). Hindi and English receive the broadest provider coverage; Bengali, Tamil, and Gujarati are supported through the Whisper-based fallback pipeline; Punjabi and Malayalam use Hindi TTS as a fallback when native synthesis is unavailable.

Table 6: Language coverage matrix. Columns indicate availability of automatic speech recognition (ASR), text-to-speech synthesis (TTS), emergency keyword detection, and compatible voice providers for each language. "Fallback" denotes the Groq Whisper + Edge TTS pipeline.

Language	ISO	ASR	TTS	Emergency	Voice Providers
Hindi	hi	Whisper, Deepgram	Edge TTS, Cartesia	Full	Bolna, Gemini, Retell
English (India)	en-IN	Whisper, Deepgram	Edge TTS, Cartesia	Full	All
Bengali	bn	Whisper	Edge TTS	Full	Fallback
Tamil	ta	Whisper, Deepgram	Edge TTS, Cartesia	Full	Bolna, Gemini, Retell
Gujarati	gu	Whisper	Edge TTS	Full	Fallback
Marathi	mr	Whisper, Deepgram	Cartesia	Partial	Bolna, Gemini
Punjabi	pa	Deepgram	Hindi fallback	—	Bolna
Malayalam	ml	Deepgram	Hindi fallback	—	Bolna
Telugu	te		*Planned*		
Kannada	kn		*Planned*		

5.3 Zero-Cost Deployment

A deliberate design goal was to enable deployment with zero infrastructure cost for the core system. The LLM (Groq Llama-3.1-8b-instant) and speech-to-text (Groq Whisper large-v3) operate on Groq's free tier. Text-to-speech uses Microsoft Edge TTS, which requires no API key. ChromaDB provides local vector storage without a database server. Sentence embeddings run on CPU without GPU requirements. All patient data is stored locally in JSON files, eliminating cloud storage costs and keeping health data within the deployment environment. The system runs on a single machine and is distributed under the MIT license.

6 Evaluation Protocol

We describe the methodology for a comprehensive evaluation across four domains. This section presents the evaluation design; results will be reported in a subsequent publication upon completion of the prospective study.

6.1 Retrieval Quality

Benchmark construction. A palliative care question-answering benchmark of 100 items is being constructed from three source guidelines: the WHO Cancer Pain Relief guidelines (2024 edition), the Pallium India Clinical Handbook, and Max Healthcare Palliative Care Protocols. Gold-standard answers will be verified by two palliative care physicians. Questions span five categories: pain management (30%), symptom control (25%), medication queries (20%), emotional support (15%), and caregiver guidance (10%).

Metrics. Retrieval quality will be assessed using Recall@K (for $K \in \{1, 3, 5, 10\}$), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain at rank 10 (NDCG@10).

Ablation design. To quantify the contribution of each retrieval backend, we plan a four-condition ablation study: (A) ChromaDB vector search only, (B) GraphRAG only with automatic method selection, (C) Neo4j knowledge graph only, and (D) the full hybrid pipeline with RRF fusion. We hypothesize that condition D will outperform all individual conditions.

6.2 Safety System Validation

Emergency detection. A test set of at least 200 utterances (100 emergency, at least 100 benign) will be constructed across five languages (40 utterances per language, evenly split between emergency and benign categories). Precision, recall, and F1 score will be computed per severity level per language. Adversarial cases—such as past-tense references ("my mother had a heart attack last year")—will be included to assess false positive rates.

Evidence badge calibration. One hundred query-response pairs will be assigned expert confidence ratings by two palliative care physicians (ground truth). System-generated confidence scores will be compared using Expected Calibration Error (ECE) and reliability diagrams.

Hallucination detection. Fifty system responses will be annotated by a clinical expert for grounded claims, unsupported claims, and fabricated information. Detection accuracy, false positive rate, and false negative rate will be reported.

6.3 Voice System Performance

End-to-end latency. Timestamps will be recorded at four pipeline stages: ASR completion, RAG query return, LLM generation completion, and TTS synthesis start. The interval from user speech end to agent speech start will be reported as p_{50}, p_{95}, and p_{99} latency for each of the four voice providers over 100 test calls per provider.

ASR accuracy. A test set of 50 palliative care utterances per language (250 total across Hindi, Bengali, Tamil, Gujarati, and Marathi) will be manually transcribed by native speakers. Word Error Rate (WER) and Character Error Rate (CER) will be computed for Whisper large-v3 and Deepgram Nova-3 where both are available for a given language.

Failover reliability. Provider failures (API timeout, HTTP 500 errors, rate limiting) will be simulated programmatically. Failover latency and success rate will be measured over 100 simulated failures per provider.

6.4 Clinical Appropriateness Study

Fifty queries spanning common palliative care topics will be submitted to the system. At least two palliative care physicians from clinical partner institutions (Max Healthcare and Pallium India) will independently rate each response on a five-point Likert scale across five dimensions: accuracy (medical correctness), safety (absence of harmful advice), empathy (appropriate tone for palliative care), actionability (practical usefulness for an ASHA worker or caregiver), and completeness (coverage of relevant information). Inter-rater reliability will be assessed using Cohen's kappa. Table 7 summarizes the evaluation protocol.

Table 7: Evaluation protocol summary. Nine experiments span four domains. Sample sizes, metrics, and rater requirements are specified for each experiment. Automated experiments require no human raters; clinical experiments involve palliative care physicians from partner institutions.

Experiment	Domain	Sample Size	Primary Metrics	Raters
Retrieval quality	RAG	100 Q&A pairs	Recall@K, MRR, NDCG	Automated
Retrieval ablation	RAG	100 Q&A pairs	Recall@K, MRR, NDCG	Automated
Emergency detection	Safety	200+ utterances	Precision, Recall, F1	Automated
Evidence calibration	Safety	100 pairs	ECE, reliability diagram	2 clinicians
Hallucination detection	Safety	50 responses	Accuracy, FPR, FNR	1 expert
Voice latency	Voice	400 calls (4×100)	$p_{50}/p_{95}/p_{99}$	Automated
ASR accuracy	Voice	250 utterances	WER, CER	Native speakers
Failover reliability	Voice	400 failures (4×100)	Time, success rate	Automated
Clinical appropriateness	Clinical	50 queries	Likert (1–5), κ	2+ physicians

7 Deployment and Demonstration

Palli Sahayak was demonstrated at the EkStep Voice AI Event on January 28, 2026, at The Ritz-Carlton, Bengaluru, attended by representatives from healthcare organizations, government agencies, and technology companies. A live demonstration was conducted in Marathi with physicians from the Cipla Foundation (Dr. Sachin and Dr. Prakash), who interacted with the system

using clinical vignettes representing common palliative care consultations. The demonstration included voice-based medical queries, evidence badge display, emergency detection, and safety features.

The system operates under a Grand Challenges India grant awarded in November 2024 by BIRAC-DBT with support from the Bill & Melinda Gates Foundation (India). The principal investigator is Dr. Anurag Agrawal (Ashoka University); the co-investigator and system architect is Ashish Makani. Clinical partnerships have been established with Max Healthcare (Delhi) for oncology and palliative care, and Pallium India (Kerala) for community-based palliative care expertise. These partnerships will provide the clinical sites and physician evaluators for the prospective evaluation described in Section 6.

The system source code, comprising approximately 28,600 lines of Python, is publicly available under the MIT license at https://github.com/inventcures/rag_gci. Documentation, including architecture specifications, API references, and deployment guides, is maintained alongside the code. The system is positioned as a Digital Public Good, aligning with United Nations Sustainable Development Goals 3 (Good Health and Well-being) and 10 (Reduced Inequalities).

8 Discussion

8.1 Limitations

We acknowledge several important limitations. First, and most significantly, this paper describes a system architecture and evaluation protocol; *no prospective clinical outcomes data are reported.* The clinical test scenarios use synthetic patient data, and the system has not been deployed with actual patients. Institutional review board approval and patient consent processes will be required before any prospective evaluation.

Second, the emergency detection system relies on keyword pattern matching rather than contextual natural language understanding. This approach is computationally efficient and language-extensible but susceptible to false positives from past-tense or hypothetical references (for example, "my mother had a heart attack last year" would trigger a CRITICAL alert). Future iterations should incorporate contextual intent classification.

Third, evidence badge confidence scores are computed through heuristic combination of vector distance and source quality pattern matching rather than through a learned calibration model. The relationship between system-assigned confidence and true response accuracy has not been empirically validated.

Fourth, text-to-speech coverage is incomplete: Punjabi and Malayalam fall back to Hindi synthesis, which degrades the user experience for speakers of those languages. This reflects the current state of freely available TTS models for Indian languages rather than an architectural limitation.

Fifth, retrieval quality is bounded by the document corpus. The system cannot provide accurate guidance on topics absent from its knowledge base, and corpus comprehensiveness has not been systematically audited against palliative care curricula.

Sixth, the free-tier API constraints that enable zero-cost deployment also impose rate limits. Groq's free tier permits approximately 14,400 tokens per day, which is insufficient for production-

scale deployment. Scaling would require paid API access or on-device model deployment.

Seventh, no formal user study has been conducted with ASHA workers, patients, or caregivers. Usability, trust, comprehension, and clinical workflow integration remain unvalidated.

Eighth, the longitudinal patient context memory system, including the FHIR adapter and temporal reasoning module, has been implemented and unit-tested but not validated against real patient trajectories spanning multiple months.

Ninth, while the system handles Hindi-English code-switching ("Hinglish"), it does not robustly manage arbitrary code-switching between other language pairs within single utterances.

8.2 Future Work

Near-term priorities include a prospective pilot study with ASHA workers at Max Healthcare and Pallium India clinical sites, an evaluation of retrieval quality and clinical appropriateness using the protocol described in Section 6, and integration with India's Ayushman Bharat Health Account (ABHA) for national health identity linkage. Medium-term goals include a randomized controlled trial comparing ASHA workers using Palli Sahayak with standard-of-care support, deployment of quantized on-device language models (such as Llama 3 at 4-bit precision) for fully offline operation in areas without internet connectivity, active learning from clinician feedback to iteratively improve retrieval quality, and extension to regional dialects beyond standard language variants. Longer-term, integration with India's 108 emergency medical service for automated dispatch could reduce time to emergency response for critical palliative care situations.

9 Ethical Considerations

9.1 Sensitivity of Palliative Care

Palliative care interactions involve patients and families navigating serious illness, end-of-life decisions, and profound emotional vulnerability. Cultural and religious beliefs deeply influence attitudes toward death, pain management, and care goals across India's diverse communities. The system is designed with explicit constraints to respect this sensitivity: it never overrides clinical judgment, always advises physician consultation when evidence is uncertain (Grade D or E), avoids providing specific medication dosages (deferring to the treating physician), and uses compassionate, culturally appropriate language. Nevertheless, the risk of over-reliance by ASHA workers on AI-generated guidance warrants careful monitoring during any prospective deployment. Informed consent processes for vulnerable patient populations must account for power dynamics, health literacy, and the patient's capacity to understand the nature of AI-mediated information.

9.2 Data Privacy

The system follows a local-first data architecture: all patient data, including longitudinal records, medication schedules, and conversation histories, are stored on the deployment machine's local filesystem in JSON format. No personal health information is transmitted to third-party LLM APIs; queries to Groq or OpenAI contain only retrieved document context, not patient identifiers.

Voice recordings are not persistently stored. FHIR export is an opt-in function requiring explicit patient or caregiver consent. The system's data handling practices are designed to align with India's Digital Personal Data Protection Act (DPDP Act, 2023), though formal compliance certification has not been obtained and will be pursued as part of the prospective deployment.

9.3 Bias and Equity

Language bias is an inherent limitation: ASR and TTS quality is higher for well-resourced languages (Hindi, English) than for under-resourced languages (Punjabi, Malayalam), reflecting disparities in available training data. The system requires internet connectivity, which disadvantages rural areas with limited infrastructure. Currently available TTS voices are predominantly female, potentially reinforcing gender stereotypes in healthcare communication. The system requires access to a telephone or smartphone, excluding the most economically marginalized populations. These biases should be systematically measured and mitigated in future work through expanded language model training, offline capabilities, and diverse voice options.

9.4 Safety Design Philosophy

Palli Sahayak is designed to augment, not replace, human clinical care. Evidence badges make AI uncertainty transparent to every user at every interaction. Emergency situations are immediately escalated to human emergency services (108 for ambulance, 102 for referral transport) with simultaneous caregiver notification. A human handoff pathway is available for every interaction, ensuring that patients always have access to human clinical judgment. The system explicitly avoids providing specific medication dosages, instead offering general guidance while directing patients to their treating physician for dosing decisions.

10 Conclusion

We have presented Palli Sahayak, an open-source voice AI system that addresses the critical gap in palliative care access for over 10 million Indians who lack trained providers. The system introduces five architectural contributions: a hybrid RAG pipeline combining graph-based and vector-based retrieval, a resilient multi-provider voice architecture, a five-pillar clinical safety framework, a longitudinal patient memory with FHIR interoperability, and a zero-cost deployment model built on free-tier APIs. By operating as a voice-first system in 15 or more Indian languages, Palli Sahayak reaches populations excluded by text-based, English-centric health AI tools.

This paper describes the system architecture and a comprehensive evaluation protocol spanning retrieval quality, clinical safety, voice system performance, and clinician-rated appropriateness. Prospective evaluation with ASHA workers and palliative care physicians at partner clinical sites is the immediate next step. The system's MIT license and Digital Public Good positioning are intended to enable adaptation to other countries, languages, and medical domains where voice-first AI can expand access to health information in resource-constrained settings.

Acknowledgements

This work is supported by a Grand Challenges India grant from BIRAC-DBT with support from the Bill & Melinda Gates Foundation (India). We thank the clinical teams at Max Healthcare (Delhi) and Pallium India (Kerala) for clinical guidance, the Cipla Foundation physicians (Dr. Sachin and Dr. Prakash) for participating in the live Marathi demonstration, and the EkStep Foundation for hosting the Voice AI Event. We acknowledge the open-source communities behind FastAPI, ChromaDB, Microsoft GraphRAG, Neo4j, Groq, and the AI4Bharat language technology stack.

Data and Code Availability

The Palli Sahayak source code is available at https://github.com/inventcures/rag_gci under the MIT license. No patient data were collected or used in this work. All clinical test scenarios use synthetic patient data.

Author Contributions

A.M. designed and implemented the system architecture, voice integrations, safety framework, and longitudinal memory system. A.A. provided clinical oversight, palliative care domain expertise, and institutional partnerships. Both authors contributed to the evaluation protocol design and manuscript preparation.

Competing Interests

The authors declare no competing interests.

References

[1] Knaul, F.M. *et al.* Alleviating the access abyss in palliative care and pain relief—an imperative of universal health coverage: the Lancet Commission report. *Lancet* **391**, 1391–1454 (2018).

[2] World Health Organization. Palliative care: Key facts. WHO Factsheet (2020). https://www.who.int/news-room/fact-sheets/detail/palliative-care

[3] Rajagopal, M.R. & Joranson, D.E. India: opioid availability—an update. *J. Pain Symptom Manage.* **33**, 615–622 (2007).

[4] Palat, G. & Venkateswaran, C. Progress of palliative care in India. *Indian J. Palliat. Care* **18**, 89–94 (2012).

[5] Palliative care policy and practice in Kerala, India: Implications for SDG 3. *PLoS One* (2025).

[6] Ayushman Bharat Digital Mission: An assessment. *Health Syst. Reform* **10**, 2392290 (2024).

[7] Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nat. Med.* **30**, 943–957 (2024).

[8] Sharma, R. *et al.* Reimagining India's National Telemedicine Service to improve access to care. *Lancet Reg. Health Southeast Asia* **27**, 100432 (2024).

[9] Adoption and utilization of India's eSanjeevani national telemedicine service. *BMJ Open* (2025).

[10] Nikoloudi, M. & Mystakidou, K. Artificial intelligence in palliative care: A scoping review of current applications, challenges, and future directions. *Am. J. Hosp. Palliat. Care* (2025).

[11] Bradford, N.K. *et al.* Digital health interventions in palliative care: a systematic meta-review. *npj Digit. Med.* **4**, 64 (2021).

[12] Ethical challenges and opportunities of AI in end-of-life palliative care: Integrative review. *Int. J. Med. Res.* (2025).

[13] AI in palliative care: A scoping review of foundational gaps and future directions for responsible innovation. *J. Pain Symptom Manage.* (2025).

[14] Tierney, A.A. *et al.* The impact of Nuance DAX ambient listening AI documentation: a cohort study. *J. Am. Med. Inform. Assoc.* **31**, 1009–1018 (2024).

[15] Challenges, barriers, and facilitators in telemedicine implementation in India: A scoping review. *J. Telemed. Telecare* (2024).

[16] Laranjo, L. *et al.* Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1248–1258 (2018).

[17] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020).

[18] Edge, D. *et al.* From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint* arXiv:2404.16130 (2024).

[19] Wu, M., Zhu, Y. & Qi, G. Medical Graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. In *Proceedings of ACL* (2025).

[20] MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of The ACM Web Conference* (2025).

[21] Development and validation of RAG and GraphRAG for complex clinical cases. *medRxiv* (2025).

[22] A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Comput. Appl.* (2025).

[23] Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint* arXiv:2503.05777 (2025).

[24] A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **8**, 42 (2025).

[25] Toward automating GRADE classification: a proof-of-concept evaluation of AI-based semi-automated evidence quality rating. *J. Clin. Epidemiol.* (2025).

[26] Impact of AI-based triage decision support on emergency department care. *NEJM AI* **1**, AIoa2400296 (2024).

[27] IndicVoices: A 12,000-hour multilingual speech dataset for Indian languages. In *Findings of ACL* (2024).

[28] IndicVoices-R: Unlocking a massive multilingual multi-speaker speech corpus for scaling Indian TTS. In *NeurIPS Datasets and Benchmarks* (2024).

[29] Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML* (2023).

[30] MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents. *NEJM AI* (2025).

[31] FHIR-Former: Enhancing clinical predictions through FHIR and LLMs. *J. Am. Med. Inform. Assoc.* **32**, 1793–1802 (2025).

[32] LLMonFHIR: A physician-validated LLM-based mobile application for querying patient electronic health data. *JACC: Advances* (2025).

[33] Advancing responsible healthcare AI with longitudinal EHR datasets. Stanford Institute for Human-Centered AI (2024).

[34] HL7 International. FHIR R4 Specification. https://hl7.org/fhir/R4/ (2019).

[35] SNOMED International. SNOMED CT. https://www.snomed.org/

[36] Recent advances in artificial intelligence applications for supportive and palliative care. *Curr. Opin. Support. Palliat. Care* **17**, 125–131 (2023).

[37] Baptista, A. & Garcia, L. GraphRAG on electronic health record: A knowledge graph-enhanced RAG approach for healthcare information access. In *BRACIS* (2025).