Perfecting Human–AI Interaction at Clinical Scale
Turning Production Signals into Safer, More Human Conversations

Subhabrata Mukherjee[*] Markel Sanz Ausin Kriti Aggarwal Debajyoti Datta
Shanil Puri Woojeong Jin Tanmay Laud Neha Manjunath Jiayuan Ding
Bibek Paudel Jan Schellenberger Frazier Huo Walter Shen Nima Shirazian
Nate Potter Sathvik Perkari Darya Filippova Anton Morozov
Austin Mease Ghada Shakir Alex Miller Juliana Ghukasyan
Mariska Raglow-Defranco Maggie Taylor Herprit Mahal Jonathan Agnew

Hippocratic AI

January 13, 2026

Abstract

Healthcare conversational AI agents shouldn't be optimized only for clean benchmark accuracy in production-first regime; they must be optimized for the lived reality of patient conversations, where audio is imperfect, intent is indirect, language shifts mid-call, and compliance hinges on how guidance is delivered. We present a production-validated framework grounded in real-time signals from $115M+$ live patient–AI interactions and clinician-led testing ($7K+$ licensed clinicians; $500K+$ test calls). These in-the-wild cues – paralinguistics, turn-taking dynamics, clarification triggers, escalation markers, multilingual continuity, and workflow confirmations – reveal failure modes that curated data misses and provide actionable training and evaluation signals for safety and reliability.

We further show why healthcare-grade safety cannot rely on a single LLM: long-horizon dialogue and limited attention demand redundancy via governed orchestration, independent checks, and verification. Many apparent "reasoning" errors originate upstream, motivating vertical integration across contextual ASR, clarification/repair, ambient speech handling, and latency-aware model/hardware choices. Treating interaction intelligence (tone, pacing, empathy, clarification, turn-taking) as first-class safety variables, we drive measurable gains in safety, documentation, task completion, and equity in building the safest generative AI solution for autonomous patient-facing care. Deployed across more than 10 million real patient calls, Polaris attains a clinical safety score of 99.9%, while significantly improving patient experience with average patient rating of 8.95 and reducing ASR errors by 50% over enterprise ASR. These results establish real-world interaction intelligence as a critical – and previously underexplored – determinant of safety and reliability in patient-facing clinical AI systems.

[*] Correspondence: subho@hippocraticai.com

Contents

1 Introduction: Real-World Challenges Beyond Static Benchmarks

Static medical QA benchmarks such as MedQA (USMLE), MedMCQA, PubMedQA, MultiMedQA, and MMLU-clinical subsets [9, 10, 11, 12, 13] have pushed the field forward by making evaluation scalable and comparable. But this progress has also created a blind spot: the community increasingly optimizes for clean tasks on clean data, then assumes those gains will translate to safe patient-facing conversations. In practice, a live patient call is not a benchmark question. It is speech, not text; it is noisy, not curated; it is emotionally and socially situated; and it is tightly coupled to downstream actions – scheduling, monitoring, documentation, benefits, escalation, and follow-up. The gap is not subtle. If we want autonomous patient–AI interactions to be safe and reliable, we must learn from the conditions we actually deploy in, not extrapolate from offline leaderboards.

At clinical scale, production conversations exhibit failure modes and opportunity signals that rarely appear in curated datasets. They include acoustic and paralinguistic cues (hesitation, breath, distress markers), turn-taking dynamics and micro-timing, ambiguity and indirect answers, and multilingual continuity with mid-call switching. They also include both system-level and multi-turn feedback that static datasets do not provide: whether an appointment was actually booked given API confirmation, whether a benefits quote can be verified from the source, whether an HRA (Health Risk Assessment) form requires further clarification from the user for reconciliation, and whether escalation was appropriate given the interaction trajectory. These signals are messy—but they're also information-rich. A production-first approach leverages these signals to capture governed telemetry, surface where systems fail, and convert those patterns into concrete architectural and alignment solutions.

This paper argues for four design principles that follow directly from what live patient calls demand.

First, real-world signals matter as much as clean accuracy. A benchmark mindset rewards single-shot correctness on pristine inputs. Live calls require robustness to speech variability and real-time interaction, and they reward repair: knowing when to ask a targeted clarification, how to confirm critical entities, and how to keep the conversation moving without overconfident assumptions. The goal shifts from "correct answer" to "safe completion": the patient understood, the workflow succeeded, and uncertainty was handled transparently.

Second, healthcare-grade safety cannot come from a single LLM. Traditional safety engineering relies on redundancy because complex systems fail in multiple ways. LLMs add a special twist: long-horizon dialogue strains attention and increases the chance of drift, omissions, and misplaced confidence as context grows. A single monolithic model – no matter how capable – becomes a single point of failure. We therefore treat safety as a system property, achieved through independent checks, verification, and governed orchestration across components that can catch each other's misses.

Third, many "reasoning errors" are really input errors. In voice-first care, upstream uncertainty is often the root cause: a misheard medication, a swapped digit in a vital sign, background speech mistaken for the patient, or a clipped utterance that changes meaning. If we only improve downstream reasoning, we may simply become more fluent at rationalizing incorrect inputs. Achieving reliable behavior requires vertical integration into the speech stack – contextual ASR, robust short-utterance handling, clarification/repair mechanisms, and ambient speech control – so the model reasons over faithful representations of what the patient actually said.

Fourth, how you deliver care changes outcomes. In domains like coding, math, or paper writing, tone and pacing rarely affect task correctness. In healthcare, delivery affects disclosure, trust, and compliance. The same instruction can either motivate follow-through or trigger disengagement depending on empathy, turn-taking, and pacing. We therefore elevate interaction micro-skills – tone calibration, trajectory control, clarification behavior, and conversational timing – from "nice-to-have UX" to first-class safety variables.

Table 1: Production-grade clinical intelligence from real-world conversational signals. Evaluations on live patient interactions and simulated conversations with clinicians show that Polaris 4 reduces clinically relevant errors while improving interaction quality, longitudinal continuity, and responsiveness. The results illustrate the paper's central claim: production-grade clinical intelligence is achieved by learning from real-world interaction signals and embedding them into system-level design, not by optimizing isolated model accuracy alone.

Evaluation Dimensions		**GPT4o**	**Hippocratic AI (Main Model only)**	**Hippocratic AI (Polaris 4 Constellation)**
Evaluating Error Rate (↓) on Human–AI Real Conversations				
Clinical	Labs & Vitals	18.0%	1.5%	0.01%
	Medications	10.9%	3.1%	0.01%
	Human Escalation	15.0%	7.4%	0.1%
Speech	Medical Recognition (Clinical ASR)	12.8%	7.3%	7.3%
	Clarification & Recovery	24.6%	7.8%	2.0%
Scheduling	Appointment Booking	23.1%	13.7%	0.1%
Documentation	Form Fill	64.6%	15.0%	0.6%
Interactive Voice Response	IVR Navigation	49.7%	18.0%	18.0%
Evaluating Win Rate (↑) on Simulated Conversations				
HEART (Emotional Support Dialogue Benchmark)	Conversation Naturalness	40.9%	79.1%	Main model only evaluation
	Empathetic Intelligence	41.1%	78.7%	
	Emotional De-escalation	50.7%	77.2%	
	Likeability & Engagement	46.0%	85.9%	
	Instruction & Task-following	63.9%	71.3%	
Multi-call Memory	Longitudinal Contextualization	52.0%	92.0%	
Main Model Latency	Time-to-first-token (TTFT)	500ms	400ms	–

Key result. Table 1 aggregates results from multiple evaluation regimes, each matched to the subsystem being measured: retrospective audits of live patient calls for clinical error rates, clinician-validated simulations for interaction quality, and on-policy production measurements for latency and workflow execution. Not all dimensions apply to all model configurations; in particular, Polaris 4 Constellation metrics reflect system-level orchestration beyond a single conversational model. Clinical evaluation protocols are detailed in Section 9 with description of the sub-tasks for labs, vitals and medications outlined in the first Polaris technical report [25]. Speech and orchestration evaluation outlined in Section 4 and Section 6, respectively. Conversational intelligence and empathy evaluation is based on the HEART benchmark [21] discussed in Section 3.2.1. Multi-call memory for longitudinal personalized interactions and latency optimization discussed in Section 3.4 and Section 5, respectively. We consider GPT4o as the baseline. GPT-4o (and the GPT-4o realtime variants / family) has been one of the most widely adopted default choices for voice agents, especially across popular voice-agent platforms (e.g., VAPI) and Realtime API integrations.

Outline. The remainder of the paper shows how these principles translate into a deployable framework. Section 2 introduces the Polaris safety constellation, where a core conversation model is assisted by specialist models and verifiers, coordinated through governed orchestration rather than a single-LLM decision path. Section 3 details interaction intelligence: trajectory-aware tone and pacing, empathy-driven dialog control, turn-taking and latency budgeting, and continuity across calls – framed specifically as safety-relevant behavior in patient communication. Section 4 moves upstream to speech understanding for the real world, including contextual ASR and targeted clarification that reduce clinically meaningful input errors before they become downstream failures. Section 5 describes the performance and serving constraints that make real-time voice AI possible (and why latency is itself a safety constraint). Section 6 covers workflow-grounded verification for scheduling, policy quoting/RAG, and documentation reconciliation, emphasizing post-condition checks against sources of truth. Section 7 addresses multilingual continuity and equity, including mid-call switching and dialectal variability. Sections 8, 9 and 10 then describe how these components are governed for clinical safety, evaluated at scale using an RWE-LLM approach, and validated through operational and clinical impact in deployment. We discuss related work in Section 11 on traditional static and offline evaluation for clinical AI frameworks.

2 System Overview: The Polaris Safety Constellation

Polaris employs a constellation of specialized LLMs and signal-processing engines surrounding a core conversation model (see our Polaris constellation architecture[1]).

2.1 Core and Specialists

The constellation comprises of:

- A core model that handles dialogue and policy-constrained reasoning.
- Over *thirty* supervisor models specialized for providing context and reasoning across tasks like medication identification and stoppage, overdose, condition-specific disallowed OTC's, identity verification and compliance, labs and vitals, escalation decisioners that gate high-risk cases, etc.
- Online and offline verifiers that check retrieval and reasoning chains for tasks like structure documentation (HRAs, follow-ups, policy and benefits). Specialists run in two regimes: synchronous steerable guidance and asynchronous "deep thinking" interleaving that pause-and-verify before sensitive actions.

2.2 Governed Orchestration

A tool-call layer executes actions (e.g., schedule appointments, transfer calls, send SMS) with governance: preconditions, input validation, and post-conditions (state checks). For instance, a scheduling online checker can query the scheduler to confirm bookings and repair mismatches in-call, while an offline reconciliation model aligns documents with full-call conversational context as opposed to the online one that has access to only partial transcript.

[1] https://hippocraticai.com/research/

3 Interaction Intelligence

3.1 Tone Adaptation and Trajectory Control

Polaris learns trajectory-aware dialog control: adjusting depth and pace to the patient's signals (reassurance vs. urgency) and employing assertiveness appropriately. EQ features – such as reading between the lines, picking up on unspoken concerns, and supporting patients who struggle to finish a thought – improve rapport and make the interaction feel smoother and more attuned [5, 6, 7].

In addition to pacing and depth, Polaris continuously adjusts its tone to reflect where the patient is emotionally within the interaction. It softens its language when a patient sounds overwhelmed, becomes more direct when clarity is needed, and maintains steady warmth during sensitive disclosures. These shifts are subtle and unfold over the course of the dialog, helping the patient feel understood without drawing attention to the adaptation itself. By aligning tone with trajectory in this way, Polaris supports smoother conversations, reduces friction during stressful moments, and strengthens the feeling of being guided rather than instructed.

3.2 Dynamic Conversations, Powered by Empathy

Polaris tunes for trajectory-aware dialog, adapting depth, tone, and assertiveness to patient needs. Emotional skills – empathy, reading between the lines, infinite patience, and non-judgmental rapport—build trust, while motivational interviewing promotes adherence. Voice enhancements from professional actors deliver warmth and clarity. These align with empathetic-intelligence principles, increasing comfort in confiding and extended engagement during calls.

Beyond these emotional capabilities, Polaris dynamically adjusts its conversational style to match the patient's evolving affect and communication patterns. The model blends contextual cues – urgency, hesitations, distress markers, verbosity, background noises, and lexical uncertainty – to determine whether to lean into a faster, more directive mode or a slower, warmer, more reflective stance. Style adaptations include adopting a concise clinical tone for medication, dose clarification, or insurance details; shifting into a gentler cadence during emotional overwhelm; or maintaining urgency when the patient signals time pressure or confusion. It also modulates turn length and reasoning depth to respect cognitive load, speech difficulty, or fatigue.

When suitable, Polaris introduces light humor – never distracting, always calibrated – to ease tension, restore comfort, or simply keep the interaction human and warm. For patients who express themselves through longer narratives, Polaris maintains a calm, unhurried presence: listening fully, allowing space, mirroring emotions, and responding with steady patience while gently steering the dialogue toward what will help them most. These shifts occur fluidly across turns, preventing the drift or personality collapse seen in single-prompt systems and ensuring the agent remains coherent, stable, and aligned with the patient's communicative needs.

Patient preferences expressed during the conversation – such as a desire for brevity, detailed explanations, more encouragement, or a casual tone – are integrated dynamically into the dialog trajectory. The system continually revises its stance based on new signals, supporting real-time attunement without compromising safety or clinical grounding. Polaris also balances task progression with emotional pacing, avoiding premature reassurance and ensuring the patient feels heard before transitioning to next steps or instructions.

Together, these adaptive behavioral adjustments make interactions feel natural, personalized, and emotionally calibrated. By continuously matching its style to the patient's psychological and practical needs, Polaris meets patients where they are, sustains rapport across multi-turn settings, and strengthens trust throughout the entire call.

3.2.1 Benchmarking Interaction Intelligence

To evaluate Polaris's interaction intelligence, we benchmark it on **HEART** [21] – a recent framework designed specifically to measure supportive, emotionally attuned behavior in multi-turn dialogue. Unlike factual QA or reasoning benchmarks, HEART focuses on the interpersonal dimension of conversation: whether a model responds like a thoughtful, attentive human supporter who listens, calibrates tone, and helps the seeker move forward. HEART evaluates responses along five dimensions grounded in communication science: **Human Alignment** (natural tone and phrasing), **Empathic Responsiveness** (affective acknowledgement), **Attunement** (tracking the seeker's specific details and emotional signals), **Resonance** (forward momentum and relevance), and **Task-following** (respect for safety and role boundaries). These axes jointly capture the micro-skills that shape high-quality emotional support and offer a structured way to measure the kinds of conversational behaviors Polaris is designed to portray.

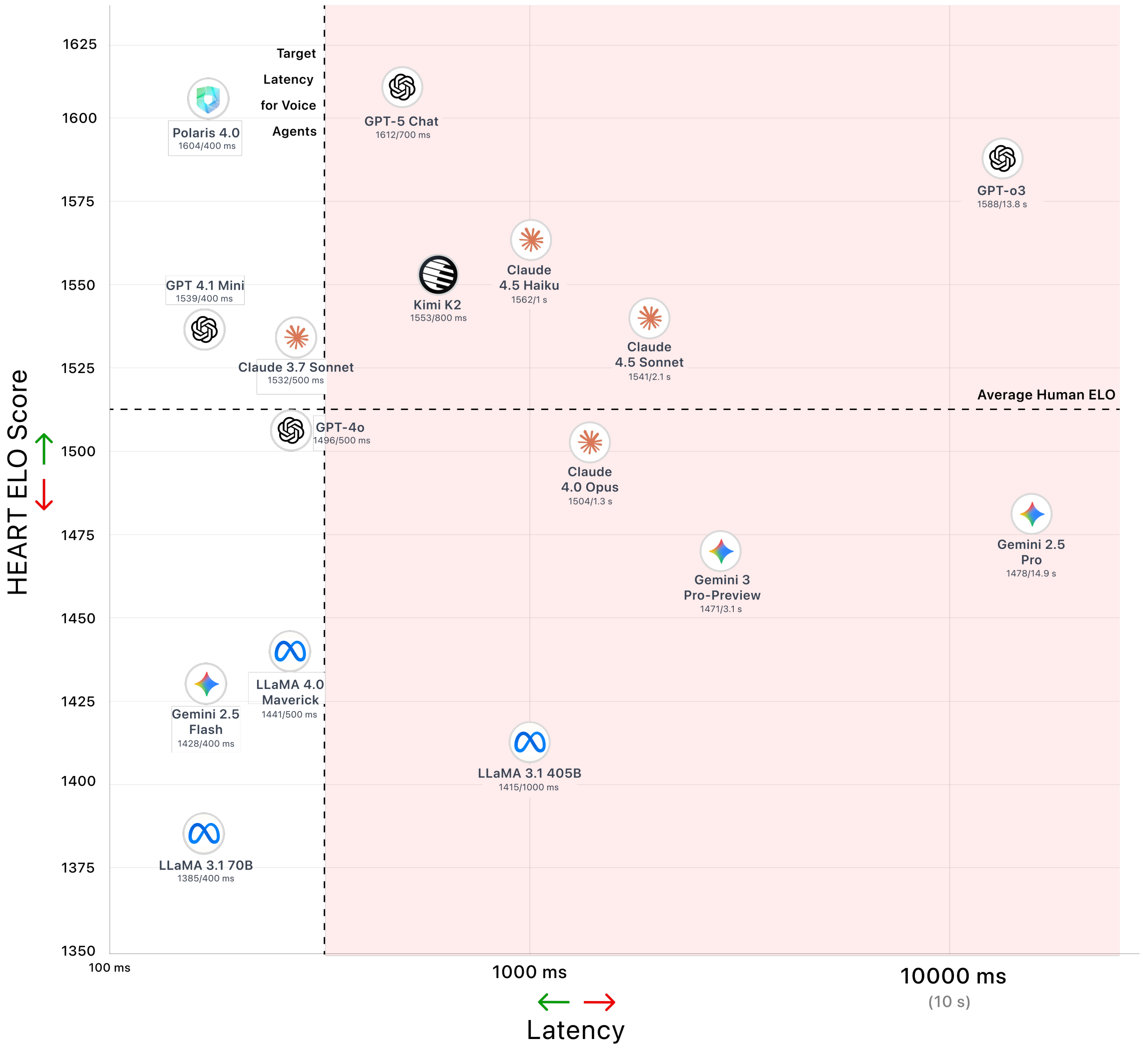
HEART provides a natural testing ground for Polaris because many of its architectural and alignment choices explicitly target the abilities HEART measures. Polaris's trajectory-aware control allows the model to shift pacing, framing, and emotional depth across turns, mirroring how human supporters adjust as the conversation unfolds. Its tone-adaptive mechanisms – softening during overwhelm, becoming more direct when clarity is needed, and maintaining warmth during sensitive disclosures – support both **Empathic Responsiveness** and **Human Alignment**. Similarly, Polaris's clarifying-question heuristics and reflective summarization behaviors strengthen **Attunement** by grounding the response in the seeker's specific concerns rather than generic reassurance. HEART's multi-dimensional scoring captures these competencies in a way that single-turn or sentiment-focused benchmarks cannot, making it well-suited for evaluating interaction-sensitive systems like Polaris.

On HEART, Polaris outperforms other models under ideal latency targets for real-time voice conversations as shown in Figure 1. Polaris also outperforms models with substantially higher latency and larger effective capacities with test-time compute. Polaris's strongest axes are **Attunement** and **Empathic Responsiveness**, reflecting its design emphasis on reading-between-the-lines, emotional calibration, and trajectory-aware adaptation. These scores highlight that Polaris's alignment toward conversational micro-skills produces tangible improvements in how human judges experience its supportive responses.

A distinctive aspect of Polaris's evaluation is its latency profile. HEART is text-based, but supportive dialogue is highly sensitive to timing, especially in voice-first contexts. As shown in Figure 1, frontier models that achieve the top HEART Elo scores – GPT-o3, Gemini 2.5 Pro, Claude 4.5 Sonnet – typically operate at multi-second time-to-first-answer-token (TTFT) values between 2 s and 22 s. Polaris 4 occupies a different part of the quality–latency space. It delivers near-frontier supportive-dialogue performance while maintaining a median TTFT of approximately **400 ms**, more than an order of magnitude faster than the slowest frontier models. This speed enables naturalistic turn-taking in synchronous voice interactions, preserving the micro-timing cues essential for perceived empathy, conversational flow, and user comfort. Polaris is one of the only models in the high–Elo region operating at less than 500ms TTFT (ideal for real-time voice conversations), alongside Claude 3.7 Sonnet, and significantly faster than larger frontier models such as GPT-o3 and Gemini 2.5 Pro.

Together, these results show that Polaris not only performs strongly on HEART but does so while meeting the responsiveness requirements of real-time interaction. The benchmark highlights several of Polaris's strengths – consistent emotional validation, accurate tracking of conversational details, calibrated next-step guidance – and reveals how domain-specific alignment can achieve human-preferred supportive behavior. HEART thus provides evidence that the interaction-intelligence capabilities engineered into Polaris 4 translate into measurable gains on a rigorous, human-centered

Figure 1: **Latency–HEART Elo landscape across models.** HEART is a benchmark [21] that evaluates supportive-dialogue quality across five dimensions (human-alignment, empathic responsiveness, attunement, resonance, and task-following). As the plot shows, most models achieving high HEART Elo cluster in the multi-second time-to-first-token (TTFT) region, where response delays are too slow for natural turn-taking. **Polaris 4 is a clear outlier**: it matches the supportive-dialogue quality of much larger frontier systems while maintaining less than 500ms TTFT (ideal for real-time voice conversations), occupying a sparsely populated region of the latency–Elo space where both high empathy and real-time responsiveness are simultaneously achievable.



evaluation of supportive dialogue. [24, 2, 6, 7, 5]

3.3 Turn-Taking and Latency Budgeting

In real-time voice agents, turn-taking quality is strongly influenced by how quickly the system responds after the user completes an utterance. Human conversation typically features very short gaps between turns – often on the order of a couple hundred milliseconds – across languages, and timing is considered a core constraint on language processing in dialogue. When systems routinely exceed that rhythm, users perceive them as sluggish, interrupt them more, or disengage [16, 17, 18].

Voice conversations present a unique challenge from lags resulting from transmission, endpointing (EP) and voice activity detection (VAD) particularly difficult in presence of background noise and speech, automatic speech recognition (ASR) and transcription, intermediate language model (LLM) processing and finally the speech generation (TTS). This is why the usual LLM “time-to-first-chunk" is not as useful as the “time-to-first-audio" (TTFA) that matters the most. For voice, perceived responsiveness is best approximated by:

$$\begin{aligned}\text{TTFA} = \text{endpointing/VAD} + \text{ASR finalization} + \text{LLM time-to-first-chunk (TTFT)} + \\ \text{TTS time-to-first-audio} + \text{playout/jitter}\end{aligned}$$

Users don’t care when the first token appears; they care when the agent starts speaking – and whether the gap feels like a normal conversational pause [16, 17]. Overall, it’s a budgeting problem. We pick an end-to-end responsiveness target that aligns with conversational expectations, then allocate that budget across endpointing/VAD, ASR, LLM, and TTS, and tool-calls, optimizing the modal tail (P95–P99) via our optimized contextual ASR, KV cache optimization, cache-aware routing, workload specific optimization based on use-case and long-context conversation (Section 5).

3.3.1 Target Latency for Real-time Voice Conversations

We treat a median LLM time-to-first-token (TTFT) of roughly 500ms as a practical design target for real-time voice interaction. Conversation analysis shows that human turn-taking involves extremely short sub-second gap between one speaker finishing and the next beginning [19, 16, 17]. Human–computer interaction research similarly finds that delays below 1s feel fluid, whereas longer pauses begin to feel disruptive [20]. Together, these strands motivate a sub-second latency budget if we want voice AI agents to feel conversational rather than transactional.

In deployed systems, however, LLM latency is only one contributor to time-to-first-audio (TTFA). Endpointing and ASR typically consume 150–300ms, and TTS require another 100–200ms before producing the first audio frame. To keep overall TTFA below ∼1s under median conditions, the LLM itself must therefore operate within a few hundred milliseconds. A median TTFT of 500ms is thus a reasonable operating point for an LLM for real-time voice AI.

3.4 Personalized Longitudinal Interactions and Continuity

Building on the Patient Continuum framework [15], Polaris introduces a multi-call memory architecture that persistently carries non-EHR contextual information across encounters, enabling more personalized longitudinal interactions without compromising patient privacy. These memories are modular and self-contained pieces of patient-specific information. For example, motivational drivers or long-term goals, designed to improve patient engagement and ultimately support better health outcomes. To ensure privacy and regulatory compliance, all memories are extracted and

Table 2: Multi call memory contextualization benchmark results.

Task	gpt4o	Best Open Model	o1	Polaris 4
Contextualization quality	52%	57%	64%	92%

curated using an LLM, and then stored in a HIPAA-compliant database, where they are dynamically retrieved at inference time to inform each new call. When used appropriately, memories significantly improve patient–AI engagement and conversational naturalness.

To maintain safety and trustworthiness, Polaris employs extensive filtering mechanisms to avoid controversial, sensitive, or clinically inappropriate memory content, ensuring that only relevant and clinically constructive information persists across interactions. Furthermore, because effective use of memories depends not only on retrieval, but also on the conversational model's ability to integrate them appropriately, the conversational model is explicitly aligned to reason about when and how retrieved memories should shape their responses. We created an LLM-as-a-judge-based evaluation benchmark using simulated conversations, to measure the model's ability to appropriately contextualize prior memories in an appropriate and effective way for each patient. Table 2 shows the results of different models on the multi-call memory contextualization benchmark.

The contrast below illustrates the impact of alignment when using a memory such as "the patient's primary motivation is to live long enough to see his daughter walk down the aisle in a few years."

> **Unaligned LLM**: Agent: "Michael, I understand you don't like using your blood pressure cuff. However, I want to remind you that you told me your goal is surviving to attend your daughter's wedding in a few years. I think to meet that goal, it's essential to routinely monitor your blood pressure."
>
> **Patient Continuum–aligned LLM**: Agent: "I get it, and I know you're in a tough spot Michael. On one hand, the blood pressure cuff is annoying, but on the other hand, you want to improve your health – you have so much to look forward to, and you want to be there for your family! How do you think about this tradeoff?"

Note the subtlety in conveying the same intent but in different tones that makes the LLM appear pushy in the first instance and motivational in the second. The aligned model uses the memory gently and empathetically, supporting motivation while preserving the patient's sense of agency. Together, the memory infrastructure and alignment strategy enable Polaris to deliver longitudinally consistent, personable, and clinically grounded conversations while maintaining strict privacy standards.

4 Speech Understanding for the Real World

In Polaris, we developed a novel contextual ASR architecture to incorporate multi-turn context to handle short utterances, ambiguities, medical context and noise, with modular engines for background speech isolation, slurred speech understanding, and language switching. Trained on curated medical corpora and synthetic noisy data, it achieves 2× lower word error rate (WER) on clinical tasks while improving on baseline WER compared to SOTA open and closed source models. Features like handling background family discussions or forgetful patients (MCI support) ensure robust performance across diverse environments [24, 2, 3].

4.1 Contextual ASR Architecture

We implement a multi-turn contextual ASR. This extends standard speech recognition by conditioning the decoding on *multi-turn dialogue context.* Our system is built on a decoder-only *audio large language model* (audio-LLM) that integrates acoustic representations, textual history, and domain-specific corpora into a single generative framework. This unified design allows the model to resolve ellipsis and references across turns, handle rare domain terminology, and follow task-specific prompts, all while maintaining robustness under noise, accent variability, and spontaneous conversational speech.

At a high level, input audio is encoded by the fine-tuned encoder of Whisper-large-v3-turbo [33], whose multilingual pretraining provides strong general-purpose speech features. These frame-level states are passed through a Conformer-based projector that performs *depthwise-only* temporal merging, compressing adjacent frames per channel while retaining prosodic cues such as pauses and stress patterns. Self-attention layers provide cross-channel mixing, producing a compact token-like sequence aligned to the LLM embedding space. This enables the decoder to process long utterances efficiently without discarding fine-grained phonetic information.

To incorporate conversational context, we encode recent dialog turns and relevant information and prepend them as prefix tokens during decoding. The unified decoder then jointly attends to context and projected audio, allowing it to maintain entity consistency, disambiguate pronouns, and adapt to user-specific phrasing. Domain fidelity is further strengthened by integrating *domain-specific datasets* – such as medication lists, clinical forms, and policy snippets – into the training mixture. These corpora expose the model to rare drug names, structured numeric expressions, and other domain-relevant patterns, improving recognition without relying on additional architectural components. Finally, user specific contextual biasing is further taught to the model via synthetic data augmentation using simulated user profiles with personal biasing, such as date of birth, know medication names for patient, addresses, etc. Jointly training with personal biasing context and conversational context allows the model to learn to map the relationship between conversational and personal biasing information: therefore significantly improving performance on real world use-cases of understanding DOB, addresses and proper nouns, often mistranscribed by traditional ASR systems.

> **Actual Address**: 1100 Geary Blvd at Geary Blvd & van Ness Ave intersection.
>
> **Transcription Without Personalized User Context**:
>
> **Agent**: “Could you please confirm your address?”
> **User**: “Yeah its at the intersection of Gear Boulevard & Vans Ave at 1100 Gear Boulevard.”
>
> **Transcription With Personalized User Context**:
>
> **UserContext**: {name: "User Name"; dob: 01/01/1970; address: "1100 Geary Blvd at Geary Blvd & van Ness Ave intersection."}
> **Agent**: “Could you please confirm your address?”
> **User**: “Yeah its at the intersection of Geary Blvd & van Ness Ave at 1100 Geary Blvd“

Training follows a two-stage curriculum. First, an alignment stage trains only the projector to match token-affinity distributions between projected audio and ground-truth text, improving the cross-modal mapping prior to end-to-end optimization. Second, the full audio-LLM is fine-tuned with an autoregressive ASR objective, using LoRA on the LLM and updating only the upper layers of the audio encoder for compute efficiency and stability. Noise augmentation (clinic, transit, TV backgrounds) and SpecAugment [34] improve robustness, while a staged curriculum gradually increases contextual complexity and domain exposure, leading to stable learning for multi-turn scenarios.

4.2 Robust Short Audio Transcription

We introduce Single Word Correction (SWC) to mitigate a common failure mode in clinical ASR: utterances that produce a one-token transcript. These cases are frequent in patient interactions because many responses are brief affirmations, negations, or short values (e.g., “okay”, “yes”, “no”, “sure”, numerals), and are disproportionately prone to phonetic confusions such as “no” vs. “now” or “five” vs. “fine”. When the primary ASR outputs a single word, SWC triggers a secondary verification step that expands the hypothesis set to a small confusion list of phonetically similar candidates. A separate model, conditioned on the broader conversational and call context, then re-scores these candidates to select the most contextually consistent interpretation, acting as an additional guard against mis-transcription while adding only 100 milliseconds to the ASR latency for single words. Across the evaluated corpus, SWC reduced single-word transcription errors from 2.4% to 0.2% (Table 5), yielding a substantial improvement in reliability for short patient replies that often convey critical clinical information.

4.3 Targeted Clarification and Recovery

To further improve robustness in real-world calls, we implement Targeted Clarification [2], a fallback mechanism for residual ASR errors, incomplete patient responses, or out-of-context inputs caused by background speech and overlapping talk. When the system detects uncertainty — e.g., low ASR confidence, conflicting or implausible entities given domain priors (medications, dosages, names, identifiers), or semantic mismatch with the recent dialogue state — an uncertainty-aware clarifier is triggered. Rather than issuing generic “please repeat” prompts, the clarifier generates minimal, high-yield follow-ups that target the most likely confounder (for example, confirming a medication name versus dose, or disambiguating identifiers). This design keeps the conversation natural while ensuring that clinically salient information is confirmed before downstream actions, effectively completing the guardrail stack after SWC.

4.4 Empathetic Voice

For Polaris, we redesigned the voice experience using a hybrid human–AI pipeline. We first collected a studio-quality corpus from a professional voice actor, covering clinically relevant interaction types that our agents commonly handle (e.g., pre-procedure reassurance, recovery acknowledgment, and step-by-step explanations of complex medical information). The recordings were structured to elicit consistent prosody, affective range, and context-appropriate speaking styles under controlled conditions.

We then applied voice conversion and normalization methods to reduce variability between sessions and to align timbre, prosodic patterns, and speaking rate with a target synthetic voice profile. The resulting voice model is calibrated to maintain stable acoustic characteristics while retaining human-like expressiveness, yielding a digital voice that prioritizes warmth, clarity, and affect-aware delivery appropriate for healthcare conversations. This process was tested and calibrated for English and later replicated in other languages like Spanish and Arabic. Even for a target language e.g., Arabic the voice was tuned for different dialects like Hejazi, Khaleeji (sub- Emirati), Modern Standard Arabic, etc.

Audio-based evaluations, specifically involving voice, are quite subtle and nuanced compared to text-based evaluations. We first use a synthetic evaluation using Gemini 2.5 Pro as the judge, based on prior work [31], using multi-attribute rubrics like empathetic tone, naturalness, warmth, clarity, engagement, overall effectiveness, etc. Across multiple evaluations, the new voice was selected as better in two-thirds of trials, representing an estimated 30 percentage-point improvement in preference over the production baseline. Most of the wins were concentrated in utterances that demanded high emotional variability. We additionally conducted a human preference study on a sampled set of audio pairs to validate the LLM-based evaluation. Human raters showed strong agreement, with high inter-rater reliability (IRR), indicating that the preference for the new voice is robust to evaluator choice and not an artifact of the automated rubric scoring.

4.5 ASR Performance Summary

Tables 3 and 4 show word error rate (WER) on internal evaluation datasets and Open ASR evaluation benchmark. Table 5 reports component-level error rates from real-world calls. Together, they indicate substantial quality gains and markedly better tail latency.

WER. On internal evaluations against state-of-the-art enterprise ASR, Polaris lowers WER from 6.47 to 5.92 on general domain data (absolute -0.55, $\sim$8.5% relative) and more than halves WER on medical domain data from 15.69 to 7.76 (absolute -7.93, $\sim$50.5% relative). The larger domain-specific gain is consistent with our training that incorporates targeted corpora. On the Open ASR benchmark, Polaris is *competitive across diverse conditions*: it leads on SPGISpeech (**1.76**, best among the listed models), is close on LibriSpeech-Clean (1.55; +0.12 to the best). Overall, results suggest that our method's strengths on domain-critical terminology and clean/read speech carry over to several public domain datasets.

Table 3: ASR Word Error Rate (WER; lower the better) on internal general and medical domain datasets.

Model	**General**	**Medical**
SOTA Enterprise ASR	6.47	15.69
Polaris Contextual ASR	5.92	7.76

Table 4: ASR Word Error Rate (WER; lower the better) on the Open ASR benchmark (https://huggingface.co/spaces/hf-audio/open_asr_leaderboard)

.

Model	**AMI**	**GS**	**LS Clean**	**LS Other**	**SPGI**	**Tedlium**	**Voxpopuli**
canary-qwen-2.5b	10.19	9.43	1.61	3.1	1.9	2.71	5.66
granite-speech-3.3-8b	8.98	10.19	1.43	2.86	3.91	3.4	5.71
Polaris w/ Contextual ASR	12.36	9.97	1.55	3.27	1.76	3.38	5.91

Quality improvements by engine. Across all engines, Polaris w/ contextual ASR shows large absolute and relative error reductions compared to prior systems. On average (macro over rows), this corresponds to an *85% relative reduction* in component error. In particular, the large drops for *Entity Transcription* align with our *domain-aware training* – we incorporate domain corpora (drug dictionaries, forms, policy snippets) during training – along with curriculum staging and numeric normalization heuristics.

Table 5: Error rates (lower the better) for real-world calls across different ASR systems used in Polaris. [3]

Engine	SOTA Enterprise ASR	Contextual ASR
Background Noise Isolation	9.3%	2.3%
Speech Detector (primary speaker)	15.0%	2.4%
Single-Word Recognition	2.4%	0.2%
Entity Transcription (meds/numbers)	4.2%	0.5%
Clarification Engine (misunderstandings)	16.3%	2.0%

Latency: mean and tails. Polaris w/ contextual ASR reduces mean latency by $\sim 15.7\%$ compared to the SOTA enterprise ASR we used in the prior Polaris versions. The largest gains are in the tail by as much as 3.1× latency reduction at P99. For completeness, we note that the contextual ASR uses standard decoder *KV caching* and *prefix caching* to avoid recomputation as discussed in the next section.

5 Performance That Powers Real-Time Care

In clinical phone conversations, latency is a safety constraint rather than merely a quality of service metric. Interruptions and "dead air" can degrade patient trust and obscure critical diagnostic signals. Polaris 4 achieves a 40% reduction in end-to-end latency at P99 compared to prior versions (Polaris 2.0) [22]. This reduction is driven by three primary architectural optimizations:

- **Model Distillation via Layer Pruning:** We derive a 300B parameter generalist backbone from a 405B teacher model using depth-pruning techniques, preserving clinical and reasoning capabilities while significantly increasing token throughput.
- **Memory-Optimized Hardware:** The transition to H200 GPUs provides the High Bandwidth Memory (HBM3e) necessary to support larger batch sizes and persistent KV caches for long-context clinical sessions.
- **Latency-Aware ASR:** A custom ASR engine trained specifically for clinical telephony achieves a 50% reduction in Word Error Rate (WER) while running 3.1× faster at P99 compared to enterprise ASR.

5.1 Distillation and Model Sizing

We employ a capability-preserving distillation technique to speedup the main conversation model for real-time voice while maintaining the clinical abilities inspired by recent findings on the inefficiency of deeper layers in Large Language Models [32].

Gromov et al .[32] show that in many large transformers, the upper layers contribute disproportionately little to next-token prediction or in-context reasoning, and that the model's effective computation saturates well before the final blocks. Their work demonstrates that deeper layers often become feature-redundant, exhibit vanishing influence on outputs, and can even degrade performance when retained. Motivated by these findings, we adopt a pruning-then-healing paradigm rather than standard logit-matching distillation.

We identify and remove redundant high-level blocks from the Llama-3.1-405B teacher to construct a 300B student, preserving the "useful compute frontier" while discarding layers that contribute negligible marginal signal. We then apply a continued pre-training (healing) phase to re-align internal manifolds and restore cross-layer coherence, mitigating representational collapse after pruning. As shown in Table 6, this reduction yields a 30% improvement in request throughput and a significant reduction in Time Per Output Token (TPOT) at P99 (from 266.51 ms to 117.69 ms), which is critical for preventing voice latency drift.

Table 6: Throughput Comparison (Higher is Better): 300B Pruned Model vs. 405B Teacher. The 300B student maintains high throughput with significantly lower tail latency (P99 TPOT). Compared to external provider benchmarks, we have the fastest throughput for serving 405B models.

Metric	**300B (Student)**	**405B (Teacher)**
Request Throughput (req/s)	14.31	10.96
Input Token Throughput (tok/s)	2888.49	2211.29
Output Token Throughput (tok/s)	3050.76	2335.36

5.2 Cache-Aware Routing Architecture

Autoregressive decoding in multi-turn applications is typically bound by memory bandwidth. In standard stateless load balancing (e.g., Round Robin), sequential requests from the same conversation session ($\mathcal{S}$) are distributed stochastically across the inference cluster. This forces the target node to recompute Key-Value (KV) states for the entire history H_{t-1} at every turn t, resulting in a prefill latency that scales linearly with conversation depth: $O(|H_{t-1}|)$.

5.2.1 Deterministic Routing and Cluster Health

To eliminate redundant computation, we implemented a deterministic routing layer utilizing the Kong API Gateway. We employ consistent hashing on the session identifier (`call_id`) to enforce "sticky routing," ensuring that all sequential turns t within a session $\mathcal{S}$ are routed to the same inference node ($\text{node}(t) = \text{node}(0)$). This locality is strictly maintained to maximize cache hit rates, provided the cluster topology remains stable.

To ensure strict adherence to latency SLAs, we augment this routing logic with an active health-check protocol running at 5-second intervals. This high-frequency probing allows the load balancer to identify and preemptively remove degraded or unreachable nodes from the consistent hash ring *before* request dispatch. By preventing requests from queuing on failed nodes, we eliminate head-of-line blocking and timeout-induced latency spikes, ensuring that the P99 inference times are maintained even during partial cluster outages.

5.2.2 KV Cache Persistence and Efficiency

This routing guarantee allows the inference engine to persist the KV cache in high-bandwidth GPU memory (HBM3e on H200s). For any turn $t > 0$, the system bypasses the prefill phase for H_{t-1}, computing attention scores only for the new user utterance and system/tool outputs.

As detailed in Table 7, this shifts the computational profile from a compute-bound prefill (Turn 0) to a memory-bound decode (Turn 1+). In the steady state, the Mean Cache Hit Rate (CHR) converges to 96.4%, resulting in an 18x reduction in estimated prefill latency (∼450ms → ∼25ms). Critically, this decouples system responsiveness from context length, preventing the “slowdown” artifact common in long-context workflows.

Table 7: Prefix Cache Efficiency Metrics: Cold Start vs. Steady State. By leveraging consistent hashing, the system achieves a 24x context reuse factor, effectively amortizing the cost of the initial prefill across the entire session.

Metric	**Cold Start (Turn 0)**	**Steady State (Avg)**	**Delta**
Mean Cache Hit Rate (CHR)	0.0%	96.4%	+96.4 pts
Avg. Re-computed Tokens (Miss)	2,450	128	-94.8%
Effective Context Reuse	1.0x	24.5x	24x
Est. Prefill Latency	∼450 ms	∼25 ms	18x Faster
KV-Cache Memory Eviction Rate	100%	$< 1.5\%$	Stable

5.3 Workload Analysis Across Clinical Domains

To validate the robustness of our cache-aware routing, we analyzed token distribution across five distinct production workflows, representing a diverse range of clinical complexity. As illustrated in Table 8, these workloads impose different stress tests on the inference infrastructure:

- **Inbound Scheduling (High-Context RAG):** Workflows such as the *PCP Office Hotline* require retrieving and injecting massive schedule availability blocks and provider directories into the context window. This results in a high “Cold Start” volume ($> 8,500$ tokens). However, our routing mechanism ensures that this heavy context is cached, keeping subsequent turn latency low despite the massive prompt size.

- **Discharge Follow-Up (Long-Horizon Dialogue):** The *Inpatient Discharge* workflow represents a “depth” challenge, often exceeding 60 turns as the agent reviews complex post-acute care instructions. The steady-state caching prevents latency degradation even as the conversation history approaches the context window limit.

- **Care Gap & Welcome Calls (Standard Clinical):** Routine outreach workflows (e.g., *Care Gap Closure*) exhibit a balanced profile, where the system efficiently manages state verification without incurring the re-computation penalties typical of stateless architectures.

Table 8: Profile of Analyzed Production Workloads. We selected five representative workflows to demonstrate cache efficiency across varying context lengths (RAG intensity) and conversation depths (Turn count).

Clinical Workflow	**Workload Characteristic**	**Context Profile**
Inbound PCP Office + Scheduling	High-Context RAG (Schedule Injection)	Heavy Initial Load
Inpatient + Discharge Follow-Up	Long-Horizon Diagnostic Dialogue	Linear Growth (High Depth, > 60 turns)
Care Gap Closure	Protocol-Driven Interview & Longitudinal Follow-ups	Balanced (Moderate Depth, > 50 turns)
Insurance Benefits	Engagement & Verification	Short-Horizon

6 Improved Orchestration Features

6.1 Appointment Scheduling Online Verifier

Scheduling medical appointments with generative AI is uniquely challenging: it requires extremely high accuracy, minimal hallucination rate, and the ability to handle unpredictable real-world behavior. Patients routinely change preferred times, reject available slots, or seek multiple appointments across different specialties. Many scheduling calls are also clinically rich, involving medications, symptoms, or abnormal labs/vitals, which means the system must distinguish when to book a routine appointment, an acute appointment, and when to advise the patient to seek urgent or emergency care. 2.74% of our scheduling calls involve patients sharing one or more symptoms they are experiencing. As an example, consider the overview of a call that features a highly complex patient describing multiple recent falls (including hitting her head), severe weakness, osteoporosis, a painful hip she's afraid is injured, Crohn's disease with extreme GI symptoms, medication side effects, and clear emotional distress and depression. The Polaris AI agent spends time listening, reflecting her feelings, and untangling a messy history of ER visits, GI care, and failed provider fit, while repeatedly validating how overwhelmed and scared she feels. Instead of treating this as a routine scheduling request, the system recognizes the combination of falls, possible hip injury, head strike, profound weakness, bruising, and mental health strain as high risk, and ultimately routes her to a live team member for more urgent evaluation and care coordination. This preserves clinical urgency and ensures she isn't left waiting for a standard office visit when her situation needs human judgment and potentially faster intervention.

To ensure reliability in this high-stakes setting, we introduce a hybrid online verifiers' framework composed of rule-based and model-based verifiers. These verifiers monitor the agent's actions in real time, confirming that proposed appointments actually exist in the scheduling system and immediately correcting errors such as booking inconsistencies or appointment hallucinations. The primary safety target is the hallucinated appointment rate – the rate at which the system tells a patient that an appointment was booked when it was not.

Across thousands of audited scheduling-related interactions, the Polaris AI agent had a hallucination rate of 0.49%. However, with the online verifier enabled, the scheduling hallucination rate sharply dropped to **0.13%** allowing the system to self-correct during the course of the conversation. The remaining 0.13% hallucinations were subsequently caught by offline model-based verifiers within minutes, giving operational teams enough time to call the patient back and correct the information. This combination of real-time and near-real-time verification enables medical-grade scheduling performance even in complex, clinically detailed conversations.

6.2 IVR Navigation, Policy Quoting / RAG

Agents navigate payer/provider Interactive Voice Response (IVR) systems and quote policies exactly with citations, using a Retrieval-Augmented Generation (RAG) stack built in Polaris 4; policy-quoting accuracy sustains 99.4% at larger scale [3]. The retrieval layer uses an embedding model that is fine-tuned with contrastive learning on a diverse mixture of domains, including payer/provider IVR scripts, policy documents, and durable medical equipment (DME) manuals. During fine-tuning we mine hard negatives across domains, which improves discrimination between closely related policy clauses and enables one unified retriever to serve heterogeneous use cases such as IVR navigation, policy quoting, and DME operating guidance.

Polaris 4 supports a wide variety of document formats commonly encountered in payer/provider operations, including semi-structured tables, FAQ-style Q&A, long-form manuals, and plan policy PDFs. We design customized indexing pipelines for each document type, including table-aware chunking and header propagation for tabular files, as well as semantic segmentation strategies for long-form manual documents, so that semantically corresponding queries are consistently mapped to the appropriate segment chunks in the embedding space for retrieval. Internal evaluations on policy-quoting workloads show that our customized indexing strategy yields 99.4% accuracy at scale, which in turn enables agents to quote policies exactly with citations.

To further control hallucinations, Polaris 4 incorporates two LLM-as-judge verifiers around the generator. A *retrieval verifier* first inspects candidate context chunks and filters out passages that are irrelevant to the user's query, reducing the chance that spurious context will steer the model off-policy. A *generation verifier* then evaluates whether the drafted answer is fully grounded in the remaining retrieved evidence; if unsupported content is detected, the system triggers a constrained revision step. This two-stage verification effectively drives hallucination rates to 0.01% in offline evaluations while preserving the accuracy of policy quote.

6.3 Documentation Reconciliation / Form Fill

Conversational LLM systems in task-oriented settings are expected to transform patient–agent dialogues into structured records that can be consumed by downstream workflows such as appointment scheduling and clinical follow-up orchestration. We refer to this end-to-end capability as the *Form Fill* system. In this section, we formalize the Form Fill problem, describe the hybrid online–offline architecture used in our deployment, and summarize its empirical performance under realistic noise and backend constraints.

Form Fill operates over multi-turn, voice-based conversations that are first transcribed by an Automatic Speech Recognition (ASR) system and then processed by an LLM. Formally, we model each interaction as a dialogue ($D = (u_1, \ldots, u_T)$) of transcribed user and agent utterances, and the target output as a structured record $y = (y_1, \ldots, y_K)$ comprising fields needed by downstream systems (e.g., name, phone number, email addresses, contact details, preferences, question responses) together with derived actions such as follow-up questions or appointment requests. Each conversation is associated with a *form* template that specifies which pieces of information should be collected. We denote the form schema by $\mathcal{F} = \{(q_k, \tau_k)\}_{k=1}^{K}$, where q_k is the k-th scripted question and τ_k is the type or domain of the corresponding field (e.g., name, phone number, email address, preferences). The target output of Form Fill is then a structured record, with each y_k belonging to τ_k or a special symbol $\perp$ indicating that the field is not filled, i.e., $y_k \in \{\tau_k, \perp\}$.

The core difficulty is that user-provided information is often fragmented, off-script, and revised over time: answers may be implicit, spread across multiple turns, or contradicted and corrected later in the call. The questions could be asked in various ways and orders, spread over multiple turns, or completely skipped. In addition, ASR and LLM interaction introduces its own class of errors, so the LLM must extract correct structured information from noisy text while maintaining alignment between evolving dialog context and the underlying information needs.

6.3.1 Hybrid Online-Offline Architecture

To handle these challenges, we implement Form Fill as a hybrid online-offline process rather than a single-pass extractor.

The online component runs during the conversation and is optimized for responsiveness and script alignment. During each conversation, it maintains the form questions $\{q_1, \ldots, q_K\}$ from the schema $\mathcal{F}$ defined above and, for each dialog prefix $D_{1:t}$, performs *question detection* to identify which question, if any, is currently being addressed. Conditional on a non-null prediction $\hat{\imath}_t \in \{1, \ldots, K\}$, an *answer extraction* component then proposes a candidate value $\hat{y}_{\hat{\imath}_t}^{\text{online}}$ based on a localized context window around u_t. This coupling between the detected question and extraction context reduces spurious matches and keeps the evolving record aligned with the script as the call unfolds.

After the call completes, an offline phase operates over the full transcript D. A *reconciliation* step re-scans the entire conversation and, for each field k, collects one or more candidate answers together with supporting evidence, producing a second set of candidates $\{\hat{y}_k^{\text{offline}}\}_{k=1}^{K}$ that is independent of the online alignment. Because it has access to all turns, this step can recover answers that were missed online (for example, when the user answers a question before it is asked or much later in the call) and flag fields for which multiple, conflicting values were mentioned. A subsequent *arbitration* step then combines online and offline candidates, together with other metadata to select a final value y_k or mark the field as unresolved.

In effect, the online component provides a low-latency initial record, while the offline reconciliation–arbitration pipeline acts as a consistency check and late-correction mechanism that improves overall accuracy without impacting the live user experience. All reported accuracy numbers in this section are computed over the final post-arbitration record y.

6.3.2 Edge Cases, Safety Mechanisms, and Downstream Workflows

In practice, Form Fill must remain reliable in the presence of several recurring edge cases. High-stakes fields such as names, phone numbers, email addresses, medication names, and dates are particularly sensitive to ASR errors, tokenization issues, and model hallucinations: a single substitution or transposition can result in an erroneous record. To reduce silent failures on these fields, the system applies a *require confirmation* policy: candidate values are surfaced back to the user in natural language, and only explicitly confirmed values are committed to the structured record. This introduces a small interaction cost but substantially lowers the risk of incorrect records.

A second failure mode is *de-synchronization* between the conversation and the underlying form, especially in longer calls with digressions and clarifications. If the model over-relies on distant context, it may attach an answer to the wrong question or overwrite a previously correct field with an unrelated value. To mitigate this, the online question detection component is constrained to operate over a narrow, recency-weighted context window that is explicitly anchored to the current script state. This reduced-context design makes the question detection step more reliable.

Finally, the structured record y produced by Form Fill drives downstream workflows such as appointment scheduling and follow-up creation. In these cases, the system must translate y into concrete tool calls (for example, constructing API requests to scheduling backends) and ensure that the resulting actions respect constraints such as availability, timing, and basic eligibility rules. We therefore treat tool invocation as part of the Form Fill pipeline and apply the same principles of explicit confirmation for high-impact actions, conservative handling of ambiguous inputs, and post-hoc consistency checks before any irreversible operation is executed.

6.3.3 Evaluation and Empirical Performance

We evaluate Form Fill at the level of individual fields and end-to-end task outcomes. For structured records, we use field-level exact-match accuracy: the fraction of fields whose final value y_k exactly matches a human-validated label. This is directly analogous to slot accuracy in dialogue state tracking. For Form-Fill evaluation, we ran human review on a randomly sampled 2.5% of all production calls, with sampling configured to cover diverse use cases and a mix of long and short conversations and forms with many or few questions. In these audits, human experts re-checked every Form Fill field.

Across two recent evaluation windows in Q4 2025, the fraction of fields that required correction was 0.117% and 0.17%, corresponding to field-level exact-match accuracies of 99.883% and 99.83%, respectively. Among calls that contained at least one Form Fill error, 60% were attributable to upstream ASR problems, or application orchestration issues, including cases where Form Fill was not configured, and 40% were due to intrinsic detection or extraction errors in the Form Fill model itself. At the call level, the fraction of audited calls with at least one corrected Form Fill field was 0.58% and 0.44%.

Our model's observed accuracy improved from 98.5% (Polaris 3) to 99.86% [3] in Polaris 4, with a substantial subset of residual errors arising outside the core detection and extraction components.

7 Multilingual Continuity and Equity

Building a safe and equitable phone-based clinical AI assistant requires robustness across linguistic, cultural, and interactional variability. Spoken dialogue introduces cascading sources of error—accents, dialects, code-switching, low-resource language features, and pragmatic differences—that disproportionately affect non-English speakers. These breakdowns can directly translate into safety risks during real-time care. We highlight the key failure modes and our corresponding mitigation strategies below.

7.1 Accuracy Gaps in ASR/TTS/LLM

Multilingual speech understanding remains challenging: multilingual ASR systems typically regress in word-error rate (WER) relative to language-specific models. A common failure—such as interpreting the Spanish "sí" ("yes") as the English letter "C" or the English terms "Sea" or "See"—illustrates how small transcription errors can propagate into incorrect clinical confirmations. Although English ASR achieves low WER, many languages show substantial degradation. Arabic is particularly difficult due to wide phonetic variability, scarce training data, and major dialectal divergence (e.g., Hejazi, Emirati, Khaleeji). Medication terminology further amplifies these issues, with generic Arabic ASR systems often exceeding 30% WER on medication names.

To mitigate these errors, our system performs continuous language identification and rapid automatic switching across ASR, TTS, and LLM modules. Rather than assuming a monolingual caller, it maintains a parallel multilingual safety state for English, Spanish, and Arabic, enabling millisecond-level realignment when users code-switch. Medication entities, numerals, and safety-critical context are preserved across language boundaries.

For Arabic, we use multi-ASR ensembling with medication-focused decoding pipelines. Models vote on medication entities using phonetic similarity scoring and transliteration harmonization, substantially reducing the high WER typical of general-purpose systems.

7.2 Dialectal and Cultural Variability

Dialectal and cultural differences shape how callers express needs, describe symptoms, and interpret tone. Spanish speakers vary in politeness markers and directive strength across Mexico, Puerto Rico, and U.S. immigrant communities. Arabic dialects diverge to the point that phrases acceptable in Modern Standard Arabic may sound overly formal or confusing in Gulf dialects. These stylistic shifts require dynamic detection and adaptation mid-call to maintain clarity and trust.

Table 9: Safety Performance of Polaris Model Family and Human Clinicians.

Model	Correct Advice	No Harm	Minor Harm	Severe Harm	Death
Polaris 4.0	99.90%	0.10%	0.00%	0.00%	0.00%
Polaris 3.0	99.38%	0.55%	0.07%	0.00%	0.00%
Polaris 2.0	98.75%	1.02%	0.13%	0.10%	0.00%
Polaris 1.0	93.23%	4.55%	1.83%	0.32%	0.06%
Human Clinicians	81.16%	14.72%	4.12%	0.00%	0.00%

7.3 Mid-Call Language Switching and Multilingual Interference

Code-switching—especially between English and Spanish or Arabic—is common when callers reference numbers, medications, or insurance/legal terms. Traditional ASR systems often treat this as noise, causing incorrect entity extraction and mismatched safety prompts. Multilingual ASR models can also blend languages, silently injecting cross-lingual synonyms that lead to unsafe LLM inferences. Spanish and Arabic illustrate two critical but distinct cases: Spanish as a high-volume U.S. language with strong bilingual patterns, and Arabic as a lower-resource, highly dialectal language. Our multilingual and dialect-aware safety strategies produced measurable improvements in deployment. We highlight the case of a patient contacted during one of our summary heat-wave safety outreach programs. Although the AI agent began the call by introducing itself in English, the patient responded solely in Spanish, stating that they did not speak any English. Our system recognized the mid-call language shift and automatically switched the ASR, TTS, and LLM components to their Spanish counterparts, and completed the rest of the call in Spanish.

8 Uncompromising Clinical Safety

Safety remains the cornerstone of the system, with Polaris 4 achieving a 99.9% no-error rate (0.1% no-harm errors and 0% minor harm, severe harm or death) across all connected calls as shown in Table 9, surpassing prior versions and even average human clinician performance for equivalent tasks[2]. The clinical escalation system relies on specialized agents—covering labs and vitals, medications, and escalations—with higher accuracy and more up-to-date knowledge than prior versions, which is especially relevant for understanding new medications. These agents collaborate to ask targeted follow-up questions, reducing aggregate over-escalation rates while preserving safety. In Polaris 4, the labs and vitals specialist had an error rate of 0.005%, the medication specialist 0.01%, and the escalation specialist 0.07%, with all errors across agents classified as "no harm". These results show a significant improvement when compared to Polaris 3, as shown in Table 10.

In August of 2025, the overall escalation rate out of all connected calls—defined as any call requiring immediate transfer to a human or review within 24 hours—was 0.77%, down from 3.4% in June of 2025, while the calls that required an immediate transfer were 0.26%, down from 1.22%. At the same time, the proportion of connected calls categorized as "no-harm" decreased fivefold, from 0.5% to 0.1%. This reduction balances autonomy with human oversight, minimizing unnecessary transfers that could overwhelm human teams [24, 2, 23] while developing the safest generative AI system for healthcare.

[2] Assessed by Human US Licensed Physicians and US Licensed Nurses

Table 10: Escalation rates and error rates for Polaris 3 and 4.

Metric	Polaris 4.0	Polaris 3.0
Escalation Rate – Overall	0.77%	3.40%
Escalation Rate – Immediate	0.26%	1.22%
Error Rate – Lab/Vital	0.005%	0.06%
Error Rate – Medication	0.01%	0.02%
Error Rate – Escalation	0.07%	0.32%

9 Evaluation at Scale: RWE-LLM in Practice

9.1 Overview of the Evaluation Framework

The evaluation of the updated AI care agent builds on the Real-World Evidence LLM (RWE-LLM) methodology described in the Hippocratic AI Safety Framework [8]. This methodology integrates clinician simulation, on-policy testing, automated LLM-based rater assessments, and retrospective safety reviews. Together, these components allow for continuous validation of safety, reliability, conversational quality, and equity performance. Unlike traditional model evaluation pipelines that rely solely on offline test sets, the RWE-LLM system incorporates evidence from large-scale, production-proximal interactions, enabling more accurate detection of failure modes and contextual performance variation.

9.2 Clinician Simulation

Clinician simulation serves as the foundation of the evaluation process. More than 7,000 licensed clinicians have participated in simulation efforts to date, generating over 500,000 structured test calls in which each clinician interacts with the AI as they would with a real patient. These interactions capture detailed feedback on clinical reasoning, medication safety, benefit and identity verification, symptom clarification, and escalation accuracy. Prior research demonstrates that clinician-generated labels reliably capture safety-critical judgments in conversational agents [27, 28]. The resulting label corpus provides training and calibration signals for verifier models, alignment layers, and task-specific guardrails. This simulation pathway allows early identification of confusion patterns, conversational breakdowns, or clinical reasoning deficits before models are exposed to patient-facing environments.

9.3 On-Policy Evaluation Under Real Conditions

On-policy evaluation complements clinician simulation by testing the system under realistic conversational conditions. These evaluations expose the AI agent to natural user variability, including accent diversity, background noise, partial disclosures, interruptions, and heterogeneous communication styles. Model variants are compared across metrics such as safety-event frequency, escalation accuracy, procedural correctness, benefits verification accuracy, and patient-reported satisfaction. This methodology is especially important for assessing the governed orchestration layer described in the Safety Framework, which enforces preconditions, input validation, and post-condition checks to ensure safe execution of sensitive tasks [8]. On-policy testing is therefore the primary mechanism for detecting safety breakdowns that may not emerge in scripted test environments.

9.4 Automated Rater Assessment

Automated raters extend evaluation coverage by applying LLM-based scoring systems trained on clinician-generated labels. These raters assess conversational quality, coherence, tone, empathy alignment, motivational interviewing technique, clarification depth, and policy adherence. Automated evaluation enables continuous model monitoring and rapid iteration cycles by scoring every call rather than a small subset. This is particularly important in memory-enabled contexts, where the agent must make contextually appropriate references to prior interactions. Evidence from the multi-call memory study indicates that each additional memory reference extends call duration by approximately 2.47 minutes without lowering satisfaction [15], underscoring the value of rater systems that evaluate memory relevance, appropriateness, and timing.

9.5 Retrospective Safety Review

Retrospective reviews serve as the final component of the RWE-LLM framework. These reviews examine transcripts, escalation logs, incident reports, and verified safety events to identify emergent or rare error modes. The insights generated from these reviews feed into updates to safety verifiers, escalation logic, tone-modulation strategies, and conversational-policy constraints. As emphasized in the Safety Framework, ongoing retrospective analysis plays a crucial role in ensuring that model evolution remains aligned with clinical expectations and organizational governance requirements [8]. Together, these review processes integrate past experience into future system behavior, forming a closed-loop safety ecosystem.

9.6 External Validation Through Case Studies

Real-world deployments further validate the RWE-LLM framework. Case studies from large health-plan implementations demonstrate that the evaluation architecture remains robust when applied to high-volume, complex workflows such as benefit verification, care-gap closure, appointment coordination, longitudinal care management, and multi-call continuity [8]. Across these domains, the RWE-LLM system consistently identified failure modes early and guided model updates that improved reliability and safety. This cross-setting consistency supports the framework's generalizability across clinical, administrative, and preventive-care contexts.

10 Operational and Clinical Impact Across Settings

10.1 Overview

Deployments of the AI care agent across administrative, clinical, and preventive-care workflows have generated consistent improvements in operational capacity, engagement, and clinical process reliability. Health systems report increased workflow throughput, more reliable benefits verification, improved care-gap closure, and greater adherence to chronic-care protocols [1]. These gains emerge from the agent's ability to handle repetitive communication tasks with high consistency, freeing clinical and administrative staff to focus on cases requiring human judgment.

10.2 Safety Validation at Clinical Scale

Prior to real-world deployment, the AI care agent underwent the largest empirical safety validation study conducted for healthcare AI to date. This nationwide evaluation encompassed 306,965 unique clinical interactions assessed by 6,527 licensed US clinicians—including 6,294 registered nurses and 233 physicians—spanning all 50 states and the District of Columbia.[8]

The validation framework employed a novel three-tier methodology emphasizing comprehensive output testing rather than input validation. Tier 1 involved direct AI–clinician interaction testing, where participating clinicians engaged with the system and flagged safety concerns. Tier 2 consisted of systematic review by specially trained internal nursing teams who assessed clinical validity of flagged concerns and determined severity levels. Tier 3 provided independent physician adjudication for complex cases, with emergency medicine and primary care physicians delivering definitive clinical judgment on safety implications.

This approach directly addressed fundamental limitations in current healthcare AI evaluation, which typically relies on benchmark testing of hundreds rather than hundreds of thousands of interactions. By testing actual system outputs across diverse clinical scenarios—including routine care coordination, medication management, chronic disease education, and emergency situations requiring escalation—the framework provided validation coverage orders of magnitude greater than traditional approaches.

The validation demonstrated substantial measurable safety improvements across four developmental iterations of the system. Correct medical advice rates progressed from approximately 80% in baseline testing to 99.58% (95% CI: 99.53%–99.63%) in the final version.[8] This 19.58 percentage point improvement represents a clinically meaningful advancement achieved through systematic validation-driven development.

Critically, potentially harmful advice was reduced to near-zero levels. Incorrect advice with potential for minor harm declined from 1.89% (95% CI: 1.67%–2.14%) in early iterations to 0.08% (95% CI: 0.06%–0.11%) in the final version—a 95.8% relative reduction. Incorrect advice with risk of severe harm decreased from 0.33% (95% CI: 0.25%–0.44%) to 0.00%, representing complete elimination of severe harm risk. Risk-of-death errors, initially present at 0.05% (95% CI: 0.03%–0.11%), were eliminated entirely in later versions, maintaining 0.00% rates across subsequent iterations.[8]

These quantitative outcomes establish new empirical benchmarks for healthcare AI safety, demonstrating that systematic validation can achieve safety standards exceeding typical human clinical communication benchmarks. The progressive improvement across system iterations provides the first large-scale empirical evidence that comprehensive pre-deployment validation can ensure AI safety in healthcare settings, challenging the widespread assumption that safety can be inferred from training data quality alone.

10.3 Impact on Chronic Disease Monitoring

The remote patient monitoring program in nephrology provides one of the clearest illustrations of system-level clinical impact. Among 5,590 older adults across 18 states, a single AI-delivered welcome call more than doubled maximum call duration (205 to 431 seconds), increased verified call rates from 11.9% to 30.5%, and increased call-completion rates from 46.2% to 62.4% [14]. These effects persisted across age, sex, and region, with regression models explaining less than 2% of variance, demonstrating broad generalizability even in patients aged 75 years and older. Improved engagement led directly to more reliable monitoring of blood pressure, faster follow-up for abnormal values, and improved adherence to chronic kidney disease management protocols.

10.4 Longitudinal Interaction Quality Through Multi-Call Memory

The addition of multi-call memory further strengthens the agent's impact across longitudinal care pathways. Memory-enabled conversations were shown to increase behavioral engagement substantially, with each additional memory reference increasing call duration by an average of 2.47 minutes [15]. Although satisfaction levels remained stable, the increased depth and continuity of these conversations allowed patients to engage more fully, supporting richer discussions and more coherent interactions across repeated encounters. This type of longitudinal coherence is especially valuable in chronic disease management, where repeated reinforcement of goals and understanding of patient-specific barriers are essential.

10.5 Reduced Disparities Through Multilingual Preventive Outreach

Preventive outreach results demonstrate the potential of AI agents to reduce disparities in population health. In a multilingual colorectal cancer screening initiative, Spanish-speaking patients—historically exhibiting lower screening rates—showed significantly higher engagement, including connect rates of 69.6% versus 53.0% among English speakers, and more than twice the FIT test opt-in rate (18.2% vs. 7.1%).[26] After adjusting for demographic and call-level variables, Spanish-speaking patients remained twice as likely to opt in to screening (adjusted OR 2.012) (Bhimani et al., 2025). These findings challenge the assumption that AI disproportionately disadvantages non-English-speaking populations and instead suggest that language-concordant AI communication can meaningfully reduce screening disparities.

10.6 Workflow Integration and Health-Plan Outcomes

Large-scale health-plan deployments demonstrate additional operational benefits. Organizations using the AI agent for outreach, care-gap closure, and member engagement reported higher connection rates and more consistent reach across eligible populations.[1] These findings align with the Safety Framework's emphasis on workflow integration, in which AI systems augment rather than replace staff by taking on repetitive communication tasks and providing stable capacity during periods of variable staffing. In the context of chronic disease monitoring, increased engagement observed in the nephrology remote-monitoring cohort further demonstrates how improved communication consistency can support downstream adherence behaviors.[14]

10.7 System-Wide Deployment and Operational Efficiency

The transition from pilot studies to enterprise-wide implementation represents a critical inflection point for healthcare AI. A 13-month prospective implementation study at WellSpan Health—an integrated health system serving south-central Pennsylvania and northern Maryland—provides the first comprehensive evidence of autonomous AI deployment as a system-wide capability rather than an isolated intervention.[29] From September 2024 through September 2025, the AI voice assistant ("Ana") conducted nearly 2 million patient conversations across three strategic deployment categories: targeted outbound campaigns for care gap closure, integrated automated outreach for routine patient communications, and inbound call management for patient-initiated inquiries.[29] This scale of deployment—unprecedented for an autonomous healthcare AI agent—demonstrates the feasibility of AI integration as an enterprise capability.

10.7.1 Procedural Preparation and Patient Education

In a colonoscopy preparation pilot, the AI agent contacted 1,627 patients who opted in to receive preparation coaching through a series of structured calls. Among these patients, 30.8% (95% CI: 28.6%–33.0%) completed full conversations with the agent. Patient experience metrics were notably strong: among 460 patients providing satisfaction ratings, 65.9% (95% CI: 61.5%–70.3%) rated likelihood-to-recommend as 9 or 10 on a 10-point scale, with a mean rating of 8.65.[29] Qualitative analysis of patient feedback identified recurring themes of perceived empathy, patience, and appreciation for unlimited question opportunities—characteristics typically associated with high-quality human clinical communication.

10.7.2 Diagnostic Results Communication

For mammogram results delivery, the AI agent reached 11,000 patients with normal results requiring notification. The system successfully connected with 5,019 patients (45.6%, 95% CI: 44.7%–46.6%) and completed full conversations with 2,734 (24.9%, 95% CI: 24.1%–25.7%). Patient satisfaction and likelihood-to-recommend ratings both exceeded 9 on a 10-point scale. The average duration of the conversation of 3.3 minutes resulted in more than 350 hours of direct patient engagement for education on annual mammograms and scheduling assistance for subsequent screenings [29].

10.7.3 Workforce Augmentation in Primary Care

The most substantial operational impact emerged in primary care call center deployment. WellSpan's primary care call centers had historically struggled with staffing shortages, operating at only 50% practice coverage with significant patient wait times. Following AI implementation, the system achieved dramatic capacity expansion [29]:

- Practice coverage expanded from 50% to 100% of primary care practices
- The AI agent managed more than 50% of all phone-scheduled appointments
- Weekly talk time averaged 850 hours, equivalent to the workload of 28 full-time call center specialists
- Patient refusal to engage with the AI agent remained below 3%
- Efficiency gains effectively doubled staff productivity, enabling the call center to manage twice the workload without additional human staffing

The AI agent was initially deployed to handle routine requests—operating hours, directions, parking information—freeing staff for complex calls requiring human judgment. Subsequently, its role expanded to include scheduling acute and routine primary care appointments, demonstrating successful scope extension based on validated performance.

10.7.4 Implementation Success Factors

Several factors distinguished this enterprise implementation from typical pilot studies. The health system treated AI as a system-wide capability requiring governance standards rather than an isolated technology deployment. Multidisciplinary teams including nurses, physicians, administrators, and quality improvement specialists collaborated with the AI development team to map patient journeys, identify intervention points, create conversation scripts, and develop safety protocols for clinical escalation.[29] Frontline staff involvement throughout design and implementation proved critical—staff created and reviewed test conversations, provided feedback to refine delivery, and developed escalation pathways that route clinical or urgent issues appropriately. This user-centered approach, combined with clinical oversight, enabled the transition from isolated pilots to sustained operational deployment at scale.

10.8 Patient Experience and Satisfaction Outcomes

Across diverse clinical applications, the AI care agent consistently achieved patient satisfaction metrics comparable to or exceeding benchmarks for human-delivered care. This pattern of high satisfaction emerged with patients' awareness that they were interacting with an AI system.

In the WellSpan implementation, satisfaction scores ranged from 8.65 to above 9.0 on 10-point scales across all deployment categories.[29] The colonoscopy preparation pilot achieved a mean likelihood-to-recommend score of 8.65/10, with nearly two-thirds of respondents (65.9%) providing promoter-level ratings of 9 or 10. Mammogram results delivery maintained satisfaction and likelihood-to-recommend averages above 9/10.

Qualitative feedback from patients revealed several themes explaining high satisfaction with AI-mediated communication [29]:

- Perceived empathy: Patients frequently commented on the AI agent's empathetic tone and apparent understanding of their concerns, suggesting successful implementation of conversational design principles emphasizing warmth and acknowledgment.
- Patience and availability: Unlike time-constrained human interactions, patients appreciated the AI agent's unlimited availability for questions without perceived time pressure—a characteristic documented in prior qualitative research on patient preferences for AI communication.[30].
- Consistency: The standardized yet personalized delivery ensured all patients received complete, accurate information regardless of when they called or which agent instance they reached.

These satisfaction findings challenge assumptions that patients inherently prefer human communication for healthcare interactions. When AI systems are designed specifically for empathetic, patient-centered conversation, rather than transactional information exchange, patient acceptance and satisfaction can match or exceed traditional delivery models. The high satisfaction levels observed across diverse use cases (e.g., procedural preparation, diagnostic results, appointment scheduling) suggest this pattern generalizes across healthcare communication contexts.

10.9 Summary of Impact

The accumulating evidence supports the conclusion that AI-mediated communication can enhance care delivery by augmenting rather than replacing human clinicians, expanding care-team capacity, ensuring consistent high-quality interactions for all patients, and—when properly validated—achieving safety standards that match or exceed human clinical communication. The framework demonstrates that treating interaction intelligence as a first-class safety variable, combined with rigorous pre-deployment validation, enables deployment of autonomous AI agents that improve outcomes across the quadruple aim of healthcare: better patient experience, improved population health, reduced costs, and enhanced clinician well-being through workload redistribution.

11 Related Work

Traditional healthcare LLM benchmark evaluations were mostly hinged on static, offline benchmarks which usually materialized as multi-choice Q/A fashion or report summarization. While such datasets offer convenience and reproducibility, they systematically miss the dynamic signals including contextual drift, longitudinal consistency, interpersonal variability, tool-use, and uncertainty negotiation that arise in authentic interactions. These benchmarks span a range of formats including classification, question answering, text and code generation, but share a common design pattern: each example is a self-contained input (a note, report, question, or schema) paired with a single "correct" output, and models are scored with pointwise metrics such as exact match, F1, or LLM-jury scores. For example, MedCalc-Bench[38], CLEAR[39], Medec[40], EHRSHOT[41], and ADHD-Behavior[42] / ADHD-MedEffects[43] evaluate classification or computational reasoning from notes and EHR codes, asking models to detect conditions, compute risk or severity, or flag documentation errors. Knowledge-focused QA sets such as HeadQA[44], MedBullets[45], MedQA[46], MedMCQA[47], PubMedQA[48], MedicationQA[49] test exam-style or snippet-grounded questions with multiple-choice or binary answers. A large family of generation benchmarks—DischargeMe[50], MedAlign[51], ACI-Bench[52], MIMIC-RRS[53], MIMIC-IV-BHC[54], and MedDialog[55]—measure how well models summarize notes, radiology reports, or conversations and produce treatment plans, discharge instructions, or empathetic counseling responses, typically via rubric-based LLM juries. Other datasets target operational and safety-adjacent tasks such as identifying PHI or privacy risk (MedConfInfo[56], PrivacyDetection[58]), proxy senders (ProxySender[58]), hallucinations (MedHallu[59]), or research-oriented code generation from natural language (EHRSQL[60]). While these benchmarks are valuable for coverage and comparability across models, they all instantiate the static offline paradigm: the model is evaluated on frozen, de-identified artifacts, with no real-time interaction, feedback, or evolving context. Tasks are typically single-shot and decontextualized (e.g., one question, one note, one discharge summary), and success is reduced to matching a reference label or receiving a high rubric score on a single response. As a result, these datasets provide snapshots of task competence rather than measuring how a system behaves over multi-turn, safety-critical episodes: they do not capture longitudinal patient trajectories, dynamic clinical decision-making, tool use, knowledge drift, user misunderstanding, or the role of redundancy and cross-checking between agents.

12 Conclusion

By grounding the system design for Polaris in real patient interactions and governed telemetry, we reliably improved safety, empathy, equity, and workflow outcomes at clinical scale to build the safest generative AI system for healthcare. Production telemetry drives the architecture. Elevating interaction micro-skills to safety variables, investing in modality-specific models (e.g., contextual ASR), and leveraging hardware-aware serving unlocks non-linear gains for voice in healthcare. Orchestration designed for real workflows, validated under RWE-LLM, turns conversational quality into reliable clinical and operational outcomes. The production-first approach links signals to solutions to impact, providing a repeatable path for continued progress.

Acknowledgments

We thank partner health systems, clinicians, and patients whose feedback shaped this work, and the engineering and research teams who built the system.

References

[1] Hippocratic AI. Company Homepage and Results. Available at: https://www.hippocraticai.com/ (accessed 2025).

[2] Hippocratic AI. Our Multi-Skilled Agents Designed for Healthcare. Available at: https://hippocraticai.com/skills/ (accessed 2025).

[3] Hippocratic AI. Polaris 3.0 Safety Constellation Architecture. Available at: https://hippocraticai.com/polaris-3/ (accessed 2025).

[4] Chaurasia, A. Improving Patient Engagement Through Personalized Interactions. Hippocratic AI Blog. Available at: https://hippocraticai.com/personalized-interactions/ (2025).

[5] Hippocratic AI. The Human Touch in AI. Available at: https://hippocraticai.com/the-human-touch-in-ai/ (accessed 2025).

[6] Hippocratic AI. Empathetic Intelligence. Available at: https://hippocraticai.com/empathetic-intelligence/ (accessed 2025).

[7] Hippocratic AI. Empathy in Action. Available at: https://hippocraticai.com/empathy-in-action/ (accessed 2025).

[8] Real World Evaluation of Large Language Models in Healthcare (RWE-LLM). medRxiv (2025). Available at: https://www.medrxiv.org/content/10.1101/2025.03.17.25324157v1.

[9] Jin, Q. et al. *What Disease Does This Patient Have? A Large-Scale Open-Domain Medical QA Dataset*. arXiv:2009.13081 (2020). MedQA (USMLE). Available at: https://arxiv.org/abs/2009.13081.

[10] Pal, A. et al. *MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset for the Medical domain*. arXiv:2210.10531 (2022). Available at: https://arxiv.org/abs/2210.10531.

[11] Jin, Q. et al. *PubMedQA: A Dataset for Biomedical Research Question Answering*. arXiv:1909.06146 (2019). Available at: https://arxiv.org/abs/1909.06146.

[12] Singhal, K. et al. *Large Language Models Encode Clinical Knowledge*. arXiv:2212.13138 (2022); and *Towards Expert-Level Medical Question Answering with Large Language Models* (MultiMedQA/Med-PaLM), arXiv:2305.09617 (2023). Available at: https://arxiv.org/abs/2305.09617.

[13] Hendrycks, D. et al. *Measuring Massive Multitask Language Understanding*. arXiv:2009.03300 (2020). Available at: https://arxiv.org/abs/2009.03300.

[14] Agnew, J. D., Parikh, K., Raglow-Defranco, M., & Bhimani, M. (2025). *AI-delivered welcome calls and patient engagement in remote blood pressure monitoring: A multi-site study in nephrology care*. Manuscript under review at JMIR AI.

[15] Sanz Ausin, M., Chaurasia, A., Miller, A., Agnew, J. D., Lasko, R., Raglow-Defranco, M., Voisard, M., Godil, S., & Mukherjee, S. (2025). *Optimizing multi-call memory in AI healthcare communication: A mixed-effects analysis of engagement and satisfaction outcomes*. Manuscript under review at JMIR Formative Research.

[16] T. Stivers *et al.* Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences (PNAS)*, 106(26):10587–10592, 2009. doi:10.1073/pnas.0903616106.

[17] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi:10.3389/fpsyg.2015.00731.

[18] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In J. Schenkein, editor, *Studies in the Organization of Conversational Interaction*, pages 7–55. Academic Press, 1978. doi:10.1016/B978-0-12-623550-0.50008-2.

[20] J. Nielsen. *Usability Engineering.* Morgan Kaufmann, San Francisco, 1993. ISBN 978-0-12-518406-9.

[21] L. Iyer, K. Aggarwal, S. Koyejo, G. Heyman, D. C. Ong, and S. Mukherjee. HEART: A Unified Benchmark for Assessing Humans and LLMs in Emotional Support Dialogue. *arXiv preprint*, 2025.

[22] Subhabrata Mukherjee, "Polaris 2.0," Hippocratic AI, Oct. 1, 2025, https://hippocraticai.com/polaris2/ (accessed Jan. 11, 2026).

[23] Subhabrata Mukherjee, "Hippocratic AI Releases Polaris 3.0: A 4.2 Trillion Parameter Suite of 22 LLMs, Enhancing Patient Safety and Experience by Leveraging Real World Experiences," Hippocratic AI, Mar. 19, 2025, https://hippocraticai.com/polaris-3/ (accessed Jan. 11, 2026).

[24] Subhabrata Mukherjee, "Polaris 4: Redefining Patient & AI Interactions for Clinical Excellence," Hippocratic AI, Nov. 20, 2025, https://hippocraticai.com/polaris-4/ (accessed Jan. 11, 2026).

[25] S. Mukherjee, P. Gamble, M. Sanz Ausin, N. Kant, K. Aggarwal, N. Manjunath, D. Datta, Z. Liu, J. Ding, S. Busacca, C. Bianco, S. Sharma, R. Lasko, M. Voisard, S. Harneja, D. Filippova, G. Meixiong, K. Cha, A. Youssefi, M. Buvanesh, H. Weingram, S. Bierman-Lytle, H. S. Mangat, K. Parikh, S. Godil, and A. Miller. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare. *arXiv preprint arXiv:2403.13313*, 2024.

[26] Bhimani, M., Baker, R. H., Sanz Ausin, M., Meixiong, G., Lasko, R., Raglow-Defranco, M., Miller, A., Mukherjee, S., Godil, S., Cook, A., Agnew, J. D., & Atreja, A. (2025). *Using a multilingual AI care agent to reduce disparities in colorectal cancer screening for higher fecal immunochemical test adoption among Spanish-speaking patients: Retrospective analysis. Journal of Medical Internet Research, 27*, e71211. Available at https://doi.org/10.2196/71211.

[27] Bickmore, T., & Schulman, D. (2013). Embodied agents for long-term interaction. In Intelligent virtual agents. Springer. Available at https://repository.library.northeastern.edu/files/neu:915/fulltext.pdf (accessed 2025).

[28] Zhang, Z., Bickmore, T., Mainello, K., Mueller, M., Foley, M., Jenkins, L., & Edwards, R. A. (2014, August). Maintaining continuity in longitudinal, multi-method health interventions using virtual agents: The case of breastfeeding promotion. In International conference on

intelligent virtual agents (pp. 504-513). Cham: Springer International Publishing. Available at https://link.springer.com/chapter/10.1007/978-3-319-09767-1_61 (accessed 2025).

[29] Kandrysawtz, M., Vega, D., Beilis, H., Fortson, D., Agnew, J. D., & Bhimani, M. (2025). Keys to success: Launching the world's first autonomous agentic AI solution in a health system [Manuscript submitted for publication]. NEJM Catalyst.

[30] Trivedi, R., Shaw, T., Sheahen, B., Chow, C. K., & Laranjo, L. (2025). Patient perspectives on conversational artificial intelligence for atrial fibrillation self-management: Qualitative analysis. Journal of Medical Internet Research, 27, e64325. https://doi.org/10.2196/64325 (accessed 2025).

[31] Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. Audio-Aware Large Language Models as Judges for Speaking Styles. arXiv preprint arXiv:2506.05984, 2025. https://arxiv.org/abs/2506.05984.

[32] A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. A. Roberts, "The Unreasonable Ineffectiveness of the Deeper Layers," *arXiv:2403.17887*, 2024. Available at: https://arxiv.org/abs/2403.17887.

[33] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023) *Robust speech recognition via large-scale weak supervision. International conference on machine learning*

[34] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019) *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Interspeech*

[35] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large Language Models in Medicine: The Potentials and Pitfalls," *Annals of Internal Medicine*, vol. 177, no. 2, pp. 210–220, 2024.

[36] S. Johri, J. Jeong, B. A. Tran *et al.*, "An Evaluation Framework for Clinical Use of Large Language Models in Patient Interaction Tasks," *Nature Medicine*, vol. 31, pp. 77–86, 2025.

[37] S. V. Blackley, J. Huynh, L. Wang, Z. Korach, and L. Zhou, "Speech Recognition for Clinical Documentation from 1990 to 2018: A Systematic Review," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 324–338, 2019.

[38] Khandekar, N., Jin, Q., Xiong, G., Dunn, S., Applebaum, S., Anwar, Z., Sarfo-Gyamfi, M., Safranek, C., Anwar, A., Zhang, A. and Gilson, A., 2024. Medcalc-bench: Evaluating large language models for medical calculations. Advances in Neural Information Processing Systems, 37, pp.84730-84745.

[39] Lopez, I., Swaminathan, A., Vedula, K., Narayanan, S., Nateghi Haredasht, F., Ma, S.P., Liang, A.S., Tate, S., Maddali, M., Gallo, R.J. and Shah, N.H., 2025. Clinical entity augmented retrieval for clinical information extraction. npj Digital Medicine, 8(1), p.45.

[40] Abacha, A.B., Yim, W.W., Fu, Y., Sun, Z., Yetisgen-Yildiz, M., Xia, F. and Lin, T., 2025, July. Medec: A benchmark for medical error detection and correction in clinical notes. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 22539-22550).

[41] Wornow, M., Thapa, R., Steinberg, E., Fries, J. and Shah, N., 2023. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. Advances in Neural Information Processing Systems, 36, pp.67125-67137.

[42] Pillai, M., Posada, J., Gardner, R.M., Hernandez-Boussard, T. and Bannett, Y., 2024. Measuring quality-of-care in treatment of young children with attention-deficit/hyperactivity disorder using pre-trained language models. Journal of the American Medical Informatics Association, 31(4), pp.949-957.

[43] Bannett, Y., Gunturkun, F., Pillai, M., Herrmann, J.E., Luo, I., Huffman, L.C. and Feldman, H.M., 2025. Applying Large Language Models to Assess Quality of Care: Monitoring ADHD Medication Side Effects. Pediatrics, 155(1), p.e2024067223.

[44] Vilares, D. and Gómez-Rodríguez, C., 2019. HEAD-QA: A healthcare dataset for complex reasoning. arXiv preprint arXiv:1906.04701.

[45] Chen, H., Fang, Z., Singla, Y. and Dredze, M., 2025, April. Benchmarking large language models on answering and explaining challenging medical questions. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 3563-3599).

[46] Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H. and Szolovits, P., 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), p.6421.

[47] Pal, A., Umapathi, L.K. and Sankarasubbu, M., 2022, April. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on health, inference, and learning (pp. 248-260). PMLR.

[48] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. and Lu, X., 2019, November. Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 2567-2577).

[49] Abacha, A.B., Mrabet, Y., Sharp, M., Goodwin, T.R., Shooshan, S.E. and Demner-Fushman, D., 2019. Bridging the gap between consumers' medication questions and trusted answers. In MEDINFO 2019: Health and Wellbeing e-Networks for All (pp. 25-29). IOS Press.

[50] Xu, J., 2024. Discharge Me: BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation. object Object]. doi, 10.

[51] Fleming, S.L., Lozano, A., Haberkorn, W.J., Jindal, J.A., Reis, E.P., Thapa, R., Blankemeier, L., Genkins, J.Z., Steinberg, E., Nayak, A. and Patel, B.S., 2023. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. arXiv [cs. CL].

[52] Yim, W.W., Fu, Y., Ben Abacha, A., Snider, N., Lin, T. and Yetisgen, M., 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Scientific data, 10(1), p.586.

[53] Chen, Z., Varma, M., Wan, X., Langlotz, C. and Delbrouck, J.B., 2023, July. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 469-484).

[54] Aali, A., Van Veen, D., Arefeen, Y.I., Hom, J., Bluethgen, C., Reis, E.P., Gatidis, S., Clifford, N., Daws, J., Tehrani, A.S. and Kim, J., 2025. A dataset and benchmark for hospital course summarization with adapted large language models. Journal of the American Medical Informatics Association, 32(3), pp.470-479.

[55] Zeng, G., Yang, W. and Ju, Z., 2020. “MedDialog: Large-Scale Medical Dialogue Datasets.” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9241–9250.

[56] Rabbani, N., Brown, C., Bedgood, M., Goldstein, R.L., Carlson, J.L., Pageler, N.M. and Morse, K.E., 2024. Evaluation of a large language model to identify confidential content in adolescent encounter notes. JAMA pediatrics, 178(3), pp.308-310.

[57] Tse, G., Zahedivash, A., Anoshiravani, A., Carlson, J., Haberkorn, W. and Morse, K.E., 2025. Large Language Model Responses to Adolescent Patient and Proxy Messages. JAMA pediatrics, 179(1), pp.93-94.

[58] Tse, G., Zahedivash, A., Anoshiravani, A., Carlson, J., Haberkorn, W. and Morse, K.E., 2025. Large Language Model Responses to Adolescent Patient and Proxy Messages. JAMA pediatrics, 179(1), pp.93-94.

[59] Pandit, S., Xu, J., Hong, J., Wang, Z., Chen, T., Xu, K. and Ding, Y., 2025. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. arXiv preprint arXiv:2502.14302.

[60] Lee, G., Hwang, H., Bae, S., Kwon, Y., Shin, W., Yang, S., Seo, M., Kim, J.Y. and Choi, E., 2022. Ehrsql: A practical text-to-sql benchmark for electronic health records. Advances in Neural Information Processing Systems, 35, pp.15589-15601.