Juno AI: A Retrieval-Grounded Conversational Platform for Equitable Healthcare Information Access and Navigation

Jay Kapoor University of Pittsburgh, Pittsburgh, PA, USA · Juno AI *Corresponding author*

Abstract

Access to trustworthy health information remains uneven, particularly for people in low-resource and low-literacy settings who must interpret symptoms, judge urgency, and locate appropriate care with little guidance. General-purpose search engines and unmoderated forums often compound this difficulty, surfacing alarming or contradictory material that increases anxiety without improving understanding. This paper presents Juno AI, a conversational platform designed to make health information clearer and the path to care easier to follow. Juno combines a large language model with retrieval-augmented generation over a curated corpus of authoritative sources, so that every response is grounded in vetted content rather than the model's unconstrained recall. A layered safety system classifies user intent, screens for emergency indicators, and constrains the assistant to an informational, explicitly non-diagnostic role, escalating to in-person or emergency care when warranted. A lightweight recommendation component maps a user's stated concern to plain-language explanations and to the most appropriate next step in the care pathway. We describe the system's design philosophy, architecture, and interaction model, and report a formative evaluation combining heuristic review with a small qualitative user study. Participants found the assistant easier to understand and less alarming than conventional symptom search, and reported greater clarity about what to do next. We discuss accessibility implications, deployment constraints in bandwidth- and device-limited contexts, and the ethical commitments—safety, transparency, privacy, and equity—that a consumer-facing health tool must uphold. Juno is positioned not as a replacement for clinicians but as an accessible first point of orientation that complements existing care infrastructure.

Keywords: conversational AI, healthcare accessibility, retrieval-augmented generation, digital health, health navigation, large language models, human-centered design

---

1. Introduction

1.1 Background

Health systems increasingly assume that patients can find, read, and act on health information on their own. Appointment booking, symptom checking, insurance navigation, and post-visit instructions are routinely delegated to self-service channels. For people with strong digital literacy and reliable connectivity, this shift can be convenient. For everyone else, it introduces a new barrier between a person and the care they need: the burden of interpretation. The first encounter a person has with the health system is now frequently not a clinician but a search box.

Artificial intelligence has advanced rapidly in clinical and consumer health over the past decade, from image-based diagnostic models to general-purpose language systems capable of fluent medical conversation [1], [2]. Topol characterized this period as a convergence of human and machine intelligence in medicine, with the potential to restore time and attention to care rather than displace it [1]. Yet most high-profile applications target clinicians and institutions. Comparatively little attention has been paid to the moment before care—when a worried person, often without medical training, is simply trying to understand what is happening and what to do.

1.2 The healthcare accessibility problem

Healthcare accessibility is commonly framed in terms of physical access to facilities or affordability of services. An equally consequential but less visible barrier is *informational*: the gap between the language of medicine and the language of everyday life. This gap is widest precisely where care is hardest to reach—among people with limited formal education, those who speak languages underserved by digital tools, and those for whom a single clinical visit involves significant time and cost.

The default tools available to fill this gap perform poorly against this standard. General search engines optimize for relevance and engagement rather than reassurance or correctness, and they readily surface the most severe possible explanation for a benign symptom. Online forums offer empathy but inconsistent accuracy. The result is a familiar pattern in which a routine query escalates into unwarranted alarm, while the practical question—*is this something I should worry about, and where should I go?*—goes unanswered. Mobile and digital health interventions have shown that well-designed tools can change health behavior at scale [3], but the design of the everyday information layer has lagged behind the sophistication of the underlying models.

1.3 Motivation

Juno AI began from a simple observation drawn from work in community and clinical settings: people do not primarily need a diagnosis from a machine. They need orientation. They need to know whether a symptom is urgent, what it likely means in plain terms, and what concrete step to take next. These are tractable problems for a carefully constrained conversational system, and they are problems that conventional search handles badly.

The motivation for Juno is therefore narrower and, we argue, more responsible than the ambition of automated diagnosis. The system is designed to be an accessible, calm, and trustworthy first point of contact that explains, reassures where appropriate, flags genuine warning signs, and directs people toward the right kind of care. This paper describes the platform's design, the engineering choices that make it safe and grounded, and a formative evaluation of an advanced prototype.

---

2. Literature Review

2.1 AI in healthcare

The modern wave of medical AI was catalyzed by deep learning results that matched specialist performance on narrow perceptual tasks, most visibly in dermatology and radiology [2], [4]. Rajkomar and colleagues offered an influential framing of machine learning's role across the clinical pipeline, distinguishing tasks where automation is appropriate from those where human judgment must remain central [5]. The transformer architecture [6] and the language models built on it [7], [8] subsequently extended these capabilities from images to text, enabling systems that can read clinical notes, answer medical questions, and hold extended conversations.

Recent work has shown that large language models can encode a surprising amount of clinical knowledge and perform competitively on medical question-answering benchmarks [9]. These results are promising but come with an important caveat: fluency is not the same as reliability. The same models are prone to *hallucination*—the generation of plausible but unsupported statements—which is acceptable in low-stakes settings and dangerous in health [10]. This tension between capability and trustworthiness is the central design problem that any consumer health AI must confront.

2.2 Digital health platforms

The broader field of e-health, articulated early by Eysenbach as the intersection of medical informatics, public health, and the internet [11], established that technology could expand access as well as deliver care. Mobile health interventions later demonstrated measurable effects on behavior change and adherence across diverse contexts, including low-resource settings [3]. These platforms succeeded when they were designed around real constraints —intermittent connectivity, shared and low-end devices, and varied literacy—rather than assuming the conditions of a well-resourced user.

The lesson for an information-access tool is that accessibility is a property of the whole system, not only of the model. A technically capable assistant that requires a fast connection, a modern smartphone, and fluent reading in a dominant language reproduces the very inequities it claims to address.

2.3 Conversational health systems

Conversational agents have a longer history in health than the current language-model era suggests. Laranjo and colleagues' systematic review documented a range of dialogue systems for health and noted that most were narrowly scripted and rarely evaluated for safety or real-world impact [12]. Relational and embodied agents were shown to support behavior change and engagement, particularly when designed to build rapport over time [13]. In mental health, Woebot demonstrated that a structured conversational agent could deliver evidence-based techniques and sustain engagement among young adults [14].

These successes were matched by sober assessments of the risks. Powell argued that conversational AI in health frequently fails an implicit trust test, projecting more competence than it possesses [15]. Nadarzynski and colleagues found that public acceptability of health chatbots hinged on accuracy, transparency, and the perception that the tool would not overstep its role [16]. Taken together, the literature suggests a clear design brief: conversational health systems are valued when they are accurate, honest about their limits, and explicitly bounded—and distrusted when they imitate a clinician they cannot replace.

---

3. Research Gap

The systems surveyed above tend to fall into one of two camps, and the space between them is where Juno is positioned.

The first camp comprises **clinically oriented or diagnostic systems**. These are powerful but typically target trained users, require structured input, and carry regulatory and liability burdens that make them ill-suited to casual, first-contact use by the general public. They answer the question *what is the diagnosis?*—a question that is often the wrong one to ask a machine.

The second camp comprises **general-purpose conversational models and search engines**. These are widely accessible but ungrounded. They will answer almost any health question with confidence, drawing on unverified recall, and they offer no built-in mechanism to detect emergencies, to refuse out-of-scope requests, or to point a user toward appropriate care. Their fluency makes their errors more persuasive, not less.

What is missing is a system designed for the specific, common, and underserved task of **health orientation**: helping a non-expert understand a concern in plain language, judge its urgency safely, and identify the right next step—while remaining transparently non-diagnostic and grounded in verified sources. Such a system must also treat accessibility as a first-order constraint, functioning for users with limited literacy, limited connectivity, and limited devices. The remainder of this paper describes a platform built explicitly to fill this gap.

---

4. Proposed System — Juno AI

4.1 Product overview

Juno AI is a conversational platform that helps people understand health information and find their way to appropriate care. A user describes a concern in their own words; Juno responds with a plain-language explanation grounded in vetted sources, an honest sense of how urgent the situation appears, and a concrete suggestion for what to do next. The defining commitment of the system is restraint: Juno explains and orients, but it does not diagnose, prescribe, or impersonate a clinician.

4.2 Intended users

The primary users are members of the general public seeking to make sense of a symptom, a diagnosis they have already received, or a step in the care process such as preparing for a visit. Particular attention is paid to users in low-resource settings—those with limited health literacy, those who are more comfortable in a regional language than in English, and those accessing the service on modest devices and connections. A secondary group consists of community health workers and frontline staff, who may use Juno as a plain-language reference while supporting others. Designing for the most constrained user first tends to produce a tool that serves everyone better.

4.3 Core features

Juno's functionality is organized around the orientation task rather than around an attempt at clinical breadth.

Feature	Purpose	Accessibility consideration
Plain-language explanation	Translate medical concepts into everyday terms	Calibrated reading level; avoids jargon
Urgency	Distinguish routine concerns	Conservative defaults; escalates rather

awareness	from warning signs	than reassures when uncertain
Care-pathway navigation	Suggest the appropriate type and setting of care	Accounts for locally available options
Source-grounded answers	Anchor responses to authoritative content	Citations shown so users can verify
Conversational memory	Maintain context within a session	Reduces repetitive re-entry of information
Multilingual interaction	Serve users in their preferred language	Lowers the literacy and language barrier

4.4 Workflow

A typical interaction proceeds in four stages. The user states a concern conversationally. Juno interprets the intent and retrieves relevant grounded material. It composes a response that explains the concern, characterizes its urgency, and proposes a next step, with sources attached. Finally, the conversation continues as needed, with Juno maintaining context and remaining within its informational role throughout. At any point, if the input suggests a medical emergency, the workflow short-circuits and Juno directs the user immediately toward emergency care rather than continuing an informational exchange.

---

5. Methodology

5.1 System design

Juno was developed through iterative, human-centered design. The guiding principle was *clarity over completeness*: a short, correct, and calming answer that a worried person can act on is more valuable than an exhaustive one they cannot parse. Three design rules followed from this principle. First, the assistant adopts a calm, non-alarming register, since the emotional tone of a health answer materially affects how it is received. Second, it errs toward caution, escalating to “seek care” guidance when a situation is ambiguous rather than offering false reassurance. Third, it is transparent about being an information tool and not a clinician, stating this boundary plainly and repeatedly where relevant.

5.2 Model selection rationale

A central decision was how to obtain conversational competence without sacrificing safety or grounding. Three options were considered: training a domain model from scratch, fine-tuning an open model on medical data, and using a capable instruction-tuned language model in combination with retrieval-augmented generation (RAG). Training from scratch was infeasible given data and compute constraints and would not, on its own, solve the reliability problem. Fine-tuning alone risks baking in errors and still leaves the model free to generate unsupported claims at inference time.

We selected the third approach: a capable instruction-tuned model whose outputs are grounded at inference time by retrieving relevant passages from a curated corpus and conditioning generation on them [17]. RAG directly addresses the hallucination problem identified in the literature [10]: rather than relying on the model's parametric memory, Juno answers from vetted source material it can cite. This choice also makes the system's knowledge auditable and updatable—correcting or extending Juno's coverage is a matter of editing the corpus rather than retraining a model. The model itself is treated as a replaceable component behind a stable interface, so the platform can adopt improved models as they become available without redesign.

5.3 User interaction flow

Each message passes through a fixed sequence. An intent and safety classifier first inspects the input to determine what the user is asking and to screen for emergency indicators and out-of-scope requests such as appeals for a definitive diagnosis or a prescription. Inputs flagged as emergencies are routed immediately to escalation messaging. Otherwise, the user's query is used to retrieve relevant passages from the knowledge base, which are supplied to the language model along with a system instruction that fixes Juno's role, tone, and boundaries. The generated response is then checked by a post-generation safety filter before it is returned with its supporting sources. This pipeline ensures that grounding and safety are structural properties of every response rather than behaviors the model is merely asked to exhibit.

5.4 Data considerations

Juno's knowledge base is assembled from authoritative public-health and clinical-reference sources and is curated rather than scraped indiscriminately, so that grounded answers inherit the credibility of vetted material. On the user side, the system follows data minimization: it is designed to function without accounts or persistent health profiles, conversational context is scoped to the session, and sensitive content is not retained beyond what an interaction requires. These choices reflect both a privacy commitment and a practical reality—users disclose more, and more honestly, to a tool they trust not to record them.

6. System Architecture

Juno is built as a layered system in which each layer has a single, well-defined responsibility. This separation keeps the safety-critical components small and inspectable and allows the AI components to be upgraded independently.

Figure 1. Juno AI system architecture (described). *The figure depicts four horizontal layers connected by a single request–response flow. At the top, the* **client layer** *(a responsive, mobile-first web application) captures the user's message and renders Juno's reply with its cited sources. The message travels to the* **backend service layer**, *an API server that manages sessions, conversational context, and orchestration. From there the request enters the* **AI layer**, *shown as a left-to-right pipeline: an intent-and-safety classifier, a retriever that queries a vector index built over the curated knowledge corpus, the language model that generates a grounded response, and a post-generation safety filter. A branch from the classifier bypasses the pipeline and returns escalation guidance directly when an emergency is detected. Running alongside the AI layer is the* **recommendation component**, *which maps the interpreted concern to the appropriate explanation and care-pathway suggestion. Arrows return the validated, source-attributed response back up through the backend to the client. A persistent* **knowledge base** *and* **vector index** *sit beneath the AI layer as shared, read-mostly resources.*

A simplified schematic of the flow is shown below.

```
User ── message ──▶  Client (responsive web app)
                              |
                              ▼
                     Backend API service
               (sessions, context, orchestration)
                              |
                              ▼
      ┌─────────── AI LAYER ──────────┐
      | intent + safety classifier    |──▶ emergency? ──▶ escalation
      |               |                |
      |               ▼                |
      |     retriever ◀── vector index ◀── knowledge base
      |               |                |
      |               ▼                |
      |   grounded LLM generation      |
      |               |                |
      |               ▼                |
      |   post-generation safety filter |
      └───────────────┬────────────────┘
                      |   (+ recommendation: explanation + care pathway)
                      ▼
```

```
response with citations ──▶ Backend ──▶ Client ──▶ User
```

6.1 Frontend

The client is a responsive, mobile-first web application built with a standard component framework and designed to degrade gracefully on low-end devices and slow connections. The interface is deliberately spare: a conversational view with large, legible type, generous spacing, and clearly displayed source citations. Accessibility features—adjustable text size, support for voice input, and language selection—are treated as core functionality rather than add-ons, since the intended users depend on them.

6.2 Backend

The backend is a stateless API service that mediates between the client and the AI layer. It manages session context, enforces request validation and rate limiting, and orchestrates the steps of the response pipeline. Keeping the backend thin and stateless simplifies horizontal scaling and concentrates health-relevant logic in the AI layer, where it can be reviewed in one place. Structured data such as the knowledge corpus and operational logs are stored in a conventional relational database, with a separate vector index serving retrieval.

6.3 AI layer

The AI layer implements the pipeline described in Section 5.3. Its components are intentionally modular: the classifier, retriever, generator, and safety filter communicate through defined interfaces so that any one can be replaced or strengthened without disturbing the others. Retrieval is performed by embedding the user's query and the corpus into a shared vector space and selecting the most relevant passages, which are then provided to the language model as grounding context [17]. The model is instructed, through a fixed system prompt, to answer only from the supplied material, to cite it, and to defer to professional care whenever appropriate. Because grounding and constraint are enforced structurally, the system's behavior does not depend solely on the goodwill of the underlying model.

6.4 Recommendation engine

The recommendation component is deliberately modest and hybrid rather than fully learned. A rule-based triage layer encodes conservative, well-established urgency heuristics—the kind of guidance that distinguishes a concern that warrants emergency care from one that can wait for a routine appointment. A content-matching layer then aligns the user's interpreted concern with the most relevant plain-language explanation and the appropriate setting of care. Favoring transparent rules over an opaque model in this safety-adjacent role makes the system's recommendations explainable and auditable, which matters more here than marginal gains in sophistication.

---

7. Evaluation

The prototype was evaluated formatively, with the goal of learning whether the core orientation task is served well and where the design needs to improve—not of establishing clinical efficacy, which would require controlled study beyond the scope of this work. Two complementary methods were used: expert heuristic review and a small qualitative user study.

7.1 Usability discussion

Heuristic evaluation followed established usability principles [18], [19], with reviewers examining the interface and a set of representative interactions for clarity, consistency, error prevention, and the visibility of Juno's limitations. The interface scored well on simplicity and legibility; the most valuable revisions concerned making the assistant's non-diagnostic role and its source citations more prominent, so that users could readily see both the boundary of the tool and the basis for its answers.

The qualitative user study involved a small, diverse group of participants who completed realistic tasks—interpreting a common symptom, making sense of a routine diagnosis, and preparing for a visit—and then reflected on the experience. Usability was probed using a standard perceived-usability instrument [18] alongside open-ended discussion. We report these findings qualitatively, in keeping with the study's formative aim and small sample.

Dimension	Method	Principal finding
Comprehensibility	Task + interview	Explanations were judged clearer and less jargon-laden than conventional search
Emotional tone	Interview	The calm register reduced alarm relative to symptom searching
Actionability	Task observation	Users left with a concrete sense of the next step
Trust	Interview	Visible citations and stated limits increased confidence
Perceived usability	SUS-style instrument	Favorable overall; navigation and tone praised

The most consistent theme was a contrast with the participants' prior experience of searching symptoms online: where search tended to amplify worry, Juno tended to settle it without dismissing genuine concern. Participants particularly valued knowing what to do next, a need that ordinary search rarely addresses.

7.2 Prototype validation

Beyond usability, validation focused on whether Juno's responses stayed within their intended bounds. Across the test interactions, the assistant reliably maintained its non-diagnostic stance, grounded its answers in retrievable sources, and escalated appropriately when prompts contained emergency indicators. The exercise also exposed the system's natural limits: where the curated corpus was thin, response quality fell, which reaffirmed that Juno's usefulness is a direct function of the breadth and quality of its knowledge base rather than of model fluency alone. These observations are encouraging for the orientation use case but should be read as formative evidence from a prototype, not as proof of effectiveness at scale.

7.3 Expected deployment considerations

Moving from prototype to field deployment introduces constraints that the architecture was designed to accommodate but that nonetheless require attention. Connectivity and device limitations argue for a lightweight client and tolerance of intermittent networks. Genuine multilingual support—not merely translated interface labels but grounded answers in regional languages—is essential to the platform's accessibility goals and is non-trivial to deliver well. Operating cost, dominated by model inference, must be managed for the service to remain free or low-cost to its intended users. Finally, the greatest practical leverage is likely to come from integration with existing care infrastructure, so that Juno orients users *into* the health system rather than standing apart from it.

---

8. Discussion

8.1 Accessibility impact

Juno's contribution is best understood as lowering the informational barrier to care. By translating medical concepts into plain language, offering answers in a user's preferred language, and converting a vague worry into a concrete next step, the platform addresses a form of inaccessibility that physical and financial access measures overlook. In doing so it also targets a specific harm of the status quo—the anxiety spiral of unguided symptom search—replacing it with a calmer, sourced, and actionable exchange. The aim throughout is augmentation rather than substitution: Juno is most valuable when it helps people arrive at human care better oriented, not when it tries to stand in for that care.

8.2 Limitations

Several limitations bound the claims made here. Juno is, by design, not a diagnostic system, and it should not be used as one. Retrieval-augmented generation reduces but does not eliminate the risk of incorrect or misleading output, so residual error remains an inherent risk of any language-model-based tool [10]. The platform's quality is capped by the coverage of its curated corpus and by the languages it genuinely supports; gaps in either translate directly into gaps in usefulness for exactly the users the project most wants to serve. The digital divide imposes a further limit: a tool delivered through a smartphone cannot, on its own, reach those without one. And the evaluation reported here is formative and small in scale, establishing promise rather than proof.

8.3 Ethical considerations

A consumer-facing health tool carries obligations beyond functioning correctly, and these shaped the design as much as any technical requirement. The overriding commitment is to avoid harm: conservative escalation, an explicit non-diagnostic boundary, and refusal of out-of-scope requests are all expressions of this commitment, consistent with emerging guidance on the governance of AI for health [20]. Transparency is treated as a safety feature—users are shown the sources behind an answer and told plainly what the tool is and is not. Privacy is protected through data minimization, so that seeking health information does not become a record to be exploited. Equity is the project's animating concern, but it also imposes a discipline: a tool built in the name of access must be evaluated by whether it actually reaches and serves the least-resourced users, and not only by its performance for those already well served. Finally, the design resists fostering over-reliance, steering users toward professional care rather than positioning itself as a terminus.

---

9. Future Work

Several directions would extend Juno from a validated prototype toward responsible deployment. The most important is rigorous, larger-scale evaluation with the intended populations, including clinical review of response quality and study of real-world impact on understanding and care-seeking. Expanding genuine multilingual and voice-based interaction would broaden reach to low-literacy users for whom text alone is a barrier. Integration with community health infrastructure—equipping frontline workers with Juno as a plain-language reference, and connecting users onward to telemedicine and local services—would convert orientation into completed care pathways. On the technical side, continued work on grounding and verification, including stronger detection of unsupported claims, would further narrow the residual error that any language-model system carries. Each of these steps is incremental and concrete, in keeping with the project's preference for measured progress over sweeping claims.

---

10. Conclusion

Juno AI addresses a specific and underserved problem: helping people who are not medical experts understand a health concern, gauge its urgency safely, and find the right next step. By pairing a capable language model with retrieval-augmented grounding and a structural safety pipeline, the platform offers fluent, source-anchored, and explicitly non-diagnostic guidance, with accessibility treated as a first-order design constraint rather than an afterthought. A formative evaluation suggests that this approach can make health information clearer and less alarming than conventional search while leaving users better oriented toward care. The contribution is intentionally bounded. Juno does not diagnose, and it does not replace clinicians; it lowers the barrier between a worried person and the care they need. Realizing that promise at scale will require broader evaluation, deeper language support, and integration with existing health systems—work that the architecture described here is designed to support.

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[2] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] C. Free, G. Phillips, L. Galli, et al., "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review," *PLoS Medicine*, vol. 10, no. 1, e1001362, 2013.

[4] A. Esteva, A. Robicquet, B. Ramsundar, et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[5] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[8] T. B. Brown, B. Mann, N. Ryder, et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] K. Singhal, S. Azizi, T. Tu, et al., “Large language models encode clinical knowledge,” *Nature*, vol. 620, pp. 172–180, 2023.

[10] Z. Ji, N. Lee, R. Frieske, et al., “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] G. Eysenbach, “What is e-health?,” *Journal of Medical Internet Research*, vol. 3, no. 2, e20, 2001.

[12] L. Laranjo, A. G. Dunn, H. L. Tong, et al., “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[13] T. W. Bickmore and R. W. Picard, “Establishing and maintaining long-term human–computer relationships,” *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 2, pp. 293–327, 2005.

[14] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial,” *JMIR Mental Health*, vol. 4, no. 2, e19, 2017.

[15] J. Powell, “Trust me, I’m a chatbot: how artificial intelligence in health care fails the Turing test,” *Journal of Medical Internet Research*, vol. 21, no. 10, e16222, 2019.

[16] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, “Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study,” *Digital Health*, vol. 5, pp. 1–12, 2019.

[17] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] J. Brooke, “SUS: A ‘quick and dirty’ usability scale,” in *Usability Evaluation in Industry*, P. W. Jordan et al., Eds. London: Taylor & Francis, 1996, pp. 189–194.

[19] J. Nielsen, “Heuristic evaluation,” in *Usability Inspection Methods*, J. Nielsen and R. L. Mack, Eds. New York: John Wiley & Sons, 1994.

[20] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: World Health Organization, 2021.