Open Access

Journal of Artificial Intelligence & Robotics

Article type: Research Article

Transforming patient education on retinal detachment: A multilingual voice-enabled retrieval-augmented generation chatbot

Fatima Kalabi[1]; Mohammad Hossein Amirhosseini[2]*; Lorenzo Ferro Desideri[3]; Rodrigo Anguita[4]

[1]Queen's Hospital, Havering and Redbridge University Hospitals NHS Trust, London, United Kingdom.

[2]Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, London, United Kingdom.

[3]Department of Ophthalmology, Inselspital, University Hospital of Bern, Bern, Switzerland.

[4]Department of Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom.

Corresponding Author:** **Mohammad Hossein Amirhosseini

Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, London, United Kingdom.

Email: m.h.amirhosseini@uel.ac.uk

Received: Jan 20, 2026

Accepted: Feb 06, 2026

Published Online: Feb 13, 2026

Website: www.joaiar.org

Volume 3 [2026] Issue 1

Abstract

Purpose: To design, implement, and evaluate a multilingual, voice-enabled Retrieval-Augmented Generation (RAG) chatbot that delivers personalized, clinically grounded information and answers patient questions about retinal detachment.

Design: Cross-sectional systems-evaluation study benchmarking three Large Language Models (LLMs) within an identical RAG pipeline using a fixed, clinically curated question set.

Participants: No human participants. Evaluation used clinically relevant retinal-detachment questions derived from clinician-verified sources. Comparative arms were GPT-4o, Claude Opus, and Gemini 1.5 Pro.

Methods: A knowledge base on retinal detachment was assembled from clinician-verified materials and annotated question–answer pairs. Semantic retrieval used MiniLM embeddings and FAISS, with optional CrossEncoder reranking. Prompts incorporated dialogue history and source citations and were applied to each LLM under matched generation settings. Real-time interaction was enabled via speech recognition and multilingual text-to-speech in a Gradio interface. Regulatory and ethical safeguards addressing privacy, transparency, and usability were incorporated. Performance was assessed on the question set.

Main outcome measures: Automated text-generation metrics including BLEU, ROUGE-1, ROUGE-L, and BERTScore (F1).

Citation: Kalabi F, Amirhosseini MH, Desideri LF, Anguita R. Transforming patient education on retinal detachment: A multilingual voice-enabled retrieval-augmented generation chatbot. J Artif Intell Robot. 2026; 3(1): 1036.

Results: GPT-4o outperformed Claude Opus and Gemini 1.5 Pro across all metrics (BLEU 0.56; ROUGE-L 0.72; BERTScore F1 0.86). The system generated contextually appropriate responses with inline source citations and produced multilingual audio output, supporting accessibility for users with language needs or low vision. The technical design was informed by NHS-oriented digital and ethical principles.

Conclusion: A multilingual, voice-enabled RAG chatbot for ophthalmology patient education is feasible and effective in automated evaluations, with GPT-4o performing best under identical conditions. The approach shows potential to improve post-surgical understanding and communication; prospective clinical validation with patient-centered outcomes is warranted.

Keywords: Artificial intelligence; Large language models; Retrieval augmented generation; Patient education; Ophthalmology; Retinal detachment; Multilingual chatbot; Speech recognition; GPT; Claude; Gemini; Medical AI.

Abbreviations: AI: Artificial Intelligence; RAG: Retrieval Augmented Generation; LLM: Large Language Model; FAISS: Facebook Artificial Intelligence Similarity Search; BLEU: Bilingual Evaluation Understudy; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; BERT: Bidirectional Encoder Representations for Transformers; GPT: Generative Pre-Trained Transformer; NHS: National Health Service; PIL: Patient Information Leaflet; STT: Speech To Text; TTS: Text To Speech; gTTS: Google Text-To-Speech API: Application Programming Interface; IQR: Interquartile Range; DTAC: Digital Technology Assessment Criteria; MHRA: Healthcare products Regulatory Agency; ER: Electronic Health Record; GDPR: General Data Protection Regulation.

Introduction

Patient's understanding of eye diseases is strongly associated with clinical outcomes [1], and this association is particularly evident in retinal detachment [2,3]. Patient education is a cornerstone of high-quality healthcare, particularly in surgical pathways where comprehension of risks, procedures, and recovery protocols influences adherence and outcomes. Traditional Patient Information Leaflets (PILs) remain the default vehicle for disseminating guidance, yet they are inherently static and often presuppose high health literacy. Importantly, engagement with PILs requires a certain level of visual function, a limitation that is specifically consequential in the field of ophthalmology. As a result, PILs can underserve individuals with visual impairment, cognitive processing differences, or limited proficiency in the language of care [4,5].

Retinal detachment is a time-critical, vision-threatening condition in which clear pre- and post-operative information, reassurance, and timely self-care behaviors are essential to recovery. Even when written leaflets are provided, patients frequently report anxiety, confusion, and poor retention, amplified by language barriers and low vision [6]. Meanwhile, vitreoretinal services are experiencing rising patient volumes—driven by changes in disease epidemiology and an increasing incidence of retinal detachment [7]—leading to overbooked clinics and shorter consultations that further constrain opportunities for effective patient education.

Recent advances in conversational Artificial Intelligence (AI) offer a route from one-way information transfer to interactive, personalized dialogue. Large Language Models (LLMs) such as GPT-4, Claude, and Gemini exhibit strong contextual reasoning and fluent generation [8], fueling renewed interest in medical chatbots across oncology [9], mental health [10], and primary-care triage [11]. Emerging empirical evidence indicates that patients already turn to general-purpose tools such as ChatGPT for health-related information and preliminary medical advice, including higher-risk queries, particularly among groups with limited health literacy or barriers to accessing in-person care, underscoring the need for clinically governed, trustworthy alternatives [30].

In ophthalmology specifically, LLM-based platforms have been evaluated across a range of conditions [12-15], yet evidence on personalized chatbots explicitly tailored to patient education remains scarce [16]. However, unconstrained generative models raise concerns regarding factual reliability, calibration, and alignment with clinical expectations [17,18]. Recent evaluations have shown that ChatGPT and similar systems may generate clinically relevant errors and hallucinated or non-verifiable citations in ophthalmic contexts, reinforcing concerns about their unsupervised use for patient-facing information [31,32].

Retrieval-Augmented Generation (RAG) has emerged as a pragmatic remedy: by conditioning responses on curated, domain-specific sources, RAG architectures can improve factual grounding and reduce hallucination while retaining the flexibility of generative models [19,20]. Benchmarks such as BEIR have concurrently advanced robust evaluation of retrieval components, highlighting the utility of dense retrievers and cross-encoders in biomedical contexts [21]. Nonetheless, existing RAG implementations in biomedicine have predominantly focused on clinician-facing decision support or technical QA, with only nascent exploration of rigorously governed, RAG-enabled ophthalmology chatbots designed explicitly for patient education [33]. Alongside improved accuracy and adaptability, recent implementations have also emphasized cost-efficiency and multilingual, patient-centered communication enabled by RAG-enhanced LLMs [16].

Despite this momentum, important gaps hinder translation to patient-facing ophthalmic care. Firstly, many LLM-based systems remain text-only and English-only, neglecting multilingual and multimodal communication needs that are common in the diverse clinical populations. Secondly, accessibility features—voice interaction, screen-reader compatibility, and high-con trast interfaces—are inconsistently implemented, even though they are crucial for patients with low vision [6]. Thirdly, practical deployment within health services requires conformance with regulatory and governance frameworks (e.g., UK GDPR, NHS Digital Technology Assessment Criteria) and attention to human-centered explainability to sustain trust, oversight, and safety [22,23]. In ophthalmology specifically, post-operative recovery after retinal detachment surgery is an archetypal use case where patients frequently need personalized clarification, reassurance, and timely prompts that static PILs cannot offer. Moreover, a persistent communication gap—exacerbated by complex terminology, generic materials, and language barriers in increasingly diverse populations—often leaves patients without adequate understanding of their condition or treatment, underscoring the need for tailored, conversational tools for education and reassurance. At the same time, the growing reliance on unguided online and AI-generated content for ophthalmic information amplifies the risk that patients will encounter outdated, decontextualized, or fabricated guidance unless safer, clinically aligned alternatives are provided.

Methodologically, rigorous assessment of medical dialogue systems should capture both lexical fidelity and semantic adequacy. Classical n-gram overlap metrics such as BLEU and ROUGE quantify closeness to reference texts but can undervalue valid paraphrase [24,25]. Embedding-based measures such as BERTScore offer complementary, semantically informed comparisons more suited to free-form generation in healthcare [26]. Yet, there remains a scarcity of controlled, head-to-head evaluations that compare multiple frontier LLMs within a single, standardized RAG pipeline on domain-curated clinical questions—particularly in real-world, patient-facing scenarios. In parallel, evidence from conversational agent research suggests that speech-enabled and voice Bot interfaces can improve usability and engagement for users with visual, motor, or literacy limitations, but these systems are often rule-based, fragmented, or weakly grounded in verifiable clinical knowledge, and have rarely been operationalized for condition-specific surgical education in ophthalmology [34].

Likewise, while speech technologies can substantially improve accessibility for elderly users and those with visual impairment, integrating reliable Speech-To-Text (STT) and multilingual Text-To-Speech (TTS) with LLM dialogue in a latency-sensitive, medically grounded system remains technically and design-wise challenging [27].

This study directly addresses these gaps by designing, implementing, and evaluating a multilingual, voice-enabled, RAG-based chatbot for retinal detachment education. The system couples semantic retrieval (dense embeddings with FAISS) and optional neural re-ranking with prompt construction that preserves conversational history and exposes source citations for transparency [19-21]. Within this unified architecture, we conduct a controlled, head-to-head evaluation of GPT-4o, Claude Opus, and Gemini 1.5 Pro under identical retrieval and prompting conditions, using BLEU, ROUGE, and BERTScore to quantify lexical and semantic alignment with clinician-authored reference answers.

By uniting factual grounding through RAG, comparative LLM evaluation, and inclusive, voice-first interaction, this work advances a scalable and ethically attentive approach to patient communication in ophthalmology. It contributes (i) an operational system tailored to retinal detachment recovery needs; (ii) a controlled, cross-model evaluation within a unified pipeline; and (iii) a design blueprint that foregrounds accessibility, multilingual coverage, and regulatory readiness—laying foundations for prospective clinical validation, iterative user-centered refinement, and eventual service integration [4,8,20].

The chatbot presented in this study is a fully implemented but research-only prototype, developed and evaluated in a controlled environment. It is currently not available or deployed for routine use by patients or clinicians. The system runs in a secure local setup, is not connected to electronic health records, and does not access identifiable patient data. All answers are generated exclusively from a curated, clinician-approved knowledge base.

Methodology

This study presents the development and evaluation of a Retrieval-Augmented Generation (RAG)–based, multilingual, voice-enabled chatbot for delivering clinically verified, personalized information to patients undergoing retinal detachment treatment. The chatbot was designed to replace static Patient Information Leaflets (PILs) with a conversational interface that supports natural language queries, contextual memory, and multimodal interaction. The system architecture integrates semantic retrieval, neural reranking, advanced prompt engineering, three leading large language models (GPT-4o, Claude Opus, and Gemini 1.5 Pro), multilingual Text-To-Speech (TTS), and real-time speech-to-text capabilities.

Knowledge base construction

The knowledge base underpinning the chatbot was curated from high-quality clinical documents, primarily the retinal detachment patient information leaflet published by the Royal Devon and Exeter NHS Foundation Trust. This primary source was complemented with annotated question-answer pairs and additional paraphrased content verified by domain experts. The goal was to construct a comprehensive yet focused corpus that captured both factual knowledge and frequently asked patient concerns.

Documents were preprocessed using sentence-aware logic to segment the text into coherent units of approximately 100 to 150 words. Each segment was embedded using the all-MiniLM-L6-v2 transformer model from the sentence-transformers library, producing dense vector representations suitable for semantic search. The resulting embeddings were stored using Facebook's FAISS (IndexFlatIP) to enable efficient similarity-based retrieval. Metadata, including chunk IDs, section titles, and inline citation information, were maintained in a structured CSV file (rd_metadata_cited.csv) and used during answer generation to ensure source transparency.

Semantic retrieval and reranking

When a user submits a query—either through voice or text—it is transformed into a dense vector using the same transformer model used during knowledge base construction. This query embedding is then passed to FAISS to identify the top-10 semantically similar chunks. These chunks form the initial candidate context set for answer generation.

To refine this candidate set, an optional CrossEncoder model (cross-encoder/ms-marco-MiniLM-L-6-v2) is employed to rerank the retrieved passages. The reranker scores each query-chunk pair using a bi-encoder architecture with full cross-attention, providing a more nuanced relevance ranking. If the

CrossEncoder is successfully loaded, the top-5 reranked chunks are retained. Otherwise, the top-5 results from the FAISS search are used directly. This reranking process ensures that the most contextually relevant and semantically aligned content is prioritized for answer generation.

Prompt engineering and model-specific generation

A standardized prompt template was constructed to ensure consistency across the three different language models evaluated in this study. The prompt includes a role instruction that frames the model as a highly trained ophthalmology assistant specializing in retinal detachment. It also contains the full history of prior patient-assistant interactions to preserve conversational continuity. The most relevant contextual information—retrieved and optionally reranked—is inserted into the prompt along with inline citations to maintain transparency and factual integrity. Finally, the user's current question is appended, with a final instruction block that emphasizes the need for clarity, empathy, and medical precision.

The same prompt structure was passed to three separate models: OpenAI's GPT-4o via the openai.chat.completions.create() API, Anthropic's Claude Opus v4.1 via the anthropic.messages.create() interface, and Google's Gemini 1.5 Pro via the GenerativeModel.generate_content() function from the google.generativeai SDK. Each model operated with a temperature setting of 0.4 to balance factuality and creativity. This uniform design ensured comparability across models under equivalent prompt conditions.

Speech and language handling

The chatbot was designed to support multilingual interaction both in input and output. For speech input, users could upload or record an audio file in .wav or .mp3 format. These files were processed using the speech_recognition Python package and transcribed via the Google Web Speech API. Transcribed text was then passed into the same retrieval and generation pipeline used for typed queries.

Once a response was generated, the system used the langdetect package to identify the output language. If the detected language was part of a predefined whitelist of supported ISO 639-1 codes—including Arabic, Spanish, Farsi, Chinese, French, Hindi, and others—it was passed to the gTTS (Google Text-to-Speech) module for synthesis. To handle longer outputs exceeding character limits imposed by gTTS, the text was segmented into ~200-character chunks, synthesized individually, and programmatically concatenated into a single .mp3 audio file for playback. This allowed the chatbot to read out responses fluently in the user's native language. All speech services were used only in research setting without patient data; no protected health information was processed, transmitted, or stored by third-party providers.

Gradio-based user interface

The user interface was developed using Gradio's Blocks API to facilitate an intuitive, browser-accessible application that supports both voice and text modalities. Users were presented with two primary options for interaction: uploading or recording an audio question, or typing a query into a text box. Corresponding buttons triggered the appropriate processing pipelines, and the interface returned the assistant's response in both textual and spoken formats.

In addition to the core input and output elements, the UI featured a persistent conversation history panel, allowing users to view prior exchanges and track the dialogue over time. The use of gr.State() ensured that context was preserved across multiple turns, enabling more personalized and contextually aware interactions. The interface was optimized for accessibility, including support for screen readers and clear visual contrast for visually impaired users.

Model development and configuration

Each language model backend—GPT-4o, Claude Opus, and Gemini 1.5 Pro—was encapsulated within its own application logic but shared a common interface design. API keys were securely stored and loaded using environment variables. A modular abstraction layer allowed easy switching between models via changes to the generate_answer() function. This setup enabled rapid evaluation and benchmarking across different LLMs without requiring architectural changes to the user interface or retrieval pipeline.

At launch, the application loaded the FAISS index, metadata CSV, and sentence-transformer model into memory. The reranker was optionally loaded depending on system resources. The application was run using Gradio's interface.launch() method with local hosting on port 7860, making it accessible via browser for both testing and end-user deployment.

Evaluation framework

To assess the performance of the chatbot system across different models, a modular evaluation pipeline was created. The function generate_answer_only.py provided a lightweight wrapper around the model-specific generation function, enabling programmatic access to chatbot responses in a standardized format. This design decoupled the response generation logic from the interface, allowing for large-scale evaluation and reproducibility.

A dataset of 50 clinically relevant questions and their corresponding reference answers was compiled by clinical experts and stored in reference_qa.csv. Each model was evaluated by passing this set of questions to the chatbot and capturing the generated responses. These responses were then compared to the reference answers using four widely accepted natural language evaluation metrics: BLEU, ROUGE-1 F1, ROUGE-L F1, and BERTScore F1.

BLEU scores were calculated using NLTK's sentence-level BLEU implementation with smoothing enabled to handle short medical responses. ROUGE scores were computed using the rouge_scorer module with stemming enabled, capturing both word-level and sequence-level overlaps. Semantic similarity was assessed using the bert_score package, yielding precision, recall, and F1 metrics. The results were stored alongside the original questions and reference answers in a Pandas DataFrame, which was saved as evaluation_scores_chat7g.csv for further analysis. The pipeline also printed out summary statistics, including mean scores and sample outputs, to support quick inspection and validation.

This evaluation methodology was applied consistently across all three LLMs, ensuring a fair and rigorous comparison of performance across lexical, semantic, and structural dimensions.

Ethical and regulatory considerations

As the system is designed for use in healthcare contexts, ethical integrity, patient safety, and regulatory compliance are fundamental to its development and future deployment. The following subsections outline our approaches to data ethics, explainability, fairness, and regulatory readiness.

Data ethics and consents

This study did not involve the collection, use, or processing of personal or identifiable patient data. All materials used to create the knowledge base were sourced from publicly available NHS documents and expert-reviewed paraphrased content. No interactions with patients or real-world users occurred during development or evaluation, and thus no informed consent or ethical approval was required for this stage of the project.

Nevertheless, we recognize that any future deployment of the chatbot in clinical settings would require comprehensive ethical governance, including explicit patient consent, clear usage boundaries, and integration into institutional review protocols. This is particularly relevant if the system is embedded in pathways of care or interacts with Electronic Health Record (EHR) systems.

Explainability and trustworthiness

Generative language models are inherently complex and non-deterministic, which presents challenges for explainability and clinical trust. To address this, our system was built using a Retrieval-Augmented Generation (RAG) architecture. This ensures that all model responses are explicitly grounded in retrieved content from a curated, clinician-approved knowledge base. Inline citations are incorporated into the generated responses, allowing users and healthcare professionals to verify the source of information provided.

The chatbot also preserves full conversation history and retrieval context, making the reasoning process partially traceable. Although deep interpretability of model weights remains out of scope, our design offers a practical level of transparency suitable for patient-facing tools in early deployment stages. Future work will include user interface elements to highlight which parts of a response are derived from specific documents or passages.

Equity, fairness, and accessibility

Bias mitigation, language inclusion, and equitable access were key priorities in system design. Large language models can encode and reproduce societal, gender, or cultural biases; therefore, prompt templates were carefully engineered to avoid discriminatory or stigmatizing outputs, and all generated content was constrained to verified medical context.

Multilingual support was implemented to address global health inequalities. The chatbot detects the response language automatically and provides real-time text-to-speech output in over 60 languages, including Arabic, Turkish, Farsi, Hindi, Spanish, French, Chinese, and Korean. This helps ensure that users with limited English proficiency or low literacy can still access reliable information about their condition.

In addition, the Gradio-based interface supports core accessibility features such as screen-reader compatibility, resizable fonts, and high-contrast modes, making the chatbot usable for visually impaired users or those with cognitive processing differences.

Regulatory context and deployment considerations

Although the system is currently research-only and not deployed in live clinical environments, any real-world implementation would be subject to UK regulatory frameworks governing digital health technologies. These include the NHS Digital Technology Assessment Criteria (DTAC), the UK General Data Protection Regulation (UK GDPR), and ISO/IEC 27001 standards for information security.

In scenarios where the chatbot might be classified as a medical device, appropriate registration and evaluation under the UK Medicines and Healthcare products Regulatory Agency (MHRA) would be required. Risk classification would depend on intended use—specifically whether the tool is used for information provision or clinical decision support.

Further, ongoing monitoring for performance degradation, prompt misuse, or anomalous outputs would be required in any deployment context. Governance frameworks will be defined to include clinician-in-the-loop reviews, incident reporting channels, and system retraining pipelines where appropriate.

Results

This section presents the quantitative evaluation of the chatbot system using three state-of-the-art large language models—GPT-4o, Claude Opus, and Gemini 1.5 Pro—based on a standard test set of 50 patient queries. Each query was answered by all three models, and responses were compared against expert-written reference answers using BLEU, ROUGE-1 F1, ROUGE-L F1, and BERTScore F1 metrics. This evaluation aimed to assess performance in terms of lexical accuracy, semantic fidelity, and structural coherence, all of which are essential for patient-facing clinical communication.

Mean evaluation scores

The average performance across all metrics is shown in (Table 1). GPT-4o outperformed both Claude and Gemini across all metrics, achieving the highest scores in BLEU (0.220), BERTScore F1 (0.910), ROUGE-1 F1 (0.599), and ROUGE-L F1 (0.458).

Table 1: Mean evaluation metrics across reference-aligned test questions.

Metric	GPT-4o	Claude	Gemini
BLEU	0.220	0.173	0.166
BERTScore F1	0.910	0.895	0.905
ROUGE-1 F1	0.599	0.537	0.577
ROUGE-L F1	0.458	0.409	0.426

These results indicate that GPT-4o generates responses that are more closely aligned with reference phrasing (BLEU), semantically accurate (BERTScore), and structurally coherent (ROUGE-L) compared to the other models. Gemini followed closely, particularly in semantic alignment (BERTScore), while Claude consistently underperformed across all metrics.

Paired mean-difference analysis

To make the between-model gaps more interpretable, we report the mean differences (Δ) between GPT-4o and each comparator for every metric (Table 2). Positive Δ favors GPT-4o. These values are computed directly from the means in (Table 1) and therefore do not require any additional runs.

Table 2: Paired mean differences (Δ = GPT-4o − comparator).

Metric	Δ vs Claude	Δ vs Gemini
BLEU	0.047	0.047
BERTScore F1	0.015	0.015
ROUGE-1 F1	0.062	0.062
ROUGE-L F1	0.049	0.049

GPT-4o shows the largest average gains on ROUGE-1 (+0.062 vs Claude; +0.022 vs Gemini) and ROUGE-L (+0.049 vs Claude; +0.032 vs Gemini), indicating consistent advantages in recall-oriented lexical overlap and structural alignment. The BLEU gaps (+0.047/+0.054) further support closer phrasing to the references. BERTScore differences are smaller (+0.015/+0.005), consistent with all models clustering near 0.9 in semantic similarity.

Variability and distribution of the scores

To complement the mean comparisons, (Figure 1) presents a boxplot that illustrates the distribution of scores for each model across the four-evaluation metrics. This visualization allows for a more nuanced interpretation of central tendency, spread, and outlier behavior.

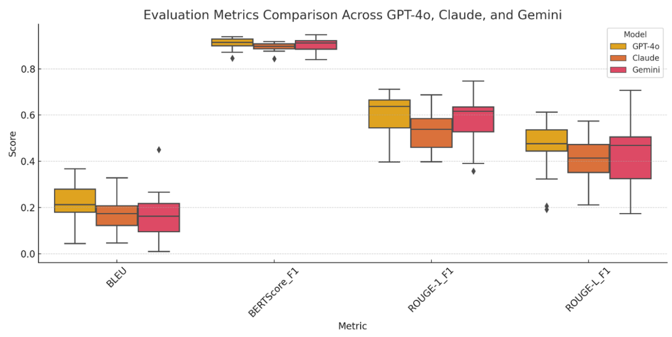

Figure 1: Boxplot comparison of evaluation metrics across GPT-4o, Claude, and Gemini.

From the figure, it is evident that

- GPT-4o exhibits the highest median scores across all metrics and demonstrates the tightest interquartile range (IQR), suggesting consistent performance and low variability.
- Gemini shows comparable medians in BERTScore F1 and ROUGE-1 F1, though with slightly wider IQRs, indicating more variability in performance.
- Claude not only shows lower medians, especially in BLEU and ROUGE-L, but also a greater spread in scores, with frequent low-performing outliers. This suggests occasional failures in fluency and alignment.

The BERTScore F1 boxplots for all three models cluster near the 0.9 mark, affirming that semantic understanding is generally strong across systems. However, the divergence becomes more pronounced in BLEU and ROUGE metrics, where GPT-4o's literal and structurally aligned generation provides it with a clear advantage.

Discussion

This study compared the performance of three state-of-the-art large language models—GPT-4o, Claude Opus, and Gemini 1.5 Pro—within a Retrieval-Augmented Generation (RAG) framework designed to replace static patient information leaflets with a dynamic, voice-enabled chatbot. Based on both mean evaluation scores and score distributions across BLEU, ROUGE-1 F1, ROUGE-L F1, and BERTScore F1, GPT-4o consistently emerged as the best-performing model, demonstrating superior semantic alignment, structural coherence, and lexical accuracy. This section interprets the results in terms of model behavior, clinical suitability, and deployment implications.

Beyond the model-comparison focus, we position the system explicitly as a multilingual, speech-enabled, RAG-based chatbot for retinal detachment patient education—a dynamic alternative to static PILs—enabling interactive, personalized dialogue across languages and bridging communication and accessibility gaps. Traditional PILs are constrained by their static, one-way format and by practical limits on producing materials for highly diverse language needs; although telephone consultations can increase patient satisfaction, they are resource-intensive and add strain to already overloaded services [28].

In addition to mean scores, paired mean differences (Δ) in (Table 2) clarify practical gaps: GPT-4o exceeds Claude and Gemini most on ROUGE-1 (Δ=+0.062 and +0.022) and ROUGE-L (Δ=+0.049 and +0.032), with smaller—but directionally consistent—gains in BLEU and BERTScore.

There are four pragmatic routes to using LLMs for patient education: (i) train a model from scratch (maximum control but currently impractical due to data/compute demands); (ii) rely on web/app-based models (lowest setup but limited customization and governance); (iii) deploy an open-source model locally (maximizes privacy/control but requires substantial engineering and hardware); and (iv) consume models via API (balanced flexibility and integration potential). In this work, we adopt the API route with RAG, retrieving clinician-approved knowledge at inference. Compared with fine-tuning (permanent parameter updates), RAG improves factual accuracy and adaptability without retraining and keeps the knowledge surface auditable and updatable.

GPT-4o: Best overall performer for clinical use

GPT-4o achieved the highest average scores across all four-evaluation metrics and exhibited the narrowest interquartile ranges in boxplot analyses. Its BLEU score of 0.220 and ROUGE-1 F1 of 0.599 reflect its strong ability to reproduce clinically accurate phrasing, while its ROUGE-L F1 score of 0.458 indicates syntactic fluency and structural alignment with reference answers. Most importantly, GPT-4o's BERTScore F1 of 0.910 shows excellent semantic preservation, confirming that it retains core medical meaning even when rephrasing slightly.

These results position GPT-4o as the most reliable and consistent model for patient-facing clinical communication. Its minimal variance and absence of low-performing outliers suggest robust generalizability across diverse patient questions. Such predictability is essential in health contexts, where inconsistency in information delivery can undermine patient trust and safety. The model's tendency to maintain both surface-level phrasing and underlying meaning makes it particularly well-suited for scenarios requiring verbatim alignment with clinician-approved content, such as pre-operative guidance and discharge instructions. These advantages are reflected in the paired deltas (Table 2), supporting GPT-4o's superiority not only in central tendency but also in practically meaningful overlap with clinician-approved references.

From a systems perspective, this model stability complements our modular architecture in which retrieval and generation are deliberately decoupled, interfaces are kept flexible, and downstream integrations (e.g., electronic health record linkage or clinician review tooling) are feasible—supporting future use in virtual clinic pathways [29]. We also implement response segmentation and audio-clip concatenation to support longer spoken outputs within text-to-speech limits, and retain conversation history to personalize interactions and sustain engagement during post-operative recovery.

Gemini 1.5 Pro: Strong semantics with slight variability

Gemini 1.5 Pro delivered semantically competent responses, achieving a BERTScore F1 of 0.905, nearly matching GPT-4o. Its performance in ROUGE-1 F1 (0.577) and ROUGE-L F1 (0.426) indicates that it is generally capable of retrieving relevant terms and structuring responses effectively. However, its lower BLEU score (0.166) and broader IQRs in the boxplot suggest a greater propensity for paraphrasing or abstract reformulation, which may introduce ambiguity in tightly regulated clinical scenarios.

While Gemini is clearly capable, its stylistic variability may be better suited for engagement-oriented or health education use cases where interpretive flexibility is beneficial. For instance, it could serve well in multilingual symptom checkers or wellness advice platforms. However, for high-stakes applications requiring strict adherence to standardized medical phrasing, fine-tuning or stricter prompting may be necessary to ensure lexical fidelity.

Claude opus: Creative but unpredictable

Claude Opus scored the lowest across all metrics—BLEU (0.173), ROUGE-1 F1 (0.537), ROUGE-L F1 (0.409), and BERTScore F1 (0.895)—and exhibited the highest variability in the boxplot visualization. These results point to inconsistencies in alignment and phrasing, with frequent outliers that suggest model instability or drift in certain contexts.

The performance of Claude Opus indicates a greater degree of abstraction and conversational creativity, which may be valuable in open-ended or coaching scenarios but is less desirable in settings requiring rigid factual accuracy and phrasing consistency. Its outputs, while often semantically valid, lacked the precision and structural reliability demanded in healthcare communications, especially when responses are expected to be verifiable, auditable, and directly traceable to evidence-based sources.

Implications for safe clinical AI deployment

The evaluation results have clear implications for model selection in clinical AI applications. GPT-4o's superior performance across all metrics and its tight score distribution highlight it as the preferred model for safe, reliable, and high-fidelity deployment in patient-facing digital health tools. Its consistent alignment with clinician-approved content, minimal variance, and robust generation makes it suitable for integration into hospital-facing chatbots, post-operative guidance systems, or accessible alternatives to printed leaflets.

Gemini remains a strong contender for domains where expressive variation is acceptable or even desirable—such as patient education across multiple languages or non-critical mental health support. Claude, while the least suited for structured clinical delivery, could play a role in supportive health applications that prioritizes emotional tone and conversational breadth over strict factual anchoring.

These findings reinforce the notion that no single model is universally best across all medical use cases, and that task-specific fine-tuning, safety constraints, and deployment context must guide model selection. Operationally, the modular RAG-first design supports governance (versioned, auditable sources; guardrails tuned without retraining) and smooth integration with virtual clinic processes [29]. The system is designed to be consistent with NHS digital guidance and UK data-protection obligations; formal compliance would require a separate assurance process prior to any deployment.

Added value of distributional analysis

The inclusion of boxplot analysis provided deeper insights into model reliability. While average scores offer a baseline understanding of model performance, the variability and outliers revealed by boxplots exposed Claude's inconsistencies and Gemini's stylistic spread. In contrast, GPT-4o's tight interquartile ranges and lack of low-scoring outliers signal its readiness for deployment in safety-critical settings. These insights underscore the importance of going beyond simple averages when evaluating AI models for healthcare use.

Limitations and future directions

This study's evaluation was based solely on automated metrics comparing generated answers to expert-authored references. While such metrics are useful proxies for fluency, alignment, and semantic accuracy, they do not capture critical human factors such as perceived trust, emotional tone, or patient satisfaction. Follow-up studies will incorporate human-in-the-loop evaluation with patients and clinicians to validate real-world effectiveness and safety.

Additionally, the evaluation was conducted in English only, although the system is multilingual by design. Future work will involve large-scale multilingual testing to assess consistency, fluency, and semantic drift across supported languages such as Arabic, Farsi, Spanish, French, and Hindi. In particular, outcomes may vary due to automatic speech-recognition accuracy and language-specific training biases; these will be explicitly evaluated in multilingual trials.

Real-world deployment that touches patient-specific data will require strong privacy protections and clear escalation/oversight routes for clinicians; where the use case moves toward clinical decision support, formal regulatory approval will be required. Finally, while our exemplar is retinal detachment, the same architecture is readily adaptable to other ophthalmic conditions and to other surgical pathways where dynamic, accessible education is required.

Conclusion

This study presented the design, implementation, and evaluation of a multilingual, speech-enabled, Retrieval-Augmented Generation (RAG) chatbot aimed at replacing static Patient Information Leaflets (PILs) with a dynamic, conversational interface for retinal detachment patient education. By integrating curated clinical knowledge with leading-edge Large Language Models (LLMs) such as GPT-4o, Claude Opus, and Gemini 1.5 Pro, the system delivers context-aware, medically grounded, and multilingual responses through both text and speech. The use of FAISS-based semantic retrieval, neural reranking, and structured prompt engineering ensures that generated answers remain accurate, traceable, and relevant to user queries.

The experimental results, based on clinically curated questions, demonstrated that GPT-4o outperformed its counterparts in BLEU, ROUGE, and BERTScore metrics, suggesting its suitability for high-stakes healthcare applications requiring nuanced and factually consistent responses. The chatbot further addresses critical challenges in accessibility by offering support for multiple languages, real-time Text-To-Speech (TTS), screen-reader compatibility, and audio-based input, thus ensuring usability for patients with diverse linguistic and sensory needs.

This work fills several key knowledge and practice gaps identified in the literature. Notably, it is among the first to operationalize RAG for a patient-facing use case in ophthalmology, conduct a direct comparative evaluation of top-tier LLMs within a unified retrieval framework, and incorporate accessibility and regulatory compliance principles from the ground up. In doing so, it contributes not only a functional prototype but also a replicable methodology that can inform the development of explainable, trustworthy, and inclusive medical AI systems.

Future research will extend this framework to other clinical domains, incorporate user-centered evaluations with patients and clinicians, and explore the integration of the chatbot into Electronic Health Record (EHR) systems and post-operative care workflows. Additional safeguards such as explainable AI overlays, patient consent modules, and clinical validation pipelines will also be developed to prepare the system for real-world deployment in compliance with NHS Digital Standards, UK GDPR, and MHRA AI guidelines.

In summary, this work provides a robust foundation for the next generation of AI-powered health communication tools—tools that are not only intelligent but also inclusive, explainable, and safe.

Declarations

Conflicts of interest: The authors declare no conflicts of interest.

Funding statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Iskander M, Hu G, Coulon S, Seixas AA, McGowan R, Al-Aswad LA. Health literacy and ophthalmology: a scoping review. Surv Ophthalmol. 2023; 68: 78–103.

2. Anguita R, Ting MYL, Makuloluwa A, Charteris DG. Causal factors for late presentation of retinal detachment. Eye (Lond). 2023; 37: 185–186.

3. Anguita R, Roth J, Makuloluwa A, et al. Late presentation of retinal detachment: clinical features and surgical outcomes. Retina. 2021; 41: 1833–1838.

4. Bickmore TW, Pfeifer LM, Jack BW. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009: 1265–1274.

5. Williams AM, Muir KW, Rosdahl JA. Readability of patient education materials in ophthalmology: a single-institution study and systematic review. BMC Ophthalmol. 2016; 16: 133.

6. Lim J, Kim W, Kim I, Lee E. Effects of visual communication design accessibility guidelines for low vision on public and open government health data. Healthcare (Basel). 2023; 11: 1047.

7. Nielsen BR, Alberti M, Bjerrum SS, la Cour M. The incidence of rhegmatogenous retinal detachment is increasing. Acta Ophthalmol. 2020; 98: 603–606.

8. Bommasani R, Hudson D, Adeli E, et al. On the opportunities and risks of foundation models. arXiv. 2021; arXiv:2108.07258.

9. Bibault JE, Chaix B, Nectoux P, et al. Healthcare ex machina: are conversational agents ready for prime time in oncology? Clin Transl Radiat Oncol. 2019; 16: 55–59.

10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavioral therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): randomized controlled trial. JMIR Ment Health. 2017; 4: e19.

11. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, Beam AL. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health. 2024; 6: e555–e561.

12. Anguita R, Makuloluwa A, Hind J, Wickham L. Large language models in vitreoretinal surgery. Eye (Lond). 2024; 38: 809–810.

13. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. Application and accuracy of artificial intelligence-derived large language models in patients with age-related macular degeneration. Int J Retina Vitreous. 2023; 9: 71.

14. Schumacher I, Ferro Desideri L, Bühler VMM, et al. Performance analysis of an emergency triage system in ophthalmology using a customized chatbot. Digit Health. 2025; 11: 20552076251320298.

15. Anguita R, Downie C, Ferro Desideri L, Sagoo MS. Assessing large language models' accuracy in providing patient support for choroidal melanoma. Eye (Lond). 2024; 38: 3113–3117.

16. Sabaner MC, Anguita R, Antaki F, et al. Opportunities and challenges of chatbots in ophthalmology: a narrative review. J Pers Med. 2024; 14: 1165.

17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023; 29: 1930–1940.

18. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023; 620: 472–480.

19. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. 2020; 33: 9459–9474.

20. Zhao X, Liu S, Yang SY, Miao C. MedRAG: grounded medical question answering with retrieval-augmented generation. In: Proceedings of the ACM on Web Conference. 2025: 4442–4457.

21. Thakur N, Reimers N, Ruckle A, et al. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Advances in Neural Information Processing Systems. 2021; 34: 5925–5940.

22. Holzinger A, Beimann C, Pattichis C, et al. What do we need to build explainable AI systems for the medical domain? arXiv. 2017; arXiv:1712.09923.

23. NHSX. Digital Technology Assessment Criteria (DTAC) for health and social care. 2021. Available at: https://transform.england.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/

24. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311–318.

25. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004: 74–81. Available at: https://aclanthology.org/W04-1013/

26. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations. 2020. Available at: https://openreview.net/forum?id=SkeHuCVFDr

27. Duong A, Valero M. Usability of voice assistants in healthcare: a systematic literature review. In: Salvi D, Van Gorp P, Shah SA, eds. Pervasive Computing Technologies for Healthcare. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Vol 572. Cham: Springer; 2024.

28. Shahid SM, Anguita R, da Cruz L. Telemedicine for postoperative consultations following vitrectomy for retinal detachment repair during the COVID-19 crisis: a patient satisfaction survey. Can J Ophthalmol. 2021; 56: e46–e48.

29. Anguita R, Ahmed S, Makuloluwa A, Hind J, Roth J, Wickham L. Prospective validation of a virtual post-operative clinic in vitreoretinal surgery. Eye (Lond). 2024; 38: 3258–3262.

30. Alanezi F. Factors influencing patients' engagement with ChatGPT for accessing health-related information. Crit Public Health. 2024; 34: 1–20.

31. Hua H, Kaakour A, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol. 2023; 141: 819–824.

32. Ming S, Yao X, Guo X, et al. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: cross-sectional study. J Med Internet Res. 2024; 26: e60226.

33. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025; 32: 605–615.

34. Anisha SA, Sen A, Bain C. Evaluating the potential and pitfalls of AI-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: scoping review. J Med Internet Res. 2024; 26: e56114.