# Advances in AI for Health and Medicine

A series of research initiatives demonstrate the rapid evolution of Artificial Intelligence, particularly Large Language Models (LLMs), across the full spectrum of healthcare—from specialist-level clinical reasoning to longitudinal disease management and personalized consumer health. Models built on the Gemini architecture, such as the Articulate Medical Intelligence Explorer (AMIE), MedGemma, and the Personal Health Large Language Model (PH-LLM), are achieving performance comparable to, and in some cases exceeding, that of human clinicians in rigorous evaluations.

The primary value of these AI systems often lies in their ability to complement and augment human expertise rather than replace it. Studies show that clinicians assisted by AI demonstrate improved diagnostic accuracy, create more comprehensive management plans, and achieve greater consensus. This complementary effect is a consistent theme, highlighting the potential for human-AI collaboration to enhance the quality and accessibility of care.

Validation of these systems is advancing beyond simple accuracy benchmarks to encompass multi-axis evaluations, including blinded head-to-head comparisons with physicians, standardized Objective Structured Clinical Examinations (OSCEs), and qualitative user studies. Methodologies are increasingly focused on real-world applicability, assessing not just correctness but also clinical helpfulness, safety, guideline adherence, and the quality of human-AI interaction.

Technologically, the field is advancing toward more sophisticated, agentic systems that leverage long-context reasoning, multimodal data integration (text, imaging, sensor data), and grounding in authoritative knowledge sources like clinical practice guidelines. This enables capabilities beyond single-point diagnostics, supporting complex, multi-visit disease management and personalized coaching based on continuous data from wearable devices. These developments signal a significant shift towards AI systems that can reason dynamically, manage longitudinal patient journeys, and empower both providers and individuals with data-driven insights.

## I. The Evolving Capabilities of Medical AI

The application of AI in health has expanded from narrowly focused diagnostic tasks to encompass a continuum of care, demonstrating proficiency in specialist-level reasoning, primary care management, and personal wellness coaching.

### 1.1. Specialist-Level Diagnostic and Clinical Reasoning

AI models are demonstrating capabilities that match or augment those of medical specialists in complex diagnostic scenarios.

- **Ophthalmology:** In a study using 100 real-world clinical vignettes, the AMIE system (a medically fine-tuned LLM based on Gemini) showed standalone diagnostic performance comparable to ophthalmologists.

Critically, when clinicians reviewed AMIE's structured output, they tended to rank the correct diagnosis higher, reached greater inter-rater agreement, and enriched their investigation plans. This improvement was observed even when AMIE's top choice was incorrect, indicating a complementary effect where the AI's structured reasoning helps clinicians re-evaluate their own thinking.

- **Oncology:** An evaluation using 50 synthetic breast cancer scenarios found that AMIE outperformed trainees and fellows but was inferior to attending oncologists. The study highlighted the AI's potential to democratize access to subspecialty expertise, particularly in under-resourced settings, by aiding in initial triage and treatment decisions.
- **Complex Differential Diagnosis:** In a study of 302 challenging case reports from the *New England Journal of Medicine*, an LLM optimized for differential diagnosis (DDx) exhibited standalone performance exceeding that of unassisted board-certified internal medicine physicians (Top-10 accuracy of 59.1% vs. 33.6%). When used as an assistive tool, the LLM improved clinician accuracy more effectively than standard internet search tools.

### 1.2. Longitudinal Disease Management

Beyond single-point diagnosis, research is focused on developing agentic AI systems capable of managing patient care over multiple encounters, grounded in established clinical guidelines.

An advanced version of AMIE was developed as an agentic system optimized for clinical management and dialogue. The system incorporates:

- **Longitudinal Reasoning:** The ability to reason over the evolution of a disease, patient responses to therapy, and multiple clinical visits.
- **Guideline Grounding:** Use of Gemini's long-context capabilities to retrieve and reason over a corpus of clinical practice guidelines (e.g., UK NICE Guidance, BMJ Best Practice) to inform management plans.
- **Agentic Architecture:** A system composed of a Dialogue Agent for patient interaction and a Management (Mx) Agent that retrieves guidelines, drafts, and refines management plans.

In a virtual OSCE study involving 100 multi-visit case scenarios, this AMIE system was non-inferior to primary care physicians (PCPs) in management reasoning and scored higher in the preciseness of treatments and its alignment with clinical guidelines.

### 1.3. Personal Health and Wellness

AI models are being developed to interpret heterogeneous, longitudinal data from wearable devices (e.g., Fitbit, Pixel Watch) to provide personalized health insights for sleep and fitness.

- **Personal Health Large Language Model (PH-LLM):** A version of Gemini fine-tuned to generate insights and recommendations from numerical time-series data. In evaluations using 857 expert-created case studies, PH-LLM approached expert performance. For fitness, its performance was not statistically different from human experts. For sleep, while experts remained superior, fine-tuning significantly improved the model's use of domain knowledge. PH-LLM also exceeded expert scores on professional multiple-choice exams for sleep medicine (79% accuracy) and fitness certification (88% accuracy).
- **Personal Health Insights Agent (PHIA):** An agentic system that uses code generation and information retrieval tools to analyze wearable data. PHIA accurately addresses over 84% of factual numerical questions and over 83% of open-ended health queries. It significantly outperforms baselines by integrating web search for domain knowledge with code generation for data analysis, demonstrating a more robust and reliable method for deriving personalized insights.

### 1.4. Consumer Health Information Seeking

Research is also exploring how to improve LLM interactions for laypeople seeking health information. The "Wayfinding AI" is a prototype designed to proactively seek context before providing answers. In a randomized study, participants rated the Wayfinding AI as more helpful, relevant, and tailored compared to a baseline AI. This context-seeking behavior led to longer, more collaborative conversations, demonstrating a design pattern that can improve the quality and user perception of consumer-facing health AIs.

## II. Technical Foundations and Model Ecosystem

The advancements in health AI are supported by a growing ecosystem of foundation models, specialized architectures, and development platforms designed to accelerate research and application development.

### 2.1. Core Models and Architectures

A suite of models, primarily based on the Gemini architecture, underpins these capabilities.

| Model / System | Base Architecture | Key Features & Purpose |
|---|---|---|
| **AMIE** | Gemini | Medically fine-tuned for conversational clinical reasoning. Evolved into an agentic system for longitudinal disease management using long-context and guideline retrieval. |
| **MedGemma** | Gemma 3 (4B & 27B) | Open, medically-tuned vision-language models for interpreting medical images and text. Part of the HAI-DEF collection. Demonstrates strong performance on medical VQA and text benchmarks. |
| **MedSigLIP** | SigLIP-400M | A standalone, medically-tuned 400M-parameter vision encoder that powers MedGemma's image capabilities. Shows performance comparable to or exceeding specialized medical image encoders. |

| | | |
|---|---|---|
| **PH-LLM** | Gemini Ultra 1.0 | Fine-tuned for reasoning over numerical time-series data from wearables to provide sleep and fitness coaching. Integrates a multimodal adapter for predicting patient-reported outcomes. |
| **PHIA** | State-of-the-art LLM | An agentic system using code generation and web search tools to analyze wearable data and answer objective and open-ended personal health questions. |

### 2.2. Foundational Platforms and Tools

To lower the barrier to entry for developing medical AI, foundational platforms provide pre-trained models and tools.

- **Health AI Developer Foundations (HAI-DEF):** A suite of pre-trained, domain-specific foundation models, tools, and recipes designed to accelerate the development of ML for health applications. It covers various modalities, including:
  - Radiology (CXR Foundation, CT Foundation)
  - Histopathology (Path Foundation)
  - Dermatology (Derm Foundation)
  - Audio (HeAR)
- **The MedGemma Collection:** As part of HAI-DEF, this suite includes the MedGemma 4B (multimodal) and 27B (text-only) models, plus the MedSigLIP image encoder. These open models serve as a starting point for building health applications.

### 2.3. Key Technical Innovations

Several key technical approaches are driving progress:

- **Agentic Systems:** Moving beyond single-shot generation, models like AMIE and PHIA operate as agents that can use tools (web search, code execution), maintain an internal state across interactions, and follow multi-step reasoning processes (e.g., draft, critique, revise).
- **Long-Context Reasoning:** Gemini's long-context window (e.g., two million tokens) is leveraged by the AMIE Mx Agent to process multiple, extensive clinical guideline documents simultaneously, enabling rich cross-document reasoning without complex retrieval pipelines.
- **Multimodality:** Models are increasingly capable of integrating diverse data types. MedGemma processes both images and text. PH-LLM was adapted to natively integrate time-series sensor data via a learned adapter, enabling it to predict subjective user outcomes from objective data.
- **Domain-Specific Fine-Tuning and Alignment:** Models like MedGemma and PH-LLM are fine-tuned on specialized datasets to enhance their domain knowledge. AMIE undergoes supervised fine-tuning (SFT) and reinforcement learning from human/AI feedback (RLHF/RLAIF) on

simulated medical dialogues and management tasks to improve its conversational and clinical capabilities.

## III. Rigorous and Multi-Axis Evaluation Frameworks

Credible evaluation is as essential as capability. The research employs diverse and sophisticated methodologies to validate AI performance, safety, and real-world utility, moving beyond simple accuracy metrics.

### 3.1. Human-AI Interaction Studies

These studies assess the complementary effect of AI on clinical practice.

- **Ophthalmology Diagnostic Reasoning:** Ophthalmologists (residents, fellows, consultants) first provided a differential diagnosis for clinical vignettes, then revised it after viewing AMIE's output. This measured the shift in diagnostic ranking, inter-rater agreement, and management plan content, showing significant improvements post-AI interaction.
- **Differential Diagnosis Assistance:** Twenty board-certified physicians evaluated 302 complex cases, randomized to receive assistance from either standard search tools or an LLM. Results showed that LLM assistance led to a greater improvement in top-10 diagnostic accuracy (from 36.1% to 51.7%) compared to search assistance (from 36.1% to 44.4%).

### 3.2. Head-to-Head Comparisons with Clinicians

Masked, randomized studies compare AI output directly against human experts.

- **Objective Structured Clinical Examination (OSCE):** To evaluate AMIE's disease management capabilities, a virtual OSCE was conducted with 21 PCPs and trained patient actors over 100 multi-visit scenarios. Specialist physicians, blinded to the source, evaluated consultation transcripts and management plans across 15 axes, including guideline alignment, preciseness, and appropriateness.
- **Specialist Rubrics:** In the oncology study, specialist-designed rubrics were used to score responses from AMIE, trainees, fellows, and attendings across axes like summarization, management reasoning, and safety. This allowed for a granular comparison of performance by expertise level.

### 3.3. Performance on Standardized Benchmarks and Exams

Models are tested against established benchmarks to quantify their domain knowledge.

| Model | Benchmark/Exam | Key Result |
|---|---|---|
| **PH-LLM** | Sleep Medicine Board Exam (AMA PRA/ABIM MOC) | **79% accuracy** (exceeds 76% by human experts and ~70% for CME credit). |

| | | |
|---|---|---|
| **PH-LLM** | Fitness Certification Practice Exam (NSCA-CSCS) | **88% accuracy** (exceeds 71% by human experts and ~70% passing grade). |
| **MedGemma 27B** | MedQA (USMLE-style) | **89.8% accuracy**, significantly outperforming its base model (Gemma 3 27B at 74.9%). |
| **MedGemma 4B/27B** | Various Text & Vision Benchmarks (MedMCQA, PubMedQA, MMLU Med, VQA-RAD) | Consistently outperforms base models and is competitive with much larger models. |

### 3.4. Novel Dataset and Benchmark Creation

Where existing benchmarks are insufficient, new datasets are created to test specific capabilities.

- **Personal Health Case Studies (PH-LLM):** 857 expert-authored case studies in sleep and fitness, combining wearable data with long-form insights and recommendations, were created to evaluate personalized coaching.
- **Personal Health Queries (PHIA):** Over 4,000 objective (fact-based) and 172 open-ended (exploratory) health questions were curated to benchmark an agent's ability to analyze wearable data.
- **RxQA (AMIE):** A 600-question benchmark derived from US and UK drug formularies and validated by pharmacists to specifically test medication reasoning.
- **EHRQA (MedGemma):** A benchmark for longitudinal reasoning over electronic health records, generated from synthetic FHIR-formatted patient records.

### 3.5. Qualitative and User-Centered Analysis

Qualitative methods provide crucial insights into user experience, trust, and the practical utility of AI systems.

- **Clinician Interviews:** Semi-structured interviews with physicians in the DDx study revealed that they found the LLM more helpful than search for cases where they were uncertain, as it could "pull some additional diagnoses that would be important to think about."
- **Consumer Health Studies (Wayfinding AI):** Mixed-methods studies (N=163) explored how laypeople interact with health LLMs. Qualitative feedback showed that context-seeking by the AI was highly valued, making the interaction feel more like a human conversation and increasing confidence in the eventual answer. This insight directly informed the design of the "Wayfinding" prototype.
- **Expert Rater Feedback (PHIA & PH-LLM):** Interviews with expert raters provided nuanced feedback. They noted that personalization was strongly linked to the use of numerical data, and that integrating domain

knowledge (e.g., comparing a user's sleep data to recommended guidelines) elevated response quality.

## IV. Key Findings and Implications

### 4.1. The Power of Augmentation and Complementation

A recurring finding is that AI is most powerful when augmenting, not replacing, human expertise.

- **Improved Clinician Performance:** In the ophthalmology study, clinician diagnostic accuracy improved across all grades (resident, fellow, consultant) after reviewing AMIE's output. Similarly, in the complex DDx study, LLM assistance led to a 15.6 percentage point increase in top-10 accuracy, compared to an 8.3 point increase with search.
- **Personalization by Expertise:** Preferences for AI responses can vary by experience. In the ophthalmology study, residents favored AMIE's comprehensive narratives, while consultants preferred more focused human answers, suggesting an opportunity to personalize AI outputs for different user needs.

### 4.2. Precision and Guideline Adherence in Management

AI systems demonstrate a high degree of precision and adherence to evidence-based standards.

- **Precise Recommendations:** In the multi-visit OSCE study, AMIE was consistently rated as more precise than PCPs in its treatment recommendations (e.g., 94% vs. 67% in visit 1) and investigation plans. This translates to more actionable directives (e.g., specific drug, dose, and duration) rather than general suggestions.
- **Guideline Alignment:** AMIE's management plans were significantly more aligned with clinical guidelines across all three visits (e.g., 89% vs. 75% in visit 1) and more frequently supported by explicit citations to the source documents. This suggests a strong potential for AI to support the implementation of guideline-directed medical therapy.

### 4.3. Transforming Unstructured Data into Actionable Insights

AI agents are proving capable of analyzing complex, unstructured, and longitudinal personal health data from wearables.

- **Objective and Open-Ended Analysis:** PHIA can answer both simple factual questions ("What was my average sleep duration?") and complex, open-ended ones ("How can I improve my fitness?"). It achieves this by combining code generation to analyze numerical data with web search to retrieve relevant domain knowledge (e.g., recommended sleep duration for an adult).

- **Predicting Subjective States:** PH-LLM demonstrated that by using a multimodal adapter to process time-series sensor data, it could predict self-reported sleep quality outcomes with performance matching specialized discriminative models and significantly outperforming text-only prompting approaches. This bridges the gap between objective sensor data and a user's subjective experience.

### 4.4. Enhancing the User Experience in Health Conversations

The design of the AI interaction profoundly impacts its utility and user perception.

- **Value of Context-Seeking:** The Wayfinding AI studies showed that users strongly prefer a conversational AI that asks clarifying questions to gather context before providing a definitive answer. This behavior was rated as more helpful, relevant, and tailored, and it fostered more collaborative conversational dynamics where users were more likely to provide clarifying information. This contrasts with baseline models, where users were more likely to pivot to new topics or end the conversation early.