HealthBot: An Open-Source AI Assistant for Longitudinal Personal Health Management

Zheng Yu [1] **Yanyuan Qiao** [2] **Qi Wu*** [1] **Yutong Xie** [3]

Abstract

Medical LLMs and VLMs perform well on clinical QA and report generation, but remain largely stateless and cannot manage personal health data over time. Existing medical agents mainly target specific benchmarks or workflows, while commercial systems such as ChatGPT Health are closed-source and cloud-dependent. We present HealthBot, an open-source personal health assistant that runs locally, stores data on-device, supports messaging-platform interfaces, and works with model-agnostic LLM backends. HealthBot integrates multimodal extraction of heterogeneous health inputs into normalized records, Hierarchical Health Context (H-Context) for budget-aware longitudinal reasoning, and tool-augmented reasoning grounded in the local archive for report interpretation, trend analysis, consultation preparation, and medication tracking. We evaluate longitudinal context on a diagnosis prediction benchmark, where models infer the current diagnosis from laboratory and vital measurements under three settings: no history, raw history, and H-Context. Across two LLM backbones, history substantially improves accuracy, and H-Context further improves over raw history while reducing context length by 39% (up to +20 pp over no-history), demonstrating the value of structured longitudinal context for personal health agents.

1. Introduction

Large language models (LLMs) and vision-language models (VLMs) have rapidly advanced medical AI. Medical LLMs have shown strong performance in clinical question answering, reasoning, and decision support (Singhal et al., 2025; Chen et al., 2023; Labrak et al., 2024; Wu et al., 2024; Zhang et al., 2023; Maity & Saikia, 2025; Ye & Tang, 2025), while medical VLMs extend these capabilities to biomedical image understanding and multimodal clinical tasks (Li et al., 2023; Zhang et al., 2024; Chen et al., 2024; Lin et al., 2025; Saab et al., 2024). However, most remain model-level systems and lack persistent, structured reasoning over an individual's longitudinal health history, which is essential for personal health assistants.

Recent medical agents have explored multi-agent decision-making (Kim et al., 2024; Wu et al., 2025), simulated clinical workflows (Schmidgall et al., 2024; Li et al., 2024b), and tool- or code-augmented reasoning over EHR tables, knowledge graphs, and domain tools (Shi et al., 2024; Zuo et al., 2025; Li et al., 2024a). Personal health agents such as openCHA and PHIA further connect conversational agents to external data sources and wearable streams (Abbasian et al., 2025; Merrill et al., 2026). Yet most systems remain benchmark-, workflow-, or modality-specific, lacking full-stack support for heterogeneous personal health documents with structured extraction, cross-source normalization, and longitudinal tracking. More recently, ChatGPT Health integrates medical records and wearable data into a dedicated ChatGPT module (OpenAI, 2026), but remains closed-source, limited-access, and cloud-dependent, limiting reproducibility and raising privacy concerns.

To address these gaps, we propose **HealthBot**, an open-source personal health AI assistant built on NanoBot (HKUDS, 2026), a lightweight messaging-agent framework inspired by OpenClaw (Steinberger, 2026). HealthBot extends this framework for longitudinal personal health management with: (1) a *multimodal knowledge extraction pipeline* that converts conversations, medical images, EHR documents, and wearable exports into normalized structured records; (2) *Hierarchical Health Context (H-Context)* $\mathcal{H}=\{P, T, S\}$, which compresses health archives into profile snapshots, rolling timelines, and historical summaries under a token budget; and (3) *tool-augmented reasoning* over locally stored health data for report interpretation, consultation preparation, trend analysis, and medication tracking.

We further evaluate whether structured longitudinal context

[1]Adelaide University, Adelaide, Australia [2]EPFL, Lausanne, Switzerland [3]MBZUAI, Abu Dhabi, UAE. Correspondence to: Qi Wu <qi.wu01@adelaide.edu.au>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

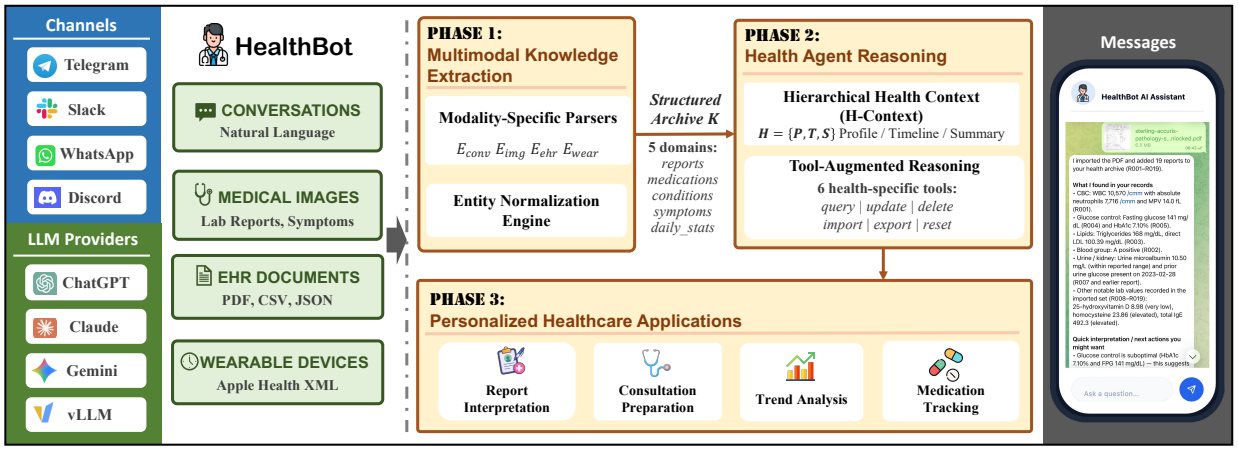


Figure 1. Overview of HealthBot. Multimodal inputs are normalized into a structured archive (Phase 1), which a health LLM agent accesses through hierarchical context $\mathcal{H} = \{P, T, S\}$ and tools (Phase 2) to support personalized healthcare applications (Phase 3).

improves diagnostic reasoning under tight input budgets. We construct a longitudinal diagnosis prediction benchmark from Synthea (Walonoski et al., 2018) and MIMIC-IV (Johnson et al., 2023), where models predict the current diagnosis from observed laboratory and vital measurements under three history settings: current-only, raw history, and H-Context. Across GPT-5.2 (OpenAI, 2025) and Qwen3-8B (Yang et al., 2025), H-Context consistently improves over raw history while reducing history context by 39%, reaching up to +20 pp over current-only.

Our contributions are threefold: (1) we present **HealthBot**, an open-source, locally hosted personal health assistant that converts heterogeneous health inputs into a structured longitudinal archive and supports tool-grounded reasoning via messaging interfaces; (2) we propose **H-Context**, a budget-aware hierarchical representation for long-term health reasoning, and validate it on a curated diagnosis prediction benchmark; and (3) we release the full codebase to support reproducible research on personal health AI assistants.

2. The HealthBot Framework

As shown in Fig. 1, HealthBot maps heterogeneous inputs—conversations, clinical document images, EHR files, and wearable exports—into a structured health archive $\mathcal{K}$ with five domains: `reports`, `medications`, `conditions`, `symptoms`, and `daily_stats`. It then supports health-agent reasoning through hierarchical context, archive-grounded tools, and downstream applications including report interpretation, consultation preparation, trend analysis, and medication tracking.

2.1. Multimodal Knowledge Extraction

This phase addresses two challenges: *format fragmentation* (the same health information may appear as images, PDFs, chat text, or wearable exports) and *semantic fragmentation* (the same metric may be named differently across sources and languages). We convert these inputs into (domain, record) pairs in five semantic domains, where Table 1 lists the required fields per domain.

Table 1. Minimum record schema per semantic domain. All records additionally carry `source_meta` preserving provenance.

Domain	Required fields
`reports`	`kind, type, ref_date, abnormal_items[{name, val, unit, flag}]`
`medications`	`name, dosage, frequency, start_date, status`
`conditions`	`name, status, diagnosed_date`
`symptoms`	`content, ref_date, status, tags`
`daily_stats`	`steps, weight_kg, heart_rate, blood_pressure, sleep, exercise`

2.1.1. Modality-Specific Parsers.

(1) *Conversation extraction* (E_{conv}). Health information in chat is often spread across multiple turns. We therefore use *batch extraction*, aggregating consecutive messages into a single LLM call to resolve cross-turn attributes (e.g., linking a medication to its dosage) while reducing redundant invocations. The extractor recognizes three action types: `new` (novel health records), `status_change` (e.g., symptom resolution), and `dosage_change`, followed by deterministic deduplication over normalized keys and timestamps.

(2) *Medical image extraction* (E_{img}). We perform a single vision-capable LLM pass that jointly identifies the document type and extracts schema-conformant fields, producing domain records such as `reports`, `medications`, and `conditions`.

(3) *EHR document extraction* (E_{ehr}). Structured files (CSV/JSON) are parsed via header/schema mapping, while PDFs are handled with a local PDF library as text parser first and a vision fallback when text extraction is insufficient.

(4) *Wearable data extraction* (E_{wear}). We stream-parse wearable data exports (e.g., Apple Health XML) and aggregate sensor readings into daily `daily_stats` summaries (steps, sleep, heart rate, etc.), converting timestamps to the user's

Table 2. Entity normalization aligns the same clinical metric across heterogeneous sources to a unified canonical key κ.

Source	Original n	Key κ	Display n_{std}
Chinese lab report	空腹血糖	`fasting_glucose`	Fasting Plasma Glucose
English report	FPG	`fasting_glucose`	Fasting Plasma Glucose
Conversation	"fasting glucose 5.8"	`fasting_glucose`	Fasting Plasma Glucose
Chinese lab report	白细胞	`wbc`	White Blood Cell Count

local timezone to ensure correct date boundaries.

2.1.2. Entity Normalization Engine.

Cross-source metric alignment is essential for longitudinal analysis: without normalization, the same clinical measurement may appear under different names and become difficult to track over time. We assign each extracted metric a three-level name (Table 2): the *original name* n from the source, a stable *canonical key* κ (English snake_case) for aggregation, and a *standardized display name* n_{std} for presentation. To prevent semantic drift, we use a lightweight *consistency loop*: during extraction, a set of existing canonical keys is provided to the LLM with the rule to reuse an existing key for synonyms and create a new key only when needed. This mechanism is applied across all extraction paths (E_{conv}, E_{img}, E_{ehr}), encouraging consistent metric identifiers across documents and conversations.

2.2. Health Agent Reasoning

After months, directly injecting all records $\mathcal{K}$ into the LLM context becomes infeasible. We address this by hierarchical context and tool-augmented reasoning.

2.2.1. Hierarchical Health Context

We design a three-layer context $\mathcal{H} = \{P, T, S\}$, injected into the system prompt as a compact view of the user's longitudinal archive for health reasoning. *Profile* P summarizes relatively stable and currently active information (e.g., demographics, active conditions/medications/symptoms, key clinical metrics). *Timeline* T lists events from a recent rolling window of the last w months in reverse chronological order. *Summary* S compresses earlier history prior to the last w months into a concise narrative capturing major events, long-term trends, and medication changes. We concatenate P, T, and S as $\mathcal{H}(q) = \textsc{Concat}(P,\ T,\ S)$ under a fixed context budget $|\mathcal{H}| \leq B_{\max}$; when the budget is exceeded, T is truncated and the agent is instructed to retrieve missing details via tools.

2.2.2. Tool-Augmented Reasoning.

The agent is equipped with six health-specific tools forming a complete read-write interface over $\mathcal{K}$: `query` (retrieves records by domain, time range, or keyword), `update` (modifies profile or records), `delete` (removes records with document cleanup), `import` (triggers E_{img}/E_{ehr} for uploaded files), `export` (packages the archive as a portable zip), and `reset` (clears all data with confirmation). The system persona encodes explicit *retrieval rules*: the agent must invoke `query` rather than relying solely on $\mathcal{H}$ when a question spans beyond w months, requires exact historical values, or involves trend analysis. This design grounds responses in the underlying records and reduces hallucinations introduced by context compression.

2.3. Personalized Healthcare Applications

Building on the structured knowledge base $\mathcal{K}$, hierarchical health context $\mathcal{H}$ and the agent's tool-grounded reasoning, HealthBot supports four categories of personalized health applications. *Report Interpretation* translates lab results into accessible explanations by identifying abnormal findings and retrieving κ-aligned historical measurements for comparison. *Consultation Preparation* generates structured visit summaries by aggregating symptoms, medications, and test results into chronological narratives to facilitate physician–patient communication. *Trend Analysis* performs cross-source longitudinal tracking, reporting comparisons with limited history and trends only when multiple measurements support them, avoiding trend claims when observations are sparse. *Medication Tracking* maintains an auditable pharmaceutical timeline where dosage updates are represented as stop-and-create event pairs.

3. Experiments

We study whether structured longitudinal context can improve LLM-based clinical reasoning when patient history grows beyond practical context budgets. While historical information is often helpful, naively concatenating all past records quickly becomes too long to fit, motivating compact representations for long-term health reasoning. We therefore evaluate the effectiveness and generalizability of HealthBot's Hierarchical Health Context (H-Context).

3.1. Experimental Setup

Benchmark. We construct a longitudinal diagnosis prediction benchmark from Synthea and MIMIC-IV. We select patients/admissions with at least two encounters, requiring the latest target encounter/admission to contain a primary diagnosis and at least five laboratory results. The benchmark contains 100 trajectories, evenly split between the two sources. Each trajectory includes demographics, up to six historical encounters with labs, Synthea vitals, and diagnoses; the target encounter provides only observed measurements, and the model predicts its diagnosis. We formulate this as four-way multiple choice, with one ground-truth diagnosis and three hash-seeded distractors, yielding a 25% random baseline.

Table 3. Diagnostic accuracy (%) across two LLM backbones. **Bold** marks the best result per model-column pair. Random baseline is 25%.

	GPT-5.2			Qwen3-8B		
Cond.	Overall	Syn.	MIMIC	Overall	Syn.	MIMIC
A: Current Only	59.0	75.5	40.4	48.0	64.2	29.8
B: + Raw Hist.	66.0	79.2	51.1	66.0	77.4	53.2
C: + H-Context	**70.0**	**84.9**	**53.2**	**68.0**	**81.1**	**53.2**
Δ A→B	+7.0	+3.7	+10.7	+18.0	+13.2	+23.4
Δ B→C	+4.0	+5.7	+2.1	+2.0	+3.7	+0.0
Δ A→C	+11.0	+9.4	+12.8	+20.0	+16.9	+23.4

Conditions and models. We compare **A: Current Only**, where the model receives demographics and target measurements only; **B: Raw History**, which adds the full prior encounter timeline verbatim; and **C: H-Context**, which adds HealthBot's $\mathcal{H} = \{P, T, S\}$ representation. We evaluate GPT-5.2 (OpenAI, 2025) via the OpenAI Responses API and Qwen3-8B (Yang et al., 2025) locally via vLLM (Kwon et al., 2023) on a single NVIDIA RTX 5090 GPU. H-Context is pre-generated using GPT-5 mini.

3.2. Main Results

3.2.1. History Matters.

Adding historical context (A→B) yields consistent improvements across both models and data sources. GPT-5.2 gains +7.0 pp overall while Qwen3-8B gains +18.0 pp, indicating that smaller models benefit even more from longitudinal information. This confirms that maintaining a persistent health knowledge base provides clinically relevant signals beyond what single-visit snapshots capture. The relative improvements from both raw history and H-Context are larger for the smaller model (e.g., $\Delta_{A\to C}$: +20.0 pp using Qwen3-8B vs. +11.0 pp using GPT-5.2 with H-Context), confirming that weaker models stand to gain the most from longitudinal health context.

3.2.2. Structured Context Outperforms Raw History.

H-Context (B→C) further improves accuracy for both models (+4.0 pp for GPT-5.2, +2.0 pp for Qwen3-8B). On Synthea, gains reach +5.7 pp and +3.7 pp respectively, suggesting that structured compression is particularly effective when history entries contain heterogeneous observation types (labs, vitals, conditions). Notably, the A<B<C ordering holds consistently across both models and both data sources, demonstrating that H-Context's benefit is *model-agnostic*. Remarkably, Qwen3-8B with H-Context (68.0%) approaches GPT-5.2 with raw history (66.0%), suggesting that structured context can partially compensate for model capacity.

Table 4. Input efficiency comparison: H-Context provides higher overall accuracy for GPT-5.2 with fewer tokens.

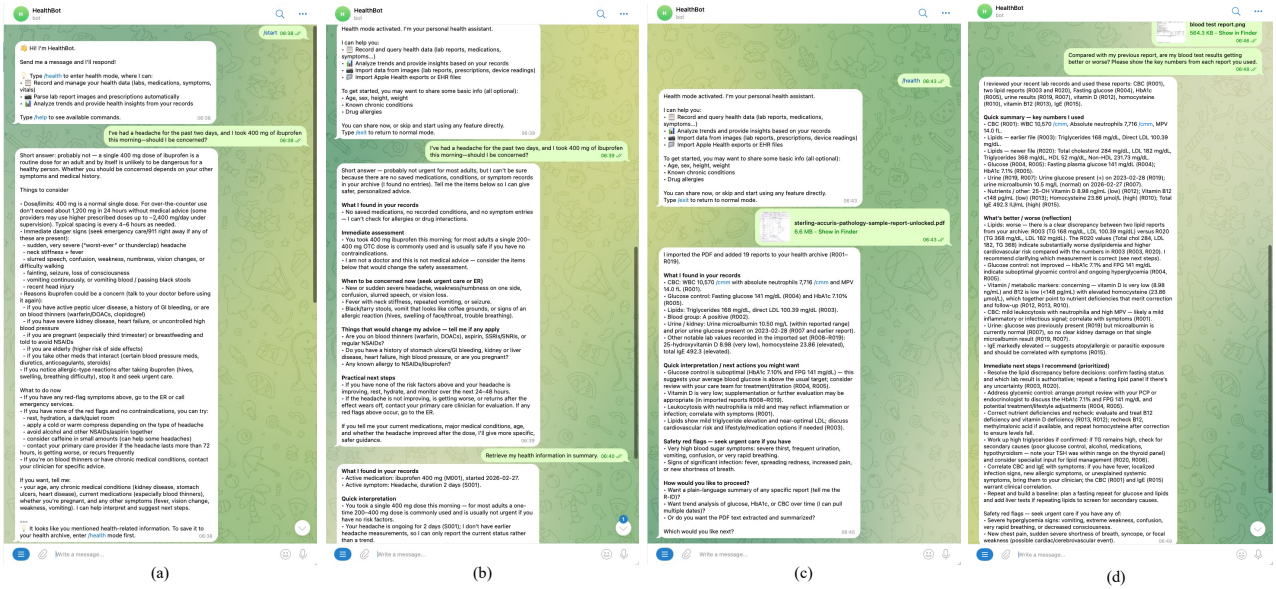
Condition	**Avg. Length**	**Accuracy (%)**
Raw History	6,215	66.0
H-Context	3,777	70.0
Change	**–39.2%**	**+4.0 pp**

3.2.3. Efficiency Advantage.

As shown in Table 4, H-Context reduces the history context by 39.2% in context length while simultaneously improving accuracy by 4.0 pp for GPT-5.2. This demonstrates that the three-layer structured compression $\mathcal{H} = \{P, T, S\}$ effectively distills clinically salient information from raw records, removing redundant or noisy data that may distract the model. We attribute H-Context's advantage over raw history to three complementary mechanisms: (1) the *Profile* layer provides a structured patient summary that foregrounds chronic conditions and demographics—information that is buried across multiple raw encounters; (2) the *Timeline* layer reformats history into a standardized schema with explicit diagnostic fields, reducing parsing burden on the LLM; and (3) the *Summary* layer synthesizes cross-encounter patterns (e.g., recurring abnormal labs, persistent symptoms, or medication-response changes) that are difficult to extract from raw chronological data.

4. Qualitative Results

Fig. 2 demonstrates HealthBot's messaging-based workflow. Panels (a–b) illustrate the mode-switching and memory-isolation design. In (a), HealthBot answers a health-related question in general mode using the same lightweight behavior as ordinary chats: it does not write to or query the dedicated health archive, but prompts the user to opt in and switch to Health mode if they want the information recorded. In (b), after switching to Health mode, HealthBot constructs structured symptom/medication records (S001/M001) and stores them in the health archive, then grounds its response by retrieving the stored records with provenance, enabling more evidence-based answers. Importantly, the conversation histories for general mode and Health mode are stored separately to avoid cross-contamination of session memory. Panels (c–d) illustrate an end-to-end longitudinal pipeline across heterogeneous inputs: in (c) the user imports a complete health check PDF from Feb 2023, which HealthBot parses into structured report entries (R001–R019); in (d) the user uploads a blood test image from Dec 2023, which is ingested as an additional report (R020). HealthBot then answers a "better or worse" longitudinal question by retrieving prior records from the archive, aligning biomarkers over time, and explicitly listing supporting values and sources via report IDs.



(a) (b) (c) (d)

Figure 2. Qualitative visualization of HealthBot interactions via Telegram. *Screenshots are shown at reduced scale; readers may zoom in for full text.*

5. Conclusion

We presented HealthBot, an open-source personal health assistant that converts heterogeneous inputs (conversations, report images/PDFs, EHR files, and wearable exports) into a structured, locally stored longitudinal health archive and supports tool-grounded reasoning. To keep long histories usable under tight context budgets, we introduced Hierarchical Health Context (H-Context), a three-layer representation (Profile, Timeline, Summary) that retains clinically relevant information with record-level provenance. On a diagnosis prediction benchmark spanning synthetic (Synthea) and real-world (MIMIC-IV) trajectories, incorporating longitudinal history improves accuracy over current-only inputs, and further structuring raw history with H-Context yields additional gains while reducing history context length by 39% across both GPT-5.2 and Qwen3-8B. Qualitative examples further demonstrate opt-in Health mode with isolated session memory, multimodal ingestion into structured records, and evidence-backed longitudinal comparisons with explicit report IDs.

Impact Statement

This work studies open-source personal health agents for longitudinal health management. While such systems may help users organize records and prepare for consultations, they also raise risks of inaccurate advice, over-reliance, privacy leakage, and use without clinical oversight. HealthBot is a research prototype, not a diagnostic tool; deployment would require validation, consent, privacy protection, uncertainty communication, and human oversight.

References

Abbasian, M., Khatibi, E., Azimi, I., Jain, R., and Rahmani, A. M. Conversational health agents: A personalized large language model-powered agent framework. *JAMIA Open*, 8(4):ooaf067, 2025. doi: 10.1093/jamiaopen/ooaf067.

Chen, Z., Hernández Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., K"opf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. MEDITRON-70b: Scaling medical pretraining for large language models, 2023.

Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J. M. J., Youssef, A., Cohen, J. P., Reis, E. P., et al. CheXagent: Towards a foundation model for chest x-ray interpretation, 2024.

HKUDS. NanoBot: The ultra-lightweight personal AI agent. https://github.com/HKUDS/nanobot, 2026. Accessed: 2026-05-09.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W.

MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79410–79452, 2024.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., and Wang, Y. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.510.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.-Q., Ma, W., and Liu, Y. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2024b.

Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., Song, X., Tang, S., Xiao, J., Lin, H., Zhuang, Y., and Ooi, B. C. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37975–37995. PMLR, 2025.

Maity, S. and Saikia, M. J. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631, 2025. doi: 10.3390/bioengineering12060631.

Merrill, M. A., Paruchuri, A., Rezaei, N., Kovacs, G., Perez, J., Liu, Y., Schenck, E., Hammerquist, N., Sunshine, J., Tailor, S., Ayush, K., Su, H.-W., He, Q., McLean, C. Y., Malhotra, M., Patel, S., Zhan, J., Althoff, T., McDuff, D., and Liu, X. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17:6070, 2026. doi: 10.1038/s41467-025-67922-y.

OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2026-05-09.

OpenAI. Introducing ChatGPT health. https://openai.com/index/introducing-chatgpt-health/, 2026. Accessed: 2026-05-09.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., Chaves, J. Z., Hu, S.-Y., Schaekermann, M., Kamath, A., Cheng, Y., Barrett, D. G. T., Cheung, C., Mustafa, B., Palepu, A., McDuff, D., et al. Capabilities of Gemini models in medicine, 2024.

Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., and Moor, M. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments, 2024.

Shi, W., Xu, R., Zhuang, Y., Yu, Y., Zhang, J., Wu, H., Zhu, Y., Ho, J., Yang, C., and Wang, M. D. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1245.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S. S. M., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. y., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Roelofs, R., Caine, B., Barral, J., Hassabis, D., Kavukcuoglu, K., Manyika, J., Dean, J., Karthikesalingam, A., and Natarajan, V. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7.

Steinberger, P. OpenClaw: Personal AI assistant. https://github.com/openclaw/openclaw, 2026. Accessed: 2026-05-09.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and

McLachlan, S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3): 230–238, 2018. doi: 10.1093/jamia/ocx079.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024. doi: 10.1093/jamia/ocae045.

Wu, X., Huang, T.-Z., Deng, L.-J., Qiao, Y., Razzak, I., and Xie, Y. A knowledge-driven adaptive collaboration of LLMs for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33495–33512, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1699.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, M., Li, M., Xue, M., Li, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen3 technical report, 2025.

Ye, J. and Tang, H. Multimodal large language models for medicine: A comprehensive survey, 2025.

Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., Wan, X., Wang, B., and Li, H. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10859–10885, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.725.

Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., Zou, J., Li, Q., Liu, H., and Sun, L. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024. doi: 10.1038/s41591-024-03185-2.

Zuo, K., Jiang, Y., Mo, F., and Lio, P. KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. In *Proceedings of the First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pp. 195–206. PMLR, 2025.