

## Article

# PHIA+: An Evidence-Grounded and Context-Efficient AI Agent Framework for Wearable Data Interpretation

Chenyu He , Yiyang Li, Yang Jin and Shiqi Zhao \* 

School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; 202312530@stu.neuq.edu.cn (C.H.); 202312561@stu.neuq.edu.cn (Y.L.); 202312566@stu.neuq.edu.cn (Y.J.)

\* Correspondence: zhaoshiqi@neuq.edu.cn

## Abstract

Wearable devices make it easy to collect long-term records of activity, sleep, heart rate, and other behavioral signals. However, translating these data into reliable personal health insights remains difficult. ReAct-style large language model (LLM) agents can compute over wearable records, yet two issues remain: health interpretations may rely on weak evidence, and long sessions accumulate tool traces that increase context pressure. We present PHIA+, an agent framework designed to address these issues. It introduces LLM-guided medical retrieval, which improves evidence reliability by selecting source-aware passages from a medical corpus, and summary-based hierarchical memory, which reduces context pressure by compressing completed reasoning traces while preserving the current one. We evaluate PHIA+ on the PHIA benchmark with objective and open-ended protocols. In the stacked objective question answering (QA) study, hierarchical memory reduces context length by a factor of 11.8 and API token use by 56.8%, while keeping final accuracy not significantly different from full-history reasoning (within 0.22 percentage points). In the open-ended QA study, under five-pass anonymized LLM judging, the RAG-augmented PHIA attains the highest mean medical source reliability score (2.92 out of 5) and the highest mean response quality score (3.84 out of 5) among five retrieval ablations. These results suggest that explicit mechanisms for evidence retrieval and long-session memory can improve source reliability and context efficiency in wearable-health agents, though the reliability of medical retrieval remains bounded by the coverage of the open-access corpus. These findings are useful to developers and researchers building practical wearable-health assistants.

**Keywords:** large language model agents; wearable health data; retrieval-augmented generation; evidence reliability; hierarchical memory; context efficiency; benchmark evaluation; health informatics



Academic Editor: Giuseppe Andreoni

Received: 1 June 2026

Revised: 9 July 2026

Accepted: 13 July 2026

Published: 19 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

Consumer wearable devices generate continuous, multidimensional, and longitudinal records of activity, sleep, heart rate, and other behavioral signals. These data can support personal health insights, but current large language models (LLMs) are not well suited to answering numerically precise questions over high-resolution wearable time-series data. Even a seemingly simple question, such as whether sleep improves after exercise, may require checking data availability, selecting appropriate activity and sleep metrics, comparing days across time, and interpreting the result with relevant health knowledge.

Personal Health Insights Agent (PHIA) [1] addressed this problem by formulating wearable data interpretation as a tool-using LLM agent task. Its ReAct-style agent combines

multistep reasoning with Python v3.11 code execution, allowing deterministic analysis over wearable-derived records rather than relying only on pre-aggregated summaries or fixed templates. It also uses web search when external health knowledge is needed for open-ended interpretation. Yet two issues remain when extending PHIA-style benchmark interactions toward a more realistic wearable-health information system.

**Evidence grounding.** Many user questions ask not just what the records show, but what those patterns may mean. For example, a sleep or activity pattern may require comparison with behavioral health knowledge, clinical ranges, or risk-aware advice. A generic web search can help, but its results are uncontrolled and vary in reliability. Search results may include consumer health pages, commercial content, outdated summaries, or sources with unclear review status. When health advice rests on such weak or uncontrolled evidence, the consequences can be serious. Recent physician-led evaluation shows that current LLM chatbots still return a non-trivial fraction of unsafe answers to patient-posed medical questions, with responses that could lead to serious harm if acted on [2]. The risk runs in both directions: automated self-triage tools have been found to under-triage genuinely urgent conditions, which can give a user false reassurance and delay needed care [3]. Fluent, confident phrasing compounds these risks, since an authoritative but unsupported claim is easily over-trusted.

**Context efficiency.** A wearable-health assistant is not a one-shot calculator. Users may return with follow-up questions and evolving health concerns. If the agent discards prior interactions, its analysis cannot build on accumulated findings and becomes less personalized. However, if it retains every previous question, code action, observation, and response in full, the context grows into a long and noisy history. Such history can distract reasoning from the current question and, in longer sessions, may even exceed the model's context window.

Moving from isolated benchmark questions toward genuinely useful, clinically relevant assistance raises the bar for what such an agent must do. Because its output can influence real health decisions, its advice has to rest on trustworthy medical evidence rather than fluent but unsupported text, and it should signal how far that evidence can be trusted. Because health concerns unfold over weeks and months, the assistant has to sustain accurate, personalized reasoning across many sessions without exhausting the model's context. Reliable evidence grounding and efficient long-session context are therefore not incidental features but prerequisites for any dependable wearable-health assistant.

We introduce PHIA+, an applied AI agent framework for interpreting wearable biomedical and behavioral records. PHIA+ keeps PHIA's ReAct-style loop for tool-mediated personal data analysis. Additionally, for open-ended advice, it adds LLM-guided medical retrieval to improve evidence grounding and source reliability. For longer sessions, it adds summary-based hierarchical memory (HiMem) to reduce accumulated context while preserving the current reasoning trace. This combination is also what distinguishes PHIA+ from existing retrieval-augmented healthcare agents: it couples active, source-aware evidence selection with long-session memory, so that grounding is sustained across many turns of personal wearable data analysis rather than supplied once for an isolated clinical question.

We frame the work around three research questions (RQs):

- RQ1.** Can summary-based hierarchical memory reduce context and token consumption in long wearable data analysis sessions while preserving objective answer accuracy?
- RQ2.** Does medical retrieval augmentation improve the grounding and reliability of open-ended wearable-health advice?
- RQ3.** Which retrieval components contribute most to answer quality: external search, local medical evidence retrieval, or confidence-aware evidence filtering?

This paper makes three main contributions.

First, to improve grounding in open-ended wearable-health advice, we design LLM-guided medical retrieval-augmented generation (RAG). It narrows retrieval to relevant health topics, reformulates the retrieval query, filters candidate passages, and compresses selected evidence into a concise observation with source and confidence metadata. As a result, the ReAct agent receives filtered, source-aware evidence observations instead of an uncontrolled list of retrieved passages.

Second, to support longer personalized sessions, we implement a summary-based hierarchical memory protocol that compresses completed reasoning traces into concise summaries while preserving durable information and keeping the current subgoal trace intact.

Third, we provide an empirical evaluation of the PHIA benchmark that separates the two intended roles of PHIA+. The stacked objective question answering (QA) study addresses RQ1 and shows that HiMem reduces context length and token use while preserving objective answer accuracy within a small margin of full-history reasoning. The open-ended QA study addresses RQ2 and RQ3 and shows that the RAG-augmented PHIA framework attains the highest mean medical source reliability score and the highest mean response quality score among five retrieval configurations.

The remainder of this paper is organized as follows. Section 2 reviews related work on wearable-health agents, medical retrieval-augmented generation, and memory management. Section 3 presents the PHIA+ method, covering the ReAct-based problem formulation, the wearable data interface, LLM-guided medical RAG, confidence-aware retrieval, and hierarchical memory. Section 4 reports the experimental setup and results for the stacked objective QA and open-ended QA studies. Section 5 discusses the findings and their implications, and Section 6 concludes the paper and outlines directions for future work.

## 2. Related Work

### 2.1. LLM Agents for Wearable Health Data Analysis

Wearable-health questions often require more than direct text generation: the system must inspect personal records, run calculations, decide whether external evidence is needed, and explain the result to a user. ReAct-style agents are a natural fit for this setting because they interleave reasoning with tool calls [4], complementing broader work on self-directed tool use [5] and feedback-based agent refinement [6]. Surveys of LLM-based autonomous agents describe agent design in terms of profiling, memory, planning, and action [7].

In wearable-health settings, PHIA [1] showed that an agent can analyze wearable-derived records through code execution and produce user-facing answers. Related wearable-health LLM systems have explored domain-specific model adaptation, health prediction from wearable sensors, and interactive personalized insight generation [8–10]. A recent study has examined hybrid machine learning (ML)–LLM and RAG-enhanced frameworks for explaining wearable-sensor fatigue predictions [11]. These systems show that LLM-based interfaces can make personal biomedical data more usable, but they leave open how an agent could obtain reliable medical evidence and manage accumulated reasoning traces during longer interactions.

A separate line of medical agent research studies agentic decision-making in healthcare settings. MedAgents and MDAgents study multi-agent collaboration for medical reasoning and decision-making [12,13], while AgentClinic and MedAgentBench evaluate LLM agents in interactive healthcare environments beyond single-turn QA settings [14,15]. These studies motivate behavior-level evaluation for health agents, but PHIA+ targets a different setting: interpreting wearable-derived personal health records and supporting longer personalized interactions.

## 2.2. Medical Retrieval-Augmented Generation

Open-ended health advice needs evidence that is both relevant and reliable. RAG can provide literature-backed context [16], and recent healthcare RAG surveys emphasize its value for reducing unsupported medical claims [17]. At the same time, retrieval can introduce noise, irrelevant evidence, or misleading passages. Self-RAG addresses this problem through adaptive retrieval and self-critique mechanisms [18], while clinical RAG studies show that query clarity and structure strongly affect retrieval quality [19]. Benchmarking work further shows that medical RAG must be evaluated systematically: MIRAGE, a large-scale medical RAG benchmark, finds that the choice of corpus, retriever, and backbone model strongly affects reliability and even reproduces “lost-in-the-middle” effects in the medical setting [20]. Prior work has also explored multi-agent RAG pipelines and confidence-aware medical collaboration [21,22].

PHIA+ follows this controlled retrieval direction in a wearable-health setting, where answers must often combine personal data analysis with medical evidence. Unlike passive top-*k* context injection, PHIA+ positions medical RAG as an evidence selection layer between a medical corpus and the wearable-health agent.

## 2.3. Context and Memory Management

Long-context reasoning remains challenging for LLM agents because tool traces, intermediate observations, and repeated instructions accumulate over time. Reflexion shows that textual feedback and memory can alter later agent decisions [6]. MemoryBank augments an LLM with a long-term memory that stores and retrieves past interactions and updates what it retains over time, so that relevant history can be recalled beyond a single context window [23]. HiAgent studies hierarchical memory for long-horizon agent tasks, where lower-level interaction traces can be summarized into higher-level memory states [24].

This line of work demonstrates that memory can be managed explicitly rather than left as an unbounded prompt history. Wearable-health interactions add a more specific requirement: the agent should retain durable personal findings while avoiding repeated replay of code, observations, and correction traces. At the same time, compressing the active reasoning process too early can discard the exact numeric values or evidence-bearing details needed to solve the current question. PHIA+ adapts hierarchical memory to this setting by compressing completed subgoal traces into concise summaries while keeping the current reasoning trace intact.

A complementary trend addresses long histories by enlarging the context window itself; methods such as LongLoRA extend pretrained models to substantially longer inputs at limited additional cost [25]. A larger window alone, however, does not remove the problem: models use long contexts unevenly, with retrieval accuracy dropping for information placed in the middle of a long input even in explicitly long-context models [26], and the token cost and latency of each turn still grow with the retained history. This cost is magnified at deployment scale: wearable devices serve very large user populations, so an approach whose per-turn context—and thus, its token and compute cost—keeps growing with session length becomes hard to sustain once multiplied across many concurrent, long-running sessions. Bounding what re-enters the context, rather than relying on an ever-longer window, is therefore the more scalable strategy for long wearable-health sessions.

## 2.4. Multimodal Healthcare Agents

A prominent recent direction equips medical AI with multimodal perception, so that a system can jointly reason over clinical text and images. Med-Flamingo adapts a multimodal few-shot learner to medicine, continued-pretrained on interleaved medical image–text data so that it can answer visual medical questions from only a few examples [27]. LLaVA-

Med aligns a general vision–language model to the biomedical domain using large-scale figure–caption data and instruction tuning, yielding an open-ended biomedical image assistant [28]. Building on such multimodal backbones, agentic systems add tool use and planning: MMedAgent learns to invoke specialized medical tools across several imaging tasks and modalities, letting one agent route a query to the most suitable expert model [29]. Recent surveys map this fast-moving landscape of medical LLM agents and identify multimodal integration, reliability, and deployment as open challenges [30].

These systems target rich clinical perception—chiefly medical imaging—whereas PHIA+ addresses a different modality and setting. It reasons over wearable-derived numeric time series rather than images, and its contributions are orthogonal to multimodal perception: reliable medical-evidence grounding for open-ended advice and context-efficient memory across long personal sessions. Multimodal clinical perception and PHIA+’s evidence-and-memory focus are complementary, and a future assistant could combine both.

Table 1 compares the design scope of PHIA+ with related health agent, RAG, and memory systems.

Qualitatively, these systems occupy different parts of the design space. The retrieval-oriented systems (MA-RAG, MedCoAct) surface external evidence but do not compress long interaction histories; the memory-oriented system (HiAgent) compresses histories but does not ground answers in curated medical evidence; and the perception- or collaboration-oriented medical agents (MMedAgent, MedAgents) add multimodal tools or multi-agent reasoning without either mechanism. PHIA+ is the only entry that brings controlled medical retrieval, an explicit confidence signal, and summary-based memory compression together in a single wearable-health agent.

**Table 1.** Positioning of PHIA+ relative to related health agent and memory/RAG systems.

| System         | Focus/Setting                             | RAG | Confidence Signal <sup>1</sup> | Memory Compression |
|----------------|-------------------------------------------|-----|--------------------------------|--------------------|
| PHIA [1]       | Wearable data code agent                  | ✗   | ✗                              | ✗                  |
| openCHA [31]   | Personalized health agent framework       | ✓   | ✗                              | ✗                  |
| MedAgents [12] | Multi-agent medical reasoning             | ✗   | ✗                              | ✗                  |
| MMedAgent [29] | Multimodal clinical tool use              | ✗   | ✗                              | ✗                  |
| MA-RAG [21]    | Multi-agent RAG                           | ✓   | ✗                              | ✗                  |
| MedCoAct [22]  | Confidence-aware medical collaboration    | ✓   | ✓                              | ✗                  |
| HiAgent [24]   | Long-horizon agent memory                 | ✗   | ✗                              | ✓                  |
| PHIA+          | Wearable data, medical RAG, long sessions | ✓   | ✓                              | ✓                  |

<sup>1</sup> The confidence signal column denotes whether a system uses an explicit confidence mechanism for RAG. In PHIA+, this signal is a supporting mechanism rather than a standalone module. Section 3.4 describes its role in detail.

### 3. Methods

#### 3.1. ReAct-Based Problem Formulation

We formulate PHIA+ as a ReAct-based agent process for interpreting wearable-derived biomedical and behavioral records. Given a user question  $q$  and the individual’s wearable-derived health records  $D_u$ , the agent interleaves reasoning, tool selection, observation, and answer synthesis. A trajectory can be written abstractly as

$$\tau = (q, h_0, a_1, o_1, h_1, \dots, a_T, o_T, y), \quad (1)$$

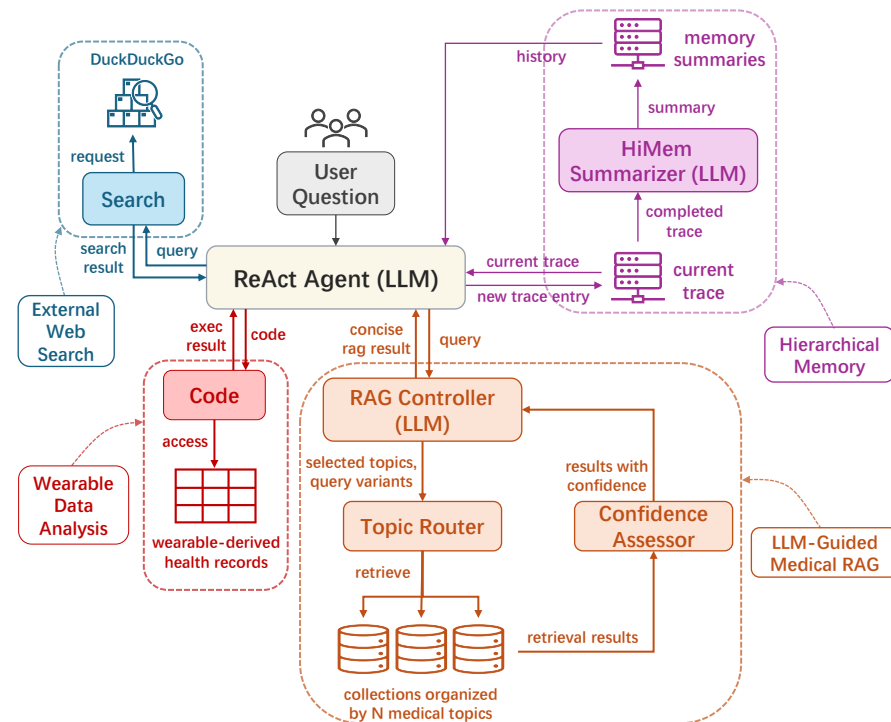
Equation (1) defines the interaction trajectory  $\tau$  for a user question  $q$ , starting from the initial reasoning state  $h_0$ . Here,  $h_t$  denotes the agent’s intermediate reasoning state,  $a_t$  is a tool action,  $o_t$  is the observation returned by that action, and  $y$  is the final answer. The two PHIA+ mechanisms operate on different parts of the trajectory: RAG changes the

evidence available through actions, while HiMem changes how completed trajectories are represented in later turns.

PHIA+ operates over the following action space:

- **Code:** Executes deterministic computation over the user's wearable-derived records.
- **Search:** Retrieves open-web evidence through DuckDuckGo.
- **Rag:** Queries LLM-guided medical RAG over the local medical literature corpus.

Figure 1 gives the method overview. Algorithm 1 describes the control flow.



**Figure 1.** Architecture of PHIA+. The ReAct agent accesses wearable-derived personal health records (Code), LLM-guided medical evidence retrieval (RAG), and DuckDuckGo web search (Search). The confidence assessor provides retrieval quality signals to support RAG filtering. HiMem maintains the current trace in full and stores completed traces as memory summaries.

#### Algorithm 1 PHIA+ ReAct control flow

**Require:** User question  $q$ , wearable-derived records  $D_u$ , memory summaries  $M$

- 1 Initialize the active ReAct trace for the initial subgoal:  $T \leftarrow \emptyset$
- 2 **while** the user request has not been resolved **do**
- 3   Build the agent context from  $q$ , data-interface information,  $M$ , and  $T$
- 4   Agent selects an action from CODE, RAG, SEARCH, or answer synthesis
- 5   **if** CODE is selected **then**
- 6     Execute the agent-generated code over  $D_u$  and append the observation to  $T$
- 7   **else if** RAG is selected **then**
- 8     Run the RAG pipeline and append the evidence observation to  $T$
- 9   **else if** SEARCH is selected **then**
- 10    Retrieve open-web evidence and append the observation to  $T$
- 11   **else**
- 12     Generate the response from the current context and append it to  $T$
- 13   **end if**
- 14   **if** the agent marks the current subgoal as completed **then**
- 15     Compress the completed trace segment with HiMem and update  $M$
- 16     Reset the active ReAct trace for the next subgoal:  $T \leftarrow \emptyset$
- 17   **end if**
- 18 **end while**

### 3.2. Wearable Data Interface

The wearable data interface connects the ReAct agent to longitudinal records derived from consumer wearables, including daily summaries, activity sessions, sleep records, heart-rate summaries, and profile attributes. In the system, these records are organized as structured tables with fields such as dates, activity types, durations, sleep stages, and heart-rate measures. Rather than placing all raw tables directly in the prompt, PHIA+ gives the agent read-only analytical access through the code tool. The agent can generate code to inspect these fields, filter records, join relevant tables, compute statistics, and return observations to the reasoning trace. This interface separates semantic task interpretation from deterministic data operations: the model decides which variables and transformations are relevant to the user question, while numerical computation is performed by code over the structured records.

Concretely, each record is normalized to a common six-field schema—participant, timestamp, category, field, value, and unit—together with provenance metadata, so that heterogeneous wearable exports share one representation. The interface exposes heart-rate fields (per-minute beats per minute, daily resting heart rate, and time-in-zone minutes), activity fields (steps, distance, calories, sedentary and active minutes, and exercise sessions), and sleep fields (per-session minutes asleep with deep, light, REM, and wake minutes and sleep efficiency, and a daily sleep score). For example, a per-minute heart-rate sample is stored as (2020-03-10T08:15:00, heart\_rate, bpm, 72, bpm), a daily resting heart rate as (2020-03-10, heart\_rate, resting\_bpm, 58.4, bpm), and a sleep session as a structured value holding minutes asleep and the deep, light, REM, and wake minutes with sleep efficiency.

Wearable health data is sensitive, and the interface is designed to limit its exposure. The experiments use de-identified research benchmark records rather than identifiable clinical data, and the analytical interface keeps the raw longitudinal tables local: the code tool computes over them on the local machine and returns only the derived observations needed for the current question to the reasoning trace, so the full raw records are never placed in the prompt or sent to the LLM API. A production deployment would additionally require informed consent, encryption in transit and at rest, access control, and minimization of the derived features shared with any external model, which we treat as deployment prerequisites rather than contributions of this study.

### 3.3. LLM-Guided Medical RAG

For open-ended wearable-health interpretation questions, PHIA+ introduces an LLM-guided medical RAG module as a tool inside the ReAct loop. This module mediates between the outer ReAct agent and a local medical evidence corpus. It consists of topic-organized medical evidence collections, a biomedical vector retriever, and an LLM-guided RAG controller that rewrites queries, filters candidate chunks, and prepares the final RAG observation. In the current implementation, the topic-organized collections cover sleep, activity, heart, nutrition, weight, mental health, general health, and medical knowledge, while the topic schema remains extensible.

Through the RAG tool interface, the available medical topics are exposed to the ReAct agent; when invoking the module, the agent supplies a retrieval query and selects one or more topics. The RAG pipeline proceeds as follows:

1. **Topic scoping.** The RAG module receives the ReAct agent's tool input: a query string and a sequence of selected health topics. Let  $t$  denote the length of this topic sequence. The query specifies what health-related evidence should be retrieved, and the selected topics specify which medical evidence collections are relevant. This stage maps the selected topics to matching evidence collections and creates the scoped search space used by the later query variants.

2. **Query rewriting.** The RAG controller rewrites the user-facing query into multiple answer-oriented declarative query variants. For example, a query such as “How long should I exercise?” can be rewritten toward statements about recommended weekly moderate-intensity exercise duration for adults. Let  $n_q$  denote the number of query variants produced by this stage.
3. **Candidate retrieval.** Candidate retrieval is performed over the selected topic collections using biomedical embeddings. For each pair of query variant and selected topic, PHIA+ retrieves the top- $k$  chunks from the corresponding topic collection:

$$k = f(n_q, t), \quad (2)$$

where  $n_q$  is the number of query variants and  $t$  is the number of selected topics. Each candidate keeps provenance information such as topic, title, section, source identifier, chunk index, and relevance score.

4. **Evidence filtering and simplification.** The candidate chunks are indexed and concatenated into a chunk-pool message, which is provided to the RAG controller LLM. The controller selects useful chunks from the pool and compresses them into a concise RAG observation. The observation returned to the ReAct agent preserves selected sources and concise medical evidence.

The query-rewriting stage is motivated by the structure of wearable-health questions and by prior findings that clinical RAG effectiveness depends on how well user-facing questions are reformulated for retrieval [19]. Users often ask interrogative advice-seeking questions, while medical evidence passages are usually written as statements about recommended durations, thresholds, mechanisms, risks, or interventions. Directly using the original question can over-match similar question forms or broad discussions; declarative query variants make retrieval more likely to find passages that state the needed recommendation or mechanism. This use of query rewriting is supported by a broader line of evidence that transforming the user query with an LLM improves retrieval for retrieval-augmented generation: reformulating the query into a retrieval-oriented form yields consistent gains in a rewrite–retrieve–read framework [32], and expanding the query with LLM-generated answer-like content improves both sparse and dense retrieval by 3–15% on standard information retrieval benchmarks [33].

Topic scoping and evidence filtering control retrieval noise. Searching the entire medical corpus can return too many chunks that are semantically similar but substantively irrelevant, especially when a question uses broad health language. Restricting retrieval to selected topics narrows the candidate space. However, irrelevant chunks can still remain within selected collections. Filtering therefore keeps useful chunks from the returned candidates, preventing a long top- $k$  list containing irrelevant chunks from being passed directly into the main ReAct context.

The function  $f(n_q, t)$  in Equation (2) is defined as a retrieval budget schedule. In this work, we instantiate it as

$$f(n_q, t) = \max\left(1, \left\lfloor \frac{B}{n_q \sqrt{t}} \right\rfloor\right), \quad (3)$$

where  $B$  is a retrieval budget constant. Narrower topic selections allow greater retrieval depth within each topic, while broader topic selections reduce per-topic depth to avoid an oversized chunk pool. Increasing  $n_q$  similarly divides the retrieval budget across more query variants. As a result, the total number of retrieved chunks stays on the order of  $B\sqrt{t}$ , independent of the number of query variants  $n_q$  and growing only sublinearly with the number of selected topics  $t$ ; enlarging the query or topic set therefore does not sharply increase the retrieved chunk count. By bounding the candidate pool in this way, the

schedule limits three burdens that grow with the query and topic sets: the retrieval cost over the topic collections, the token cost of passing the pool to the RAG controller LLM, and the semantic difficulty of selecting the useful evidence from it.

The resulting RAG observation is concise and source-aware. For questions that require personal data interpretation, PHIA+ first establishes the relevant wearable-derived facts through code observations. RAG is then used when the answer requires medical or health-behavior evidence to interpret those facts.

### 3.4. Confidence-Aware Retrieval

MedCoAct uses confidence-aware reflection to evaluate retrieval quality and regenerate queries when retrieved evidence is insufficient or low quality [22]. However, its confidence assessment is still LLM-driven, which makes the signal dependent on internal model reasoning. PHIA+ therefore uses a retrieval-specific confidence signal based on measurable properties of the returned chunks. The signal captures substantive characteristics of the chunks, computed as follows in Equation (4):

$$C = 0.30S_{\text{semantic}} + 0.15S_{\text{richness}} + 0.55S_{\text{quality}}. \quad (4)$$

$S_{\text{semantic}}$  reflects how closely the retrieved chunks match the retrieval query. It is computed from the mean vector retrieval relevance score of the returned chunks, normalized by a target relevance value of 0.6 and clipped to  $[0, 1]$ .  $S_{\text{richness}}$  measures whether retrieval returns enough usable evidence, using  $\min(n_{\text{valid}}/3, 1)$ , where  $n_{\text{valid}}$  is the number of valid returned chunks. The score saturates at three chunks so that additional near-duplicate results do not keep increasing confidence.  $S_{\text{quality}}$  reflects whether the returned passages contain clinically useful evidence rather than only generic health language. It is computed by scanning the returned passages for evidence cues, including numeric health quantities, reference or recommended ranges, and guideline-, consensus-, clinical-, or recommendation-oriented language; the resulting cue score is clipped to  $[0, 1]$ .

The design targets one question—whether the retrieved evidence is trustworthy enough to rely on when forming health advice—and each term measures a necessary condition for that trust.  $S_{\text{semantic}}$  requires the evidence to be on-topic, since off-topic passages cannot support the answer whatever their content;  $S_{\text{richness}}$  requires enough evidence, since a single or near-empty result is a weak basis for a health claim; and  $S_{\text{quality}}$  requires clinically substantive content—numeric quantities, reference ranges, or guideline-like statements—rather than generic health prose. Relevance, sufficiency, and clinical substance are jointly necessary for retrieved evidence to be actionable, so the signal combines all three rather than relying on retrieval similarity alone, and it reads them directly from the returned chunks so that it stays transparent and independent of the generator LLM it is meant to check.

The weights and thresholds are design heuristics used in this study. We assign the largest weight to  $S_{\text{quality}}$  because semantic similarity alone can retrieve passages that match the query wording but contain little clinically useful evidence; for health advice, passages with numeric quantities, reference ranges, or guideline-like statements are more likely to support actionable and reliable responses. Semantic relevance receives the next-largest weight because a passage must first be on-topic for its evidence cues to be meaningful, whereas richness receives the smallest weight and acts only as a coverage floor: it rises with the number of valid chunks but saturates at three so that additional near-duplicate results cannot inflate confidence. The specific values (0.55, 0.30, 0.15) were set a priori as design heuristics rather than tuned on the evaluation data, so they do not confer an unfair advantage to any ablation.

After candidate retrieval, we use this confidence signal inside the RAG module as retrieval quality feedback for evidence filtering and simplification. The signal helps the RAG controller distinguish strong medical evidence from weak, sparse, or generic retrieval results before the selected chunks are compressed into the final RAG observation.

Algorithm 2 summarizes this confidence computation.

---

**Algorithm 2** Retrieval confidence score for a set of retrieved chunks

---

**Require:** Retrieved chunks  $R$  with relevance scores; target relevance  $r_0 = 0.6$ ; richness cap  $n_0 = 3$ ; weights  $(w_s, w_r, w_q) = (0.30, 0.15, 0.55)$

```

1  if  $R = \emptyset$  then
2    return 0
3  end if
4   $S_{\text{semantic}} \leftarrow \text{clip}(\text{mean-relevance}(R)/r_0, 0, 1)$ 
5   $n_{\text{valid}} \leftarrow \text{number of valid chunks in } R$ 
6   $S_{\text{richness}} \leftarrow \min(n_{\text{valid}}/n_0, 1)$ 
7   $c \leftarrow \text{evidence-cue score}(R)$   $\triangleright$  numeric quantities; reference or recommended ranges; guideline-, consensus-, or clinical-oriented language
8   $S_{\text{quality}} \leftarrow \text{clip}(c, 0, 1)$ 
9   $C \leftarrow w_s S_{\text{semantic}} + w_r S_{\text{richness}} + w_q S_{\text{quality}}$ 
10 return  $C$ 
```

---

### 3.5. Hierarchical Memory

HiMem is a memory management module for accumulated ReAct traces in realistic wearable-health interactions. During such an interaction, the ReAct agent can organize the work into user-facing analytical subgoals, such as answering a follow-up question, comparing a new observation with earlier findings, or producing an intermediate health interpretation. HiMem treats each completed subgoal trace as a memory compression unit. The current subgoal is kept in working memory as a full ReAct trace; only after the agent marks the subgoal as completed is that trace segment compressed and stored in summary memory.

The hierarchy therefore has two levels. The working memory level stores the current subgoal trace, while the summary-memory level stores compressed records of completed subgoals. For later subgoals, the model context is assembled by reading the summary memory records together with the current working trace, instead of replaying every previous trace in full.

The write operation into summary memory is implemented as an LLM-guided compression step. It takes as input the subgoal description together with the complete ReAct trace produced while solving that subgoal—the intermediate reasoning, tool actions, observations, and any conclusion or interim finding it yielded—as well as any control loop feedback issued for invalid actions or failed tool calls. A compression prompt then instructs the memory writer to keep only what may be reusable later—what was resolved, the resulting finding or conclusion, the key supporting evidence, and the reasoning behind it—while discarding long code blocks, verbose tool outputs, and transient trial-and-error details. The exact retained fields are task-dependent and should match the reusable information needed in the target wearable data analysis setting. For objective wearable data subgoals, this means preserving the resolved question intent, the final numeric finding, and the key computation details, such as filters and aggregations.

## 4. Experiments and Results

### 4.1. Dataset and Implementation

We use the PHIA benchmark [1]. The original benchmark contains wearable-derived health records for 16 individuals. Following PHIA's objective personal health insights evalu-

ation, the stacked objective QA study uses the four individuals with complete 1000-question sets: 41, 171, 333, and 465. Each objective question asks for a computable value from the wearable records and is paired with a numeric ground-truth answer, yielding 4000 objective questions in total.

For open-ended evaluation, we use the PHIA open-ended question set containing 172 canonical health questions. These questions call for natural-language responses rather than exact numeric outputs, so we evaluate them with the LLM-as-judge protocols defined for the open-ended QA study. We deliberately adopt the released PHIA benchmark unchanged rather than collecting new data, because the 172 open-ended questions are not an arbitrary small sample but a curated, high-quality set: following PHIA [1], they were distilled from roughly 3000 hypothetical queries solicited from personal and consumer health experts, categorized into nine query types, reduced to a 200-query subset for evaluation feasibility, and then deduplicated by semantic similarity into 172 distinct questions that were manually verified to cover all nine types and held out from agent development to avoid over-fitting. Their small number therefore reflects deliberate selection for quality rather than a shortage of data. Adopting this curated set unchanged also keeps our evaluation consistent with the original PHIA protocol and preserves its full nine-type coverage, while the paired objective study contributes 4000 questions with numeric ground truth so that accuracy preservation is measured at scale. This design follows two related evaluation trends: medical agent benchmarks increasingly evaluate tool-mediated, context-aware behavior beyond static QA [14,15], while medical RAG evaluation shows that retrieval quality strongly shapes the reliability of medical question answering [20].

All system LLM calls in the experimental runs use MiniMax-M2.7, a commercial large language model. This includes calls from the ReAct agent, RAG controller, and HiMem summarizer. We access the model through its API and keep the same backbone across all experimental conditions. No experiment fine-tunes the LLM. The stacked objective QA study evaluates HiMem in long wearable data analysis sessions, and the open-ended QA study evaluates medical retrieval configurations for natural-language wearable-health advice.

The RAG module is implemented with LangChain. Retrieved passages come from per-topic local Chroma vector stores; the indexed text is chunked at 512 tokens with 128-token overlap and embedded with the NCBI MedCPT-Query-Encoder, and the number of chunks returned per query follows the retrieval budget schedule of Equations (2) and (3). The medical corpus indexed in these stores is described with the open-ended QA setup (Section 4.3). HiMem is implemented as an in-process store with two levels: a trajectory store that keeps the full pre-summarization reasoning traces and a summary memory that keeps the one-line conclusion of each completed step; it uses no external database. External web search is served by DuckDuckGo. All experiments were run on a single laptop with an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB memory), with query embedding computed on the GPU.

With this design, every component runs locally except two remote services: LLM inference, served by the MiniMax-M2.7 API, and web search, served by DuckDuckGo. The main local footprint is the medical index built in this study, which occupies about 2.9 GB on disk as a one-time cost; once it is built, querying it is cheap, as top-five retrieval over a topic index of roughly twenty thousand chunks takes about 15 ms per query on the above GPU. In-memory summarization and trajectory management are dictionary and serialization operations, code actions run over each user's own records, and per-session memory remains at the kilobyte scale, so all of these are negligible next to a network LLM call. End-to-end latency and monetary cost are therefore dominated by LLM API calls, which are billed per token; this is exactly the cost that HiMem reduces, cutting API token usage by 56.8% and compressing context about 11.8-fold (Section 4.2.2).

## 4.2. Stacked Objective QA Study

### 4.2.1. Setting

The stacked objective QA study evaluates HiMem as a context-efficiency mechanism for objective analysis of wearable-derived biomedical and behavioral records. It adapts the independent question objective QA setting used in the PHIA study into a controlled long-task setting for hierarchical memory. In the original PHIA setting, each objective question is solved in isolation; here, the same type of numeric wearable data question is organized into consecutive stacks so that previous reasoning traces or memory summaries accumulate within a stack. The goal is to test whether the agent can keep accuracy close to full-history reasoning while substantially reducing the context and token cost of carrying previous interaction traces. We use objective QA because each answer has a numeric ground truth, which makes accuracy preservation measurable.

Each user's 1000 questions are split consecutively into 50 stacks of 20 questions:

$$\text{stack}_i = \text{questions}[20i : 20(i + 1)]. \quad (5)$$

We compare two memory management modes:

- **Full:** previous natural ReAct traces remain in context.
- **HiMem:** previous traces are replaced by memory summaries.

In both modes, memory resets between stacks. Within a stack, the agent receives one current question at a time. After each question is completed, the run records the intermediate answer for evaluation. Since the benchmark provides explicit question boundaries and a numeric answer target for every item, this protocol treats each question as one analytical subgoal; in HiMem mode, completion of each question triggers compression of its trace, and the prompt instructs the summarizer to preserve the question intent, final numeric answer, and a concise method description covering the key computation, decisive filters, or aggregation. The answer contained in the generated summary is also recorded for summary-error analysis. After all 20 questions are completed, the agent receives a final request to return the 20 answers from its available history. This final stack answer request probes a practical long-session requirement: after many completed subgoals have accumulated, the agent should still be able to recover earlier findings from the context or memory it has retained.

We report two kinds of measurements.

First, for accuracy preservation, we separate three failure points in long-task processing. **Intermediate accuracy** (Inter. Acc.) measures current-task solving ability: whether the agent answers the active question correctly when that question is completed. **Final accuracy** (Final Acc.) measures whether the completed answers can still be extracted from the retained history at the end of the stack. We treat a final/intermediate mismatch as a history-extraction failure. For HiMem, **summary errors** (Sum. Err.) measure whether compression preserves key answer information: they count cases where the intermediate answer is correct before compression, but the answer recorded in the HiMem summary is wrong.

Second, for context efficiency, we report token usage, token ratio, context length, and context compression, which are the main HiMem outcomes in the stacked objective QA protocol. Token usage is the sum of provider-reported total tokens across all successful LLM calls in the protocol, including the ReAct answer-generation calls, HiMem summary-generation calls when enabled, and the final stack-answer extraction call. Tokens/Q is this total divided by the number of objective questions. Context length is measured separately as the character length of the context provided to the agent.

#### 4.2.2. Accuracy and Context Efficiency

Table 2 reports stacked QA accuracy results and error-source analysis.

**Table 2.** Stacked objective QA accuracy and error source diagnostics.

| Mode         | User           | Questions   | Final Acc. (%) | Inter. Acc. (%) | Mismatch | Sum. Err. |
|--------------|----------------|-------------|----------------|-----------------|----------|-----------|
| Full         | 41             | 1000        | 78.6           | 78.6            | 0        | –         |
| Full         | 171            | 1000        | 81.4           | 81.5            | 1        | –         |
| Full         | 333            | 1000        | 80.3           | 80.3            | 0        | –         |
| Full         | 465            | 1000        | 77.6           | 77.8            | 2        | –         |
| <b>Full</b>  | <b>Overall</b> | <b>4000</b> | <b>79.47</b>   | <b>79.55</b>    | <b>3</b> | <b>–</b>  |
| HiMem        | 41             | 1000        | 77.4           | 77.4            | 0        | 0         |
| HiMem        | 171            | 1000        | 83.0           | 83.0            | 0        | 0         |
| HiMem        | 333            | 1000        | 77.9           | 77.9            | 0        | 0         |
| HiMem        | 465            | 1000        | 78.7           | 78.7            | 0        | 0         |
| <b>HiMem</b> | <b>Overall</b> | <b>4000</b> | <b>79.25</b>   | <b>79.25</b>    | <b>0</b> | <b>0</b>  |

Across 4000 questions, full-history PHIA achieves 79.47% final accuracy, while HiMem achieves 79.25%. The overall difference is 0.22 percentage points in favor of full history. This small gap indicates that, in this stacked objective QA setting, replacing completed traces with memory summaries preserves overall accuracy within a narrow margin. A McNemar paired test over the 4000 questions indicates that this difference is not statistically significant (final accuracy  $p = 0.68$ ; intermediate accuracy  $p = 0.57$ ), so replacing completed traces with summaries does not significantly change objective answer accuracy in this setting.

The user-level results are heterogeneous: HiMem improves users 171 and 465, but underperforms for users 333 and 41. Final/intermediate mismatches are rare in both modes, and the summary-error count is zero, indicating that we did not observe cases where a correct intermediate answer was rewritten as a wrong answer in the HiMem summary.

Table 3 reports the main contribution of HiMem in this protocol: context and token efficiency. In HiMem mode, completed traces are compressed into concise summaries with an average context compression ratio of about  $11.8\times$  at the character level. The context length decreases from 1840 to 150 characters per question (36,700 to 2900 per stack), accompanied by a reduction in API token usage from 167.4 M to 72.3 M tokens (41,800 to 18,100 per question).

**Table 3.** Context and token efficiency in stacked objective QA.

| Mode  | API Tokens | Tokens/Q | Token Ratio  | Ctx. Chars/Stack | Ctx. Chars/Q | Comp. Ratio  |
|-------|------------|----------|--------------|------------------|--------------|--------------|
| Full  | 167.4 M    | 41,800   | $1.00\times$ | 36,700           | 1840         | –            |
| HiMem | 72.3 M     | 18,100   | $0.43\times$ | 2900             | 150          | $11.8\times$ |

Together, Tables 2 and 3 show that HiMem substantially reduces context length and total API token consumption while preserving aggregate accuracy within a small margin of full-history reasoning.

#### 4.3. Open-Ended QA Study

##### 4.3.1. Setting

The open-ended QA study evaluates RQ2 and RQ3 using the 172 open-ended PHIA questions. Each question is run as an independent interaction: the context resets between questions, and the agent answers with a standard ReAct loop and a natural-language final response. The task evaluates how medical evidence retrieval and external search affect evidence-based open-ended advice.

Since open-ended QA has no single numeric ground truth, we use LLM-as-judge evaluation, also implemented with MiniMax-M2.7. The study compares five retrieval ablations. For each question, all five answers are placed in the same judge prompt and anonymized as A–E, so the judge compares the alternatives side by side without seeing method names. To reduce position bias, we repeat the judging five times with rotated answer order, so each ablation appears once in each label position; reported scores are averaged across the five passes. This anonymized, order-rotated protocol is used for both evidence-reliability evaluation and response quality evaluation. Unlike the objective questions, the open-ended answers have no numeric ground truth; the judge scores the five anonymized answers relative to one another within a single comparison, and the five passes rotate their order to control position bias rather than acting as independent replicates. We therefore summarize the open-ended results descriptively—mean scores, win counts, and evidence support category distributions over the 172 questions—and state all comparisons as differences in mean. Formal significance testing is applied where a fixed ground truth makes it well defined, namely the stacked objective QA study (Section 4.2.2, McNemar test).

Table 4 defines the five ablations. For the RAG-enabled ablations, the number of query variants in Equation (3) is set to  $n_q = 4$  and the retrieval budget constant is set to  $B = 15$ .

**Table 4.** Retrieval ablations for open-ended QA.

| Retrieval Ablation    | Tools Available     | Role in the Comparison                                                 |
|-----------------------|---------------------|------------------------------------------------------------------------|
| PHIA (baseline)       | code + search       | PHIA-style agent with wearable data computation and open-web evidence. |
| PHIA w/o search       | code                | Tests performance without any external evidence tool.                  |
| PHIA + RAG w/o search | code + RAG          | Tests local medical retrieval without fallback to open-web evidence.   |
| PHIA + RAG w/o conf.  | code + RAG + search | RAG without retrieval-confidence feedback.                             |
| PHIA + RAG (Ours)     | code + RAG + search | Full PHIA+ method with LLM-guided medical RAG and confidence feedback. |

The local corpus used in this study was built from open-access full-text articles in PubMed Central (PMC). We organized the corpus into eight topic collections: sleep, activity, heart, nutrition, weight, mental health, general health, and medical knowledge. For each topic, we used manually specified topic queries with the PMC open-access filter. The queries included guideline, consensus, clinical-practice, recommendation, systematic-review, and review terms to bias retrieval toward clinically useful evidence. The returned PMC records were downloaded as full-text XML through Europe PMC or PMC interfaces, grouped by topic, and chunked. The chunks were embedded with MedCPT-style biomedical embeddings [34] and indexed in local Chroma vector stores.

Table 5 summarizes the medical RAG corpus used for open-ended QA.

The open-ended QA study has two evaluation layers. The trace-based evidence reliability evaluation tests whether open-ended answers rely on weak or unsupported evidence. It uses the complete ReAct trace for each answer: code observations, RAG results, web search results, intermediate ReAct messages, and the final answer. The response quality evaluation scores the final natural-language answers.

**Table 5.** Medical RAG corpus used for open-ended QA.

| Topic          | Source Files | Chunks         |
|----------------|--------------|----------------|
| Heart          | 870          | 20,237         |
| Activity       | 800          | 17,964         |
| Sleep          | 800          | 16,919         |
| Mental health  | 800          | 18,151         |
| General health | 800          | 16,738         |
| Nutrition      | 800          | 18,374         |
| Weight         | 800          | 18,234         |
| Medical        | 800          | 16,733         |
| <b>Total</b>   | <b>6470</b>  | <b>143,350</b> |

#### 4.3.2. Trace-Based Evidence Reliability

The trace-based judge receives the same question and the five anonymized answers, and each answer is accompanied by its complete ReAct trace. The judge is instructed not to reward a method merely for calling RAG or search. Instead, it evaluates whether important final-answer claims are supported by evidence visible in that answer's own trace and whether that evidence is reliable. Personal data claims should be supported by code observations; medical, guideline, risk, or intervention claims should be supported by RAG or search evidence, or expressed with appropriate uncertainty.

The evidence judge reports two dimensions: **medical source reliability** (Source), measuring whether medical or health claims are supported by credible and relevant sources, and **trace claim grounding** (Claim), measuring whether important personal data claims, medical claims, and recommendations are supported by evidence visible in the trace. Wins count the number of questions on which a method obtains the highest mean score across source reliability and claim grounding, with ties for the highest score counted for all tied methods. It also assigns each answer to one evidence support and reliability category: evidence grounded, partially grounded, weakly supported, or unsupported/fabricated. Table 6 reports the aggregated five-pass trace-based judge results.

**Table 6.** Trace-based evidence reliability in open-ended QA.

| Retrieval Ablation    | Source        | Claim         | Grounded     | Partial | Weak  | Unsup.       | Wins      |
|-----------------------|---------------|---------------|--------------|---------|-------|--------------|-----------|
| PHIA (baseline)       | 2.8721        | <b>2.9942</b> | <b>28.49</b> | 28.49   | 14.53 | 28.49        | 42        |
| PHIA w/o search       | 2.7965        | 2.9209        | 26.16        | 25.00   | 20.93 | 27.91        | 41        |
| PHIA + RAG w/o search | 2.7930        | 2.8081        | 25.58        | 24.42   | 22.09 | 27.91        | 41        |
| PHIA + RAG (no-conf.) | 2.8791        | 2.8779        | <b>28.49</b> | 24.42   | 20.35 | <b>26.74</b> | 34        |
| PHIA + RAG (Ours)     | <b>2.9227</b> | 2.9035        | <b>28.49</b> | 27.91   | 15.12 | 28.49        | <b>44</b> |

Bold indicates the best value in each column: the highest Source, Claim, Grounded, and Wins, and the lowest Unsup. (unsupported/fabricated) rate. The intermediate Partial and Weak categories are not marked.

PHIA + RAG (Ours) attains the highest mean medical source reliability score and the largest number of wins, while PHIA (baseline) has the highest mean trace claim grounding score. Three methods share the highest evidence-grounded rate, and the unsupported/fabricated rate of PHIA + RAG (Ours) remains comparable to the baseline. This indicates that the local medical RAG module can improve the credibility of medical evidence sources, but stronger evidence sources do not automatically eliminate unsupported claims or make every final claim better grounded in the trace. In several questions, the baseline answer is more conservative and therefore avoids unsupported claims, while RAG-enabled answers sometimes introduce additional medical interpretation that must be integrated carefully.

### 4.3.3. Response Quality

The second open-ended judge evaluates response quality across the same retrieval ablations. It asks whether each ablation produces safer, more complete, more clearly reasoned, more actionable, and more personalized open-ended answers. The weighted dimensions are safety (25%), completeness (20%), reasoning and evidence integration (25%), actionability (20%), and personalization (10%).

Table 7 reports the five-pass judge results for response quality. PHIA + RAG (Ours) achieves the highest average score (3.8416), followed by PHIA w/o search (3.8223) and PHIA (baseline) (3.7835).

**Table 7.** Response quality from five-pass anonymized judging.

| Retrieval Ablation    | Avg.          | Safe.         | Comp.         | Reason.       | Action.       | Pers.         |
|-----------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| PHIA + RAG (Ours)     | <b>3.8416</b> | <b>4.4802</b> | <b>3.7360</b> | 3.6605        | <b>3.7465</b> | 3.0988        |
| PHIA w/o search       | 3.8223        | 4.4116        | 3.6988        | 3.6721        | 3.7221        | <b>3.1721</b> |
| PHIA + RAG (no-conf.) | 3.7889        | 4.3419        | 3.7267        | 3.6337        | 3.7047        | 3.0872        |
| PHIA (baseline)       | 3.7835        | 4.3686        | 3.6465        | <b>3.6733</b> | 3.6453        | 3.1465        |
| PHIA + RAG w/o search | 3.7372        | 4.3640        | 3.6186        | 3.5860        | 3.5942        | 3.0709        |

Bold indicates the highest value in each column.

The per-dimension results are mixed. PHIA + RAG (Ours) obtains the best safety, completeness, and actionability scores. However, it does not dominate every dimension: PHIA (baseline) has the highest reasoning score, and PHIA w/o search has the highest personalization score. Retrieved evidence appears to help the agent cover medical caveats and practical recommendations, but it can also add integration burden when the question is mostly answerable from the user's data.

Across the two evaluation layers, PHIA + RAG (Ours) attains the highest mean medical source reliability and the highest mean response quality score. However, it does not obtain the best trace claim grounding score, which suggests that using more reliable medical sources does not guarantee that every final claim is better grounded in the trace.

## 5. Discussion

### 5.1. Further Analysis of the Stacked Objective QA Study

The stacked objective QA results show that HiMem can greatly reduce context length while preserving the key information needed later in the interaction. In practical wearable-health use, this reduction can substantially extend the interaction horizon and reduce the risk of context-length issues. Beyond avoiding context-length issues, a longer usable history may also help the agent maintain a broader view of the user's historical characteristics and recurring personal patterns across interactions.

The final/intermediate mismatch results add another angle to history extraction. Full history has three mismatches, while HiMem has none. Although the mismatch count is small, this pattern is consistent with a long-context burden: as the retained context grows, earlier findings remain available but become harder to locate and extract. HiMem reduces this burden by replacing completed reasoning traces with shorter memory summaries.

The same raw history can also be useful in this particular benchmark, which helps explain the mixed user-level accuracy pattern. The objective questions are relatively simple and highly regular in form, so the full-history mode can often benefit from previous raw solving traces, including field choices, similar filters or aggregations, and earlier error corrections. In this setting, the accumulated history is not always pure noise; it can also act as a set of worked examples for later questions with a similar structure. HiMem removes many of these low-level procedural details, which yields large context and token savings

but can also remove calculation patterns that are useful for subsequent, similarly formatted questions. This helps explain why, for some users in this benchmark, the full-history mode can achieve slightly higher intermediate accuracy instead of being consistently degraded by long-context noise.

However, the conditions that make raw history useful in this benchmark differ from the long-term use case that motivates HiMem. In real wearable-health interactions, user requests vary widely in form and intent: a user can move from numeric summaries to interpretation, comparison, planning, risk discussion, or follow-up advice. Under this kind of task diversity, detailed solution patterns from earlier tasks are usually not applicable to the next task; they are more likely to add irrelevant context than useful guidance. Compared with full trace replay, memory summaries better match this setting by retaining durable user-specific findings and discarding task-specific details. The current result should therefore be read as evidence that HiMem can substantially reduce context and token cost while preserving aggregate accuracy in a controlled long-task setting. Its further value in real, task-diverse wearable-health use—such as denoising a genuinely heterogeneous accumulated history—is a reasoned expectation rather than a demonstrated benefit: because this benchmark is short and regular, establishing such practical-deployment benefits would require prospective user studies with real users and interactions, which we leave to future work.

Two aspects of HiMem warrant further study. First, the summarization granularity is currently fixed: each completed trace is compressed into a compact record at a uniform level of detail. Finer-grained summaries retain more procedural detail at a higher token cost, whereas coarser summaries save more context but risk discarding information needed later, so systematically varying this granularity is a natural extension. Second, replacing completed traces with summaries could, in principle, allow summarization errors to accumulate over very long sessions. In the stacked objective QA runs, we observed zero summary errors across all 4000 questions (Table 2), that is, no case where a correct intermediate result was rewritten into an incorrect summary. Because these benchmark questions are short and highly regular, however, they do not fully stress this failure mode; whether such errors remain negligible under the longer and more varied sessions of real-world use is an open question that we leave to future work.

### 5.2. Evidence Contributions of Medical RAG

The open-ended QA study shows the contribution of medical RAG from two perspectives. The trace-based evidence evaluation shows that PHIA + RAG (Ours) achieves the highest mean medical source reliability score and the largest number of wins. This supports the central motivation for adding a local medical RAG module: open-ended wearable-health advice should be grounded in reliable medical sources such as peer-reviewed literature, guidelines, and clinical consensus documents, rather than depending only on generic web evidence, which may include consumer health pages, commercial content, or sources with unclear review status and clinical context.

The response quality evaluation points to a different effect. PHIA + RAG (Ours) attains the highest mean response quality score and shows its clearest gains in safety, completeness, and actionability. This suggests that evidence passages selected from reliable medical sources help the agent better cover medical caveats, explanatory context, and practical recommendations.

### 5.3. Limitations of Medical RAG

The first limitation is corpus coverage. Because the local corpus defines the evidence boundary of RAG and is drawn entirely from open-access PMC full-text articles, its reach

is bounded by the scope and licensing of PMC. A more consequential limitation is that the corpus consists mainly of primary research articles and reviews rather than systematically curated clinical reference works; its retrieved evidence therefore cannot be assumed to provide textbook-level or clinical decision support coverage.

The corpus boundary helps explain why PHIA + RAG w/o search is the weakest method, scoring lowest on source reliability, claim grounding, and nearly all response quality dimensions. When the local corpus does not contain sufficiently useful evidence for a question, the retrieved passages can become extra trace context rather than usable support. Without a search fallback, that extra context may act as noise and make the final answer harder to ground. PHIA + RAG (Ours) mitigates this failure mode by allowing the agent to seek web evidence when local retrieval is insufficient, which helps explain why it avoids the broad performance drop observed in the no-search RAG ablation.

The ablation results also reveal the cost of adding evidence tools. On trace claim grounding, reasoning, and personalization, all three RAG-enabled variants score below the two non-RAG variants. This pattern suggests that retrieval introduces an information selection burden: the agent must now weigh wearable-derived observations, retrieved medical passages, and any web search results together. PHIA (baseline) has code and search but no local RAG, so it can focus more directly on the user's data and on externally retrieved evidence without reconciling additional local corpus passages. This may explain why it obtains the best claim-grounding score and the highest reasoning score. Moreover, PHIA w/o search has the opposite constraint: without RAG or search, it can only reason from the user's wearable-derived records, which likely pushes the answer toward user-specific interpretation and explains its strongest personalization score.

Overall, the limitation is not only retrieval itself, but also how retrieval is controlled. PHIA + RAG benefits when local medical evidence is useful and search can compensate for corpus gaps; it weakens when weak local passages are still carried into the trace and compete with evidence from the user's data or from web search. Future work should therefore improve corpus coverage, retrieval stopping criteria, passage filtering, and mechanisms for tying retrieved evidence back to code-derived personal findings.

#### *5.4. Comparison with the Literature*

We now relate the findings above to prior work; because PHIA+ is evaluated on wearable-derived personal data over long, stacked sessions, we compare qualitatively rather than by raw benchmark scores. The context and token reductions we report, achieved without a significant change in objective accuracy, are in line with the memory management literature, in which explicitly deciding what an agent retains—rather than replaying an unbounded history—is what keeps long interactions tractable [6,23,24]. Our observation that a longer raw history helps only marginally and unevenly is, in turn, the behavior predicted by the “lost-in-the-middle” analysis of long-context models [26], which is why we bound context by compression instead of relying on larger windows [25].

On the evidence side, the gains we observe from source-aware medical retrieval are consistent with healthcare RAG studies reporting that literature-grounded retrieval reduces unsupported medical claims [16,17], whereas the information selection cost we observe when retrieval is uncontrolled matches the same literature's central caution: retrieval helps only when its quality and control are managed, whether through corpus and retriever choice [20] or through self-critique and selective retrieval [18]. This is precisely the gap PHIA+ addresses with confidence-aware selection and a web search fallback.

### 5.5. Evaluation and Deployment Boundaries

The open-ended QA study relies on an LLM-as-judge evaluation because the PHIA open-ended questions do not have single numeric ground-truth answers. The anonymized five-pass protocol reduces label-order bias, but it does not remove all evaluator bias and cannot be treated as equivalent to expert clinical review. The judge can compare answer quality and evidence use within the provided traces, but it cannot establish clinical correctness, patient safety, or real-world usefulness in the way that clinician review or prospective user studies would. Accordingly, we make no claim about the clinical safety or effectiveness of PHIA+; establishing either would require expert clinical evaluation and prospective user trials that lie beyond the scope of this benchmark study.

PHIA+ is an informational biomedical AI system for wearable-health data interpretation, not a clinically validated diagnostic or treatment system. Its outputs are intended to support explanation, reflection, and evidence-aware interpretation of wearable-derived records. Deployment in health or wellness products would require additional safeguards, expert review of advice quality, validation with broader user populations, and careful handling of privacy, liability, and escalation when a user's data may indicate clinically relevant risk.

## 6. Conclusions

We introduced PHIA+, a ReAct-based applied AI framework for wearable data interpretation that couples LLM-guided medical retrieval with summary-based hierarchical memory. Our experiments answer the three research questions set out in Section 1.

For RQ1, the stacked objective QA study shows that hierarchical memory reduces context length by a factor of 11.8 and API token usage by 56.8% while keeping final accuracy statistically indistinguishable from full-history reasoning (within 0.22 percentage points). For RQ2, the open-ended QA study shows that medical retrieval augmentation improves grounding and reliability: PHIA+ attains the highest mean medical source reliability score (2.92 out of 5) and the highest mean response quality score (3.84 out of 5) among five retrieval configurations. For RQ3, the ablation comparison shows that local medical retrieval contributes most to source reliability and answer quality, but chiefly when combined with confidence-aware evidence filtering and a web search fallback that compensates for gaps in the open-access corpus, whereas uncontrolled retrieval adds an information selection burden.

The two mechanisms are complementary rather than competing, addressing different bottlenecks—HiMem, the growth of context over long sessions, and medical RAG, the evidence gap in open-ended advice, where passages drawn from reliable medical sources help the agent cover caveats, explanatory context, and practical recommendations. The results distinguish medical source reliability from trace claim grounding: stronger source reliability does not necessarily imply stronger trace support for all final-answer claims.

PHIA+ advances evidence-aware and context-efficient wearable-health information systems, while remaining outside the scope of clinically validated diagnosis or treatment. These results are useful to developers and researchers building practical wearable-health assistants that must sustain long, continuous sessions in which reasoning traces accumulate toward the model's context limit and that must ground health advice in reliable medical sources rather than generic web content—settings where the context efficiency of HiMem and the evidence reliability of medical RAG are the primary concerns. Building on these findings, we see four concrete directions for future work: (i) expanding the medical corpus beyond open-access PMC articles toward systematically curated clinical reference sources, to raise the reliability ceiling of retrieval; (ii) developing adaptive retrieval-stopping and passage-filtering criteria, together with tighter mechanisms for tying

retrieved evidence to code-derived personal findings; (iii) systematically studying how summarization granularity trades off memory efficiency against the retention of procedural detail; and (iv) conducting prospective, larger-scale validation with clinician review, real user interactions, and clinically and demographically diverse populations. These steps are prerequisites for reliable deployment in higher-stakes health or wellness settings.

**Author Contributions:** Conceptualization, S.Z. and C.H.; methodology, S.Z. and C.H.; software, C.H.; validation, Y.L., Y.J. and C.H.; formal analysis, Y.L.; investigation, Y.J.; resources, S.Z.; data curation, C.H.; writing—original draft preparation, C.H.; writing—review and editing, S.Z.; visualization, Y.L.; supervision, S.Z.; project administration, S.Z.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (No. 62404037), and Hebei Natural Science Foundation (No. F2024501017), and the Fundamental Research Funds for the Central Universities (No. N2623033).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The PHIA benchmark data analyzed in this study are publicly available in the official PHIA repository at <https://github.com/yahskapar/personal-health-insights-agent>, accessed on 27 February 2026.

**Acknowledgments:** During the preparation of this manuscript, the authors used GPT-5 for language editing, code assistance, and manuscript formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

## Abbreviations

The following abbreviations are used in this manuscript:

|       |                                   |
|-------|-----------------------------------|
| AI    | Artificial intelligence           |
| API   | Application programming interface |
| HiMem | Hierarchical memory               |
| LLM   | Large language model              |
| ML    | Machine learning                  |
| PHIA  | Personal Health Insights Agent    |
| PMC   | PubMed Central                    |
| QA    | Question answering                |
| RAG   | Retrieval-augmented generation    |
| ReAct | Reasoning and acting              |
| RQ    | Research question                 |

## References

1. Merrill, M.A.; Paruchuri, A.; Rezaei, N.; Kovacs, G.; Perez, J.; Liu, Y.; Schenck, E.; Hammerquist, N.; Sunshine, J.; Taylor, S.; et al. Transforming wearable data into personal health insights using large language model agents. *Nat. Commun.* **2026**, *17*, 1143. [[CrossRef](#)] [[PubMed](#)]
2. Draelos, R.L.; Afreen, S.; Blasko, B.; Brazile, T.L.; Chase, N.; Desai, D.P.; Evert, J.; Gardner, H.L.; Herrmann, L.; House, A.V.; et al. Large language models provide unsafe answers to patient-posed medical questions. *npj Digit. Med.* **2026**, *9*, 241. [[CrossRef](#)] [[PubMed](#)]
3. Semigran, H.L.; Linder, J.A.; Gidengil, C.; Mehrotra, A. Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ* **2015**, *351*, h3480. [[CrossRef](#)] [[PubMed](#)]
4. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.

5. Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 68539–68551.
6. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 8634–8652.
7. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 186345. [\[CrossRef\]](#)
8. Khasentino, J.; Belyaeva, A.; Liu, X.; Yang, Z.; Furlotte, N.A.; Lee, C.; Schenck, E.; Patel, Y.; Cui, J.; Schneider, L.D.; et al. A personal health large language model for sleep and fitness coaching. *Nat. Med.* **2025**, *31*, 3394–3403. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Kim, Y.; Xu, X.; McDuff, D.; Breazeal, C.; Park, H.W. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. In *Proceedings of the Fifth Conference on Health, Inference, and Learning, New York, NY, USA, 27–28 June 2024*; Proceedings of Machine Learning Research (PMLR): New York, NY, USA, 2024; Volume 248, pp. 522–539.
10. Fang, C.M.; Danry, V.; Whitmore, N.; Bao, A.; Hutchison, A.; Pierce, C.; Maes, P. PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models. In *Proceedings of the 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Houston, TX, USA, 10–13 November 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. [\[CrossRef\]](#)
11. Ha, S.; Lee, T.; Seo, H.; Yoon, S.; Lee, H. A Selective RAG-Enhanced Hybrid ML-LLM Framework for Efficient and Explainable Fatigue Prediction Using Wearable Sensor Data. *Bioengineering* **2026**, *13*, 58. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; Gerstein, M. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 599–621. [\[CrossRef\]](#)
13. Kim, Y.; Park, C.; Jeong, H.; Chan, Y.S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; Park, H.W. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 79410–79452. [\[CrossRef\]](#)
14. Schmidgall, S.; Ziaei, R.; Harris, C.; Kim, J.W.; Reis, E.P.; Jopling, J.; Moor, M. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digit. Med.* **2026**, *9*, 499. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Jiang, Y.; Black, K.C.; Geng, G.; Park, D.; Zou, J.; Ng, A.Y.; Chen, J.H. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI* **2025**, *2*, AIdbp2500144. [\[CrossRef\]](#)
16. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474.
17. Abo El-Enen, M.; Saad, S.; Nazmy, T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Comput. Appl.* **2025**, *37*, 28191–28267. [\[CrossRef\]](#)
18. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*.
19. Prabha, S.; Gomez-Cabello, C.A.; Haider, S.A.; Genovese, A.; Trabilsy, M.; Wood, N.G.; Bagaria, S.; Tao, C.; Forte, A.J. Enhancing Clinical Decision Support with Adaptive Iterative Self-Query Retrieval for Retrieval-Augmented Large Language Models. *Bioengineering* **2025**, *12*, 895. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 6233–6251. [\[CrossRef\]](#)
21. Nguyen, T.; Chin, P.; Tai, Y.W. MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning. *arXiv* **2025**, arXiv:2505.20096. [\[CrossRef\]](#)
22. Zheng, H.; Shi, Z.; Yi, P. MedCoAct: Confidence-Aware Multi-Agent Collaboration for Complete Clinical Decision. In *Proceedings of the 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Wuhan, China, 15–18 December 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 4525–4528. [\[CrossRef\]](#)
23. Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; Wang, Y. MemoryBank: Enhancing Large Language Models with Long-Term Memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2024; Volume 38, pp. 19724–19731. [\[CrossRef\]](#)
24. Hu, M.; Chen, T.; Chen, Q.; Mu, Y.; Shao, W.; Luo, P. HiAgent: Hierarchical Working Memory Management for Solving Long-Horizon Agent Tasks with Large Language Model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 32779–32798. [\[CrossRef\]](#)

25. Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; Jia, J. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
26. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [[CrossRef](#)]
27. Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E.P.; Rajpurkar, P. Med-Flamingo: A Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium (ML4H)*, New Orleans, LA, USA, 10 December 2023; Proceedings of Machine Learning Research (PMLR): New York, NY, USA, 2023; Volume 225, pp. 353–367.
28. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 28541–28564.
29. Li, B.; Yan, T.; Pan, Y.; Luo, J.; Ji, R.; Ding, J.; Xu, Z.; Liu, S.; Dong, H.; Lin, Z.; et al. MMedAgent: Learning to Use Medical Tools with Multi-modal Agent. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, 12–16 November 2024; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 8745–8760. [[CrossRef](#)]
30. Wang, W.; Ma, Z.; Wang, Z.; Wu, C.; Ji, J.; Chen, W.; Li, X.; Yuan, Y. A Survey of LLM-based Agents in Medicine: How Far Are We from Baymax? In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 27 July–1 August 2025; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 10345–10359. [[CrossRef](#)]
31. Abbasian, M.; Azimi, I.; Rahmani, A.M.; Jain, R. Conversational health agents: A personalized large language model-powered agent framework. *JAMIA Open* **2025**, *8*, ooaf067. [[CrossRef](#)] [[PubMed](#)]
32. Ma, X.; Gong, Y.; He, P.; Zhao, H.; Duan, N. Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 6–10 December 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 5303–5315. [[CrossRef](#)]
33. Wang, L.; Yang, N.; Wei, F. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 6–10 December 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 9414–9423. [[CrossRef](#)]
34. Jin, Q.; Kim, W.; Chen, Q.; Comeau, D.C.; Yeganova, L.; Wilbur, W.J.; Lu, Z. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* **2023**, *39*, btad651. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.