PULSEBENCH: PERSONALIZED UNDERSTANDING OF LONGITUDINAL SENSOR EVIDENCE FOR WEARABLE HEALTH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering open-set health questions from longitudinal wearable records requires identifying relevant signals, time windows, and personal reference points, yet this end-to-end evidence discovery remains insufficiently evaluated. We introduce PULSEBench, a benchmark of 3,000 questions from 120 individuals across three public wearable datasets. Questions are grounded in observed trajectories, while a five-dimensional rubric evaluates free-text answers against question-specific evidence that is withheld from answering systems. Our evaluation reveals a persistent gap in evidence localization and supported attribution: adaptive inspection improves every evaluated backbone over direct inference, but the best-performing system reaches only 10.80 out of 20. Controlled experiments further show that longer inputs and uniformly finer temporal resolution do not reliably improve answer quality, highlighting the importance of temporal organization. Motivated by these findings, we introduce Personalized Hierarchical Memory (PHM), which organizes personal records across temporal scales and retrieves evidence conditioned on the question. PHM improves over direct inference with a single answer-generation pass, although iterative agents remain stronger overall in the shared-model comparison. PULSEBench provides a testbed for studying how models discover and use personal evidence, beyond their ability to reason over preselected observations.

1 INTRODUCTION

Wearable devices can accumulate months of sleep, cardiovascular, activity, and recovery data, creating a detailed record of an individual's health and behavior (Fang et al., 2024; Englhardt et al., 2024; Khasentino et al., 2025). But access to a long personal history does not imply that a model knows how to use it. The evidence relevant to a health question is often sparse and distributed across signals and time. Its relevance depends on the question and the individual's observed history, and its location is not supplied to the answering system. For example, answering "Why am I sleeping less lately?" requires comparing recent sleep with personal baselines, aligning changes across signals, and examining preceding events without mistaking temporal association for causation. Longitudinal health reasoning therefore depends not only on reasoning over observed data, but also on *finding the right evidence in the first place*.

Recent benchmarks have broadened evaluation of health and time-series reasoning, including statistical calculation, event localization, cross-signal inference, and long-horizon synthesis (Kong et al., 2025; Li et al., 2026). Tool-augmented agents and sensor-interaction systems further allow models to inspect personal records dynamically (Merrill et al., 2026; Zhang et al., 2026b; Yu et al., 2025). Yet an important question remains underexplored: *when the question does not specify where to look, can a model identify the signals, time periods, and personal reference points that should ground its answer?* Evaluating this ability requires assessing evidence discovery together with the resulting answer, without supplying query-specific evidence annotations.

We introduce PULSEBench, a benchmark for open-set reasoning over longitudinal wearable-health trajectories (Fig. 1). PULSEBench contains 3,000 questions from 120 individuals across three public wearable datasets, spanning eight health domains and seven task modes. Each answering system

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

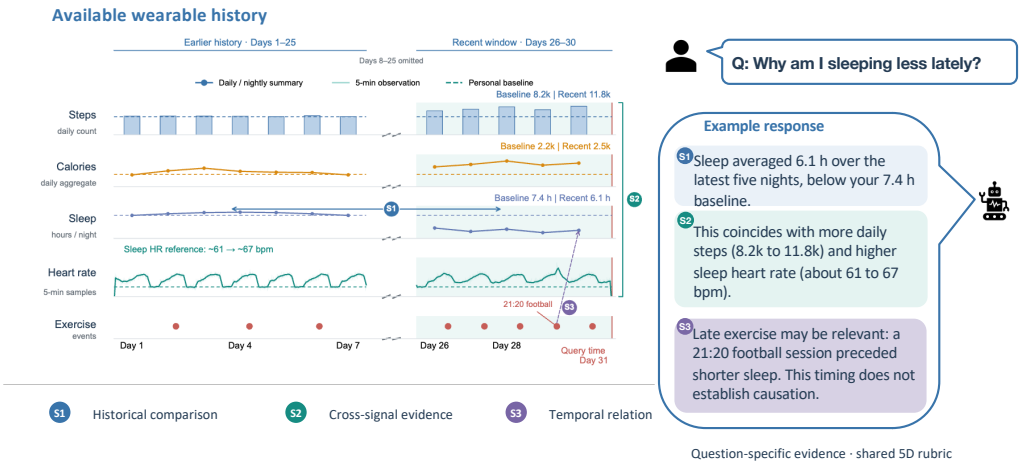

Figure 1: Evidence discovery in PULSEBench, illustrated with synthetic wearable data and an example response. Answering an open-set sleep question requires historical comparison (S1), cross-signal integration (S2), and examination of preceding events (S3), without treating temporal association as causation. Evidence highlights are not supplied to answering systems; responses are scored against question-specific evidence using a shared five-dimensional rubric.

receives the user query and has access to history available by query time, without annotations identifying the relevant evidence. The evaluator additionally receives a question-specific evidence specification and scores each free-text answer using a shared five-dimensional rubric covering accuracy and localization, multi-signal reasoning, attribution, uncertainty, and personalized context.

Our experiments identify evidence access and temporal organization as important constraints on longitudinal reasoning. Adaptive inspection improves every evaluated backbone over direct inference, yet evidence localization and supported attribution remain difficult. Simply supplying longer histories or uniformly finer observations does not reliably improve answer quality in the controlled study. These findings motivate Personalized Hierarchical Memory (PHM), which organizes personal records across temporal scales independently of the query and retrieves relevant evidence when a question arrives. PHM improves over direct inference with a single answer-generation pass, demonstrating the value of structured evidence access while leaving a performance gap to iterative analysis.

The main contributions of our work are

- We formulate open-set longitudinal wearable-health reasoning as an end-to-end setting in which signal selection and temporal localization are implicit parts of answering, and introduce PULSEBench with evidence-grounded evaluation of free-text responses.
- We characterize evidence-access and temporal-representation challenges: adaptive inspection improves all evaluated backbones, while longer or finer-grained inputs do not reliably improve answer quality in controlled experiments.
- We introduce PHM for query-conditioned retrieval from structured personal memory, improving single-pass answering and identifying the remaining gap to iterative tool-using agents.

2 Related Work

LLMs and Agents for Wearable Health. Language models support health prediction (Kim et al., 2024), personalized insight generation and coaching (Fang et al., 2024; Khasentino et al., 2025), and clinician-facing interpretation of behavioral data (Englhardt et al., 2024). SensorLM learns sensor–language representations for recognition, retrieval, and captioning (Zhang et al., 2025). Agentic

Table 1: Comparison with related longitudinal time-series and wearable-health benchmarks.

	Longitudinal Data			Question Formulation			Evidence Reasoning		
Benchmark	**Real-world**	**Users**	**History Span**	**User-facing**	**Free-form**	**Cross-signal**	**Signal Sel.**	**Temporal Loc.**	**Evidence-grounded Eval.**
Time-MQA (Kong et al., 2025)	Mixed	–	–	×	△	×	×	×	×
SensorQA (Reichman et al., 2025)	✓	60	2 d–3 mo	✓	✓	△	×	△	×
OpenLifelogQA (Tran et al., 2025)	✓	1	18 mo	✓	✓	✓	×	×	×
HEARTS (Li et al., 2026)	✓	–	Varies	×	×	△	×	×	×
LifeAgentBench (Tian et al., 2026)	✓	100	1 mo	✓	△	✓	△	×	×
WEQA (Zhang et al., 2026b)	✓	212†	10 wk†	✓	△	✓	✓	△	△
Ours	✓	120	29–244 d (med. 71 d)	✓	✓	✓	✓	✓	✓

Notes. ✓: explicitly supported; △: partially supported or supported only by a subset of tasks; ×: not part of the benchmark formulation; –: not applicable or not reported. History span denotes the duration of longitudinal records available per user, rather than the temporal window required by an individual question. †For WEQA, user count and history span refer to its longitudinal TILES-2018 component.

systems extend these capabilities: PHIA combines language-model reasoning with code execution and retrieval (Merrill et al., 2026); WEQA routes questions to analytical and predictive tools (Zhang et al., 2026b); VitalAgent integrates physiological memory with tools for reactive and proactive monitoring (Zhu et al., 2026). These systems motivate studying how evidence access affects reasoning over longitudinal personal records.

Longitudinal Health and Sensor Benchmarks. ChatTS studies understanding and reasoning over multivariate time series (Xie et al., 2025), while Time-MQA evaluates multiple time-series question-answering tasks (Kong et al., 2025). SensorQA and SensorChat address daily-life sensor questions, including retrieval from long-term multimodal records (Reichman et al., 2025; Yu et al., 2025); OpenLifelogQA targets open-ended multimodal lifelog questions (Tran et al., 2025). HEARTS evaluates health time-series capabilities, including localization and trajectory reasoning (Li et al., 2026), while LifeAgentBench and WEQA cover longitudinal lifestyle and wearable-health analysis (Tian et al., 2026; Zhang et al., 2026b). PULSEBench treats signal selection and temporal localization as implicit components of answering open-set questions, evaluating their consequences jointly with attribution, uncertainty, and personalized context.

Long Context, Memory, and Evidence Retrieval. Lost in the Middle identifies sensitivity to evidence position (Liu et al., 2024), and RULER tests effective context use beyond simple retrieval (Hsieh et al., 2024). LongMemEval evaluates persistent memory over conversational histories (Wu et al., 2025). Approaches to evidence access include iterative reasoning and action in ReAct (Yao et al., 2023), hierarchical summaries in RAPTOR (Sarthi et al., 2024), graph-based retrieval in HippoRAG and AssoMem (Jiménez Gutiérrez et al., 2024; Zhang et al., 2026a), and staged memory consolidation in LightMem (Fang et al., 2026). These methods primarily address documents or conversations. We examine history length and temporal representation in wearable trajectories, and evaluate PHM as a targeted intervention that retrieves multiscale personal evidence for single-pass answering.

Table 1 compares closely related sensor and health benchmarks. PULSEBench combines free-form questions, implicit evidence discovery, and evaluation of the resulting evidence-grounded interpretation. The distinction lies in this joint formulation, not in localization or retrieval alone.

3 BENCHMARK

3.1 TASK FORMULATION

We consider open-set question answering over longitudinal wearable health records. For a query q issued at time t_q, the available user history is

$$H_{\leq t_q} = \{(s, t, x_s(t)) \mid s \in S,\ t \in T_s,\ t \leq t_q\},$$

where S is the set of available signals, T_s is the set of observation timestamps for signal s, and $x_s(t)$ is its observed value at time t. The system must produce a free-text answer grounded in this history, without access to future observations.

Relevant evidence must be located within the trajectory: queries may name signals or suggest temporal anchors, but no annotated, query-specific evidence subset is supplied. Successful answering therefore requires *signal selection* and *temporal localization*, including personal reference periods when needed. These are implicit requirements of end-to-end answering, not separate required outputs: evidence may be identified explicitly through retrieval or tool use, or implicitly during reasoning.

3.2 Benchmark Construction

Longitudinal wearable trajectories. We use three public datasets: LifeSnaps (Yfantidou et al., 2022), PMData (Thambawita et al., 2020), and FitBit Fitness Tracker Data (Möbius, 2016). Their measurements are converted into a unified per-user representation, preserving source granularity and missingness while normalizing timestamps and enforcing the query-time cutoff (Appendix A.1).

Query construction. Questions originate from patterns in observed trajectories. More than 25 deterministic operators scan up to 30 days of preceding context for events, anomalies, trends, personal-baseline deviations, and cross-signal associations. Each triggered pattern produces a data-grounded seed with a temporal anchor that determines the query-time cutoff. An LLM turns selected seeds into user-facing questions while preserving their temporal perspective and evidence requirements. Automated critique, selection, and rephrasing filter invalid candidates and reduce repetitive or detector-specific wording.

Evidence specification and validation. A deterministic planner extracts evidence from the visible trajectory; an LLM structures it into a query-specific specification of relevant signals, temporal windows, key observations, supported attribution, and expected uncertainty. This specification is evaluation-only and never exposed to the answering model. Schema and temporal-visibility checks, together with multi-judge calibration, reject cases with unresolved evaluation failures or evidence leakage before benchmark export. The construction and validation pipeline is detailed in Appendix A.2.

3.3 Task Taxonomy and Statistics

The benchmark contains 3,000 questions from 120 individuals, covering eight *health domains* and seven *task modes*. These axes distinguish physiological content from the reasoning required to interpret it: the same signal may support lookup, personal comparison, or contextual attribution. Per-user records span 29–244 days (median: 71); this source-record duration is distinct from the evidence horizon required by an individual question (Appendix A.4). Lookup and Personal Baseline Comparison account for 36.5% and 31.8% of questions, respectively. Figure 2 visualizes the source-record spans and non-uniform domain–task coverage. Detailed definitions and counts are provided in Appendix A.

3.4 Evaluation Protocol

HealthBench uses physician-authored rubrics for open-ended health conversations (Arora et al., 2025). Here, we ground rubric-based assessment in the observed wearable evidence for each question. We use five fixed dimensions: Data Accuracy & Localization (D1, 0–4), Multi-Signal Reasoning (D2, 0–4), Attribution Quality (D3, 0–4), Uncertainty Calibration (D4, 0–3), and Context-Aware Reasoning (D5, 0–5). Their sum yields a 20-point response score. Common scoring criteria are grounded in the query-specific evidence specification; dimension definitions are given in Appendix B.1.

We use criteria-guided LLM judging (Liu et al., 2023), with Kimi-K2.5 as the primary evaluator unless otherwise stated. Each judge receives the question, fixed rubric, evidence specification, judge-only evidence context, and candidate answer, all subject to the query-time cutoff. Per-model scores average available valid judgments; coverage and missing-score handling are detailed in Ap-

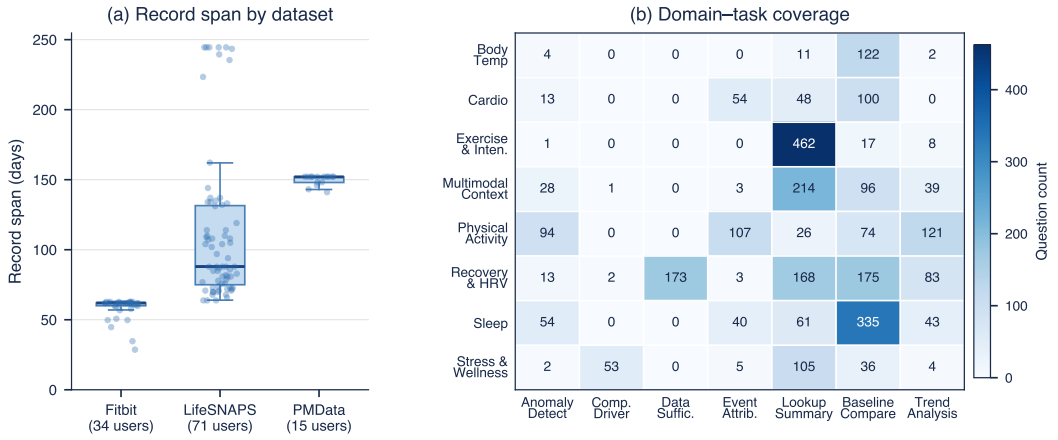

Figure 2: **Benchmark dataset statistics. (a)** Source-record spans for benchmark users, grouped by dataset. Dots denote users; boxes show medians and interquartile ranges (IQR), with whiskers extending to the most extreme observations within $1.5 \times \mathrm{IQR}$ of the box. Spans count inclusive calendar days and do not imply continuous daily coverage. **(b)** Question counts across health domains and task modes (3,000 questions in total).

pendix B.2. Alternative judges assess evaluator robustness; their scores are not averaged into the primary results.

Cross-judge consistency. To assess sensitivity to evaluator choice (Zheng et al., 2023), we compare matched response scores for the seven open-weight answer models under Direct LLM and ReAct Agent. Alternative evaluators show strong response-level agreement with Kimi-K2.5 ($\rho = 0.793$–0.795), despite differences in absolute calibration (Table 2). All three evaluators preserve ReAct's advantage over Direct, indicating that this central comparison is not specific to the primary evaluator within the evaluated open-weight setting. Across D1–D5, within-one-point agreement with the primary judge ranges from 92.2% to 99.9%; this tolerance uses raw dimension scores, whose maxima differ. Full agreement statistics are reported in Appendix B.3. These checks assess evaluator consistency, not clinical validity or agreement with independent experts.

Table 2: Evaluator robustness on seven open-weight answer models. Spearman ρ is computed from matched response scores against the primary evaluator; Direct and ReAct are macro-averages of per-model means.

Evaluator	ρ	**Direct**	**ReAct**	Δ
Kimi-K2.5	–	6.21	8.14	+1.93
DeepSeek-V4-Flash	0.793	5.99	8.56	+2.57
Qwen3.5-397B-A17B	0.795	5.05	7.72	+2.67

4 Personalized Hierarchical Memory

4.1 Overview

Personalized Hierarchical Memory (PHM) separates longitudinal memory construction from query-time evidence access. It organizes wearable observations into multi-scale personal memory, retrieves a query-conditioned context, and generates one free-text answer:

$$H_{\leq t_q} \rightarrow M_{\leq t_q} \xrightarrow[\text{query } q]{\text{structured retrieval}} E_q \rightarrow y,$$

where $M_{\leq t_q}$ is built from observations visible at the query time and E_q is the retrieved evidence. Unlike Direct LLM, context is selected for the query; unlike ReAct, retrieval is completed before an-

swer generation rather than through iterative tool use. We evaluate this design as a targeted evidence-access intervention.

4.2 Multi-Scale Personalized Health Memory

PHM retains local observations, health events, and window-level summaries, consolidates them into daily aggregates, and maintains longer-term personalized baselines and trends. Local memories describe a recent episode, while historical summaries provide the user's own reference for interpreting it. The representation supports incremental updates under the query-time cutoff, including partial-day summaries when a question occurs before the day ends.

The memory also stores temporal, hierarchical, baseline-comparison, and cross-signal relations. These describe associations rather than causal conclusions; they are not traversed in the reported configuration. Memory types, temporal resolutions, and update rules are described in Appendix C.1.

4.3 Query-Conditioned Retrieval and Answering

A query parser infers signal families, intent, and temporal cues from the question and visible history, without reading annotations saved during benchmark construction. Retrieval combines semantic candidates with rule-scored candidates from the visible memory. The rules match node attributes—including signal compatibility, temporal proximity, and evidence roles—rather than relation edges. Hybrid retrieval combines normalized semantic and rule scores; temporal-pattern queries use a rule-only policy.

Role-aware selection and signal-coverage checks precede query–evidence reranking. Context construction then deduplicates, reorders, and budgets the selected items. The reported *1-hop* configuration means direct node selection without graph-neighborhood expansion, with a retrieval-selection limit of 120 and a final context limit of 90 memory items (Appendix C.2).

The answer model receives the question, query time, and retrieved context, then generates one response without further file inspection. Single-pass refers to answer generation, not the separate query-parsing call. Neither the evaluation rubric nor the query-specific evidence specification is supplied to this pipeline. Prompt templates and context construction are documented in Appendix C.3.

5 Experiments

5.1 Experimental Setup

Answering settings. Direct LLM answers from a fixed representation of visible history; ReAct Agent uses tools to inspect records and perform computations iteratively; PHM retrieves structured memory for single-pass answer generation. All three are restricted to records available by the question-time cutoff and cannot access evaluation rubrics or query-specific evidence annotations. They share eligible source records, not identical token-level inputs. Context construction, tools, and inference settings are detailed in Appendix D.1; evaluation follows Sec. 3.4.

Evaluation scope. The benchmark-wide comparison (Sec. 5.2) evaluates Direct LLM and ReAct across ten answer models—seven open-weight and three proprietary—on the full 3,000-question benchmark. The controlled context study (Sec. 5.3) crosses three history windows with three temporal representations for two answer models on a fixed 500-question subset. The structured-memory intervention (Sec. 5.4) compares PHM with Direct and ReAct on the two shared backbones Qwen3.5-397B-A17B and DeepSeek-V4-Pro. These studies address complementary questions and are analyzed separately.

5.2 Benchmarking LLMs and Agents

Table 3 reports benchmark-wide performance for all ten answer models under both Direct LLM and ReAct Agent. Two results stand out. First, ReAct improves every evaluated model and every rubric dimension on average in both model groups. Among open-weight models, it raises the average 20-point score from 6.21 to 8.14 ($+1.93$); among proprietary models, it improves the average from 8.96

to 10.14 (+1.18). Second, substantial headroom remains: the strongest result is only 10.80 out of 20, with evidence localization and supported attribution among the weakest normalized dimensions. The Direct–Agent ordering is unchanged when each model is restricted to responses scored in both settings (Appendix B.2).

Table 3: Benchmark-wide performance on the full 3,000-question set. ReAct Agent improves all ten answer models; Δ denotes Agent minus Direct, and dimension entries report Direct / Agent.

Answer Model	Overall (/20)			Dimension Scores (Direct / Agent)				
	Direct	Agent	Δ	D1 (/4)	D2 (/4)	D3 (/4)	D4 (/3)	D5 (/5)
Open-weight models								
Qwen3.5-35B-A3B	5.63	7.91	+2.28	0.63 / 1.16	1.64 / 2.01	0.71 / 1.26	1.02 / 1.37	1.63 / 2.11
Qwen3.5-122B-A10B	6.00	8.20	+2.20	0.69 / 1.22	1.72 / 2.05	0.80 / 1.33	1.06 / 1.39	1.73 / 2.20
Qwen3.5-397B-A17B	5.67	8.43	+2.76	0.63 / 1.24	1.64 / 2.10	0.73 / 1.39	1.02 / 1.42	1.66 / 2.27
DeepSeek-V4-Flash	6.21	7.89	+1.68	0.77 / 1.10	1.75 / 2.06	0.86 / 1.24	1.06 / 1.32	1.78 / 2.17
DeepSeek-V4-Pro	6.16	8.26	+2.10	0.75 / 1.15	1.74 / 2.13	0.84 / 1.30	1.05 / 1.44	1.77 / 2.24
Kimi-K2.5	**7.02**	7.83	+0.81	0.91 / 1.09	1.88 / 2.04	1.02 / 1.24	1.25 / 1.30	1.96 / 2.16
GLM-5	6.81	**8.47**	+1.66	0.86 / 1.23	1.81 / 2.13	0.97 / 1.37	1.28 / 1.46	1.89 / 2.28
Average	6.21	8.14	+1.93	0.75 / 1.17	1.74 / 2.07	0.85 / 1.31	1.11 / 1.39	1.77 / 2.20
Proprietary models								
GPT-5.4	**9.31**	10.12	+0.81	1.31 / 1.45	2.14 / 2.30	1.62 / 1.87	1.60 / 1.69	2.65 / 2.82
Claude Opus 4.8	8.68	**10.80**	+2.12	1.14 / 1.55	2.03 / 2.36	1.43 / 1.99	1.57 / 1.97	2.51 / 2.93
Gemini 3.1 Pro	8.89	9.51	+0.62	1.52 / 1.58	1.94 / 2.14	1.62 / 1.77	1.30 / 1.30	2.52 / 2.71
Average	8.96	10.14	+1.18	1.32 / 1.53	2.04 / 2.27	1.56 / 1.88	1.49 / 1.66	2.56 / 2.82

Where Access Helps and Gaps Remain. The largest absolute rubric gain occurs in D3 (Attribution Quality): $+0.46$ for open-weight models and $+0.32$ for proprietary models. Gains in D1 (Data Accuracy and Localization) and D5 (Context-Aware Reasoning) further suggest that adaptive access helps models locate relevant observations and connect them to personalized context. Yet after normalization by each dimension's maximum, D1 and D3 remain the two lowest-scoring dimensions in both ReAct model groups. For open-weight and proprietary models, respectively, D1 averages approximately 29% and 38% of its maximum, while D3 averages 33% and 47%. These are rubric-relative scores, not answer accuracies or direct retrieval metrics. Improved access therefore does not yet ensure accurate evidence localization or well-supported attribution.

Model Capability and Evidence Access. Proprietary models substantially outperform the open-weight group under Direct inference, while ReAct yields further gains for every model. Within the Qwen3.5 family, Direct scores are non-monotonic across the 35B-A3B, 122B-A10B, and 397B-A17B variants (5.63, 6.00, and 5.67), whereas the corresponding Agent scores increase from 7.91 to 8.20 and 8.43. This within-family pattern is descriptive, not evidence of a general scaling law.

Finding 1: Adaptive inspection helps across backbones, but accurate localization and supported attribution remain unresolved.

5.3 Controlled Effects of Context Length and Temporal Representation

We ask how the amount and temporal representation of supplied history affect answer quality when questions and the Direct LLM answering mode are held fixed. On 500 questions from 115 individuals, we vary the input window over the 3, 7, or 30 days preceding query time for Qwen3.5-397B and DeepSeek-V4-Flash. These are supplied-history windows, not the evidence horizons required by the questions; a shorter window may omit relevant earlier observations. We cross each window with three representations. *Fine* retains higher-resolution observations, *Coarse* compresses the history into lower-resolution summaries, and *Multiscale* combines recent local detail with coarser historical summaries. Every condition uses the same paired cases, and Figure 3 reports results under the primary judge. Full matrices, secondary-judge results, and supporting statistics are provided in Appendix D.2.

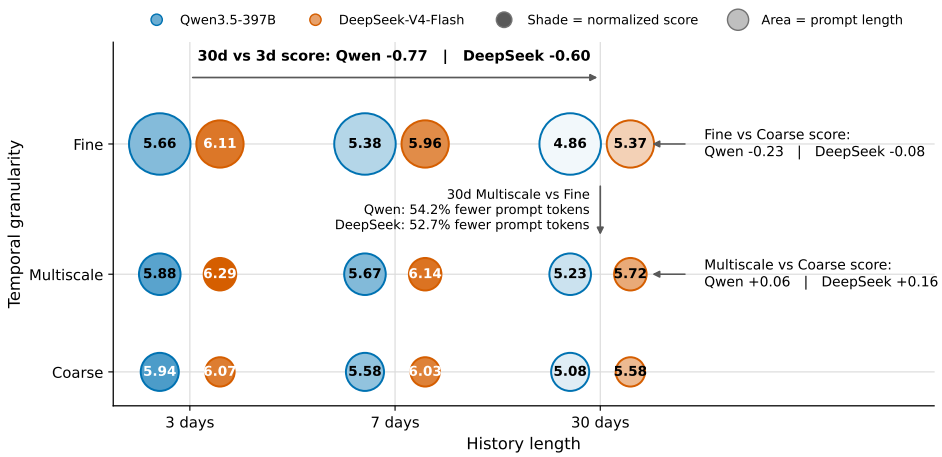

Figure 3: **Controlled effects of input history length and temporal representation on the 500-question subset.** Blue and orange bubbles denote Qwen3.5-397B and DeepSeek-V4-Flash. Labels report mean scores on the 20-point rubric; shading and area provide visual encodings of normalized score and prompt-token count, respectively. History contrasts (30d minus 3d) average over representations; representation contrasts average over history lengths. Token reductions compare *Multiscale* with *Fine* at 30 days.

Longer Inputs Lower Scores in This Study. Across all three temporal representations, the 30-day setting underperforms the 3-day setting for both models. Averaged across representations, extending the history from 3 to 30 days decreases the score by 0.77 points for Qwen3.5-397B and 0.60 points for DeepSeek-V4-Flash. Thus, supplying more history lowers mean answer scores in these conditions. Longer prompts may make relevant evidence harder to identify, but this experiment does not isolate that mechanism or establish that shorter histories suffice for every question.

Multiscale Improves Scores with Fewer Tokens at 30 Days. With 30 days of history, *Multiscale* achieves the highest score for both models. Compared with *Fine*, it improves performance by 0.37 points for Qwen3.5-397B and 0.35 points for DeepSeek-V4-Flash, while using 54.2% and 52.7% fewer prompt tokens, respectively. It also outperforms *Coarse* by 0.15 and 0.14 points. By retaining recent fine-grained observations while compressing older history, *Multiscale* provides a better trade-off between temporal detail and context length, particularly as the visible history grows.

> ***Finding 2:*** For both tested models, 30-day inputs score lower than 3-day inputs across all three representations. At 30 days, *Multiscale* scores higher than uniformly *Fine* inputs while using roughly half as many prompt tokens, supporting temporal organization over simply increasing input volume.

5.4 Structured Memory as a Targeted Evidence-Access Intervention

We next test whether structured retrieval can improve Direct LLM answering without iterative tool use during answer generation. On the two shared backbones Qwen3.5-397B-A17B and DeepSeek-V4-Pro, PHM raises the mean score from 5.91 for Direct LLM to 7.72, compared with 8.34 for ReAct (Table 4). Structured evidence access therefore improves single-pass answering, while leaving a gap to iterative analysis. This comparison concerns the complete answering pipelines on these two backbones, rather than an isolated retrieval component or a benchmark-wide ranking.

Task-Dependent Benefits. PHM outperforms both Direct LLM and ReAct on Personal Baseline Comparison, Anomaly Detection, and Event Context Attribution, with respective scores of 8.25, 8.81, and 9.38. These tasks rely on recovering personal baselines or event-related observations, consistent with PHM's role in retrieving distributed longitudinal evidence.

The ReAct Agent remains stronger on Lookup / Summary, Trend Analysis, Component / Driver Analysis, and Data Sufficiency. In particular, trend and driver questions often require explicit aggregation over a target interval or decomposition of underlying measurements, operations supported by ReAct's inspection and computation tools. This pattern is consistent with a role for iterative computation, but the task-level comparison does not isolate the source of the performance differences.

Table 4: Overall and task-mode performance on the full 3,000-question benchmark, averaged over Qwen3.5-397B-A17B and DeepSeek-V4-Pro, the two answer models evaluated in all three answering settings. Best results in each row are bolded.

Task Mode	Direct LLM	ReAct Agent	PHM
Anomaly Detection	6.40	7.90	**8.81**
Component / Driver Analysis	4.59	**7.22**	5.07
Data Sufficiency / Quality Check	5.42	**8.05**	7.62
Event Context Attribution	5.76	7.98	**9.38**
Lookup / Summary	6.14	**9.54**	7.29
Personal Baseline Comparison	6.06	7.56	**8.25**
Trend Analysis	4.93	**7.36**	6.22
Overall	5.91	**8.34**	7.72

Finding 3: Structured retrieval improves single-pass answering over Direct LLM on the shared backbones. Its task-specific strengths do not eliminate the overall gap to ReAct, which retains an advantage on tasks including trend and driver analysis.

6 Conclusion

PULSEBench evaluates whether models can find and use the signals, time windows, and personal references needed for longitudinal wearable-health reasoning. Across ten backbones, adaptive evidence access improves performance, while controlled experiments show that more history or finer resolution alone does not guarantee better answers. PHM demonstrates that structured retrieval can improve single-pass answering over Direct LLM; the remaining gap to ReAct motivates further work on combining evidence organization with iterative analysis. Together, these results highlight evidence access and temporal organization as central considerations for personalized reasoning over long wearable histories.

Limitations. PULSEBench draws on 120 individuals from three public datasets and therefore provides limited coverage of populations, devices, and recording conditions. Its domain and task distributions are uneven, so aggregate scores reflect the benchmark's composition rather than uniform coverage of health needs. Questions and evidence specifications are constructed with LLM assistance from data-grounded seeds, which may introduce phrasing and selection biases relative to naturally occurring user questions. Automated validation and cross-judge agreement help assess consistency but cannot rule out shared model biases or replace independent expert assessment. Finally, the benchmark measures evidence-grounded reasoning from observational wearable records; it does not establish causal explanations, clinical validity, or the safety and effectiveness of health recommendations in real-world use.

Statement on the Use of Generative AI

Generative AI tools assisted with literature exploration, manuscript drafting and editing. LLM-based benchmark construction and automated answer evaluation are described in the corresponding methodology sections and appendices. The authors reviewed and revised the AI-assisted content and take full responsibility for the final manuscript and associated artifacts.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. URL https://arxiv.org/abs/2505.08775.

Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–25, 2024. doi: 10.1145/3659604.

Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. PhysioLLM: Supporting personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–8. IEEE, 2024. doi: 10.1109/BHI62660.2024.10913781.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a05b72653ec5b473732129829ae04195-Abstract-Conference.html.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.06654.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A. Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, Megan Walker, Jeffrey Yu, Tim Althoff, Conor Heneghan, John Hernandez, Mark Malhotra, Leor Stern, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Jiening Zhan, Shruthi Prabhakara, Daniel McDuff, and Cory Y. McLean. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31(10):3394–3403, 2025. doi: 10.1038/s41591-025-03888-0.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-LLM: Large language models for health prediction via wearable sensor data. In *Proceedings of the Fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, pp. 522–539. PMLR, 2024. URL https://proceedings.mlr.press/v248/kim24b.html.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-MQA: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29736–29753, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1437. URL https://aclanthology.org/2025.acl-long.1437/.

Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. HEARTS: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17(1):1143, 2026. doi: 10.1038/s41467-025-67922-y.

Möbius. Fitbit fitness tracker data. https://www.kaggle.com/datasets/arashnic/fitbit, 2016.

Benjamin Reichman, Xiaofan Yu, Lanxiang Hu, Jack Truxal, Atishay Jain, Rushil Chandrupatla, Tajana Simunic Rosing, and Larry Heck. SensorQA: A question answering benchmark for daily-life monitoring. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pp. 282–289, 2025. doi: 10.1145/3715014.3722074.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8a2acd174940dbca361a6398a4f9df91-Abstract-Conference.html.

Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, et al. Pmdata: a sports logging dataset. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pp. 231–236, 2020.

Ye Tian, Zihao Wang, Onat Gungor, Xiaoran Fan, and Tajana Rosing. LifeAgentBench: A multi-dimensional benchmark and agent for personal health assistants in digital health. *arXiv preprint arXiv:2601.13880*, 2026.

Quang-Linh Tran, Binh Nguyen, Gareth J. F. Jones, and Cathal Gurrin. OpenLifelogQA: An open-ended multi-modal lifelog question-answering dataset. *arXiv preprint arXiv:2508.03583*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 2025. doi: 10.14778/3742728.3742735. URL https://arxiv.org/abs/2412.03104.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data*, 9(1):663, 2022.

Xiaofan Yu, Lanxiang Hu, Benjamin Reichman, Dylan Chu, Rushil Chandrupatla, Xiyuan Zhang, Larry Heck, and Tajana Rosing. SensorChat: Answering qualitative and quantitative questions during long-term multimodal sensor interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025. doi: 10.1145/3749496.

Kai Zhang, Xinyuan Zhang, Ejaz Ahmed, Hongda Jiang, Caleb Kumar, Kai Sun, Zhaojiang Lin, Sanat Sharma, Shereen Oraby, Aaron Colak, Ahmed Aly, Anuj Kumar, Xiaozhong Liu, and Xin Luna Dong. AssoMem: Scalable memory qa with multi-signal associative retrieval. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a921f335253add9996d5175ad30896ec-Abstract-Conference.html.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/42cd98f0e7520d4a63c34891ac1c972f-Abstract-Conference.html.

Yuwei Zhang, Tong Xia, Bianca Emmerich, Yu Yvonne Wu, Dimitris Spathis, Xin Liu, Daniel McDuff, and Cecilia Mascolo. WEQA: Wearable health question answering with query-adaptive agentic reasoning. *arXiv preprint arXiv:2606.18147*, 2026b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Di Zhu, Yu Yvonne Wu, Hong Jia, Aaqib Saeed, Vassilis Kostakos, and Ting Dang. VitalAgent: A tool-augmented agent for reactive and proactive physiological monitoring over wearable health data. *arXiv preprint arXiv:2605.29483*, 2026.

A Benchmark Details

A.1 Source Datasets and Preprocessing

We construct the benchmark from three public wearable and lifelog datasets: LifeSnaps (Yfantidou et al., 2022), PMData (Thambawita et al., 2020), and FitBit Fitness Tracker Data (Möbius, 2016). The final benchmark contains 1,607 questions from 71 individuals in LifeSnaps, 994 questions from 15 individuals in PMData, and 399 questions from 34 individuals in FitBit Fitness Tracker Data, for a total of 3,000 questions from 120 individuals.

The source datasets contain heterogeneous physiological and behavioral measurements, including heart rate, resting heart rate, HRV, sleep and sleep stages, physical activity, exercise, stress and wellness, body temperature, SpO_2, respiratory rate, and related daily or event-level summaries. Because signal availability differs across datasets and individuals, benchmark instances preserve the partially observed nature of the original trajectories rather than requiring a common set of modalities.

We convert each dataset into a unified per-user CSV representation while preserving the original temporal granularity of each signal, ranging from high-frequency measurements to daily summaries and event-level records. Timestamps are normalized to a common datetime representation. Missing measurements are retained without interpolation, and continuous signals are not resampled or subjected to additional anomaly cleaning. A provenance index records the correspondence between processed files and their source records.

All benchmark construction and evaluation obey a strict query-time cutoff: only observations available at or before the question time may be exposed to either the answering system or the evaluation

pipeline. Records crossing the cutoff, such as sleep-stage episodes, are clipped when necessary to prevent future-data leakage.

A.2 Query Generation and Evidence Specification

Candidate generation. Questions are generated from patterns detected in each individual's trajectory rather than independently sampled templates. We scan the trajectory using daily anchors with up to 30 days of preceding context and apply more than 25 deterministic operators covering patterns such as anomalies, temporal trends, periodicity, sleep changes, personal-baseline deviations, cross-signal associations, recovery, and contextual events. Each triggered operator produces a data-grounded seed together with an associated temporal anchor. The question time is deterministically derived from the triggering event or anchor and defines the latest observation visible to the answering system.

Natural-language question construction. Data-grounded seeds are converted into natural-language drafts using an LLM. The drafting stage preserves the temporal perspective implied by the question time while replacing detector-specific or benchmark-specific terminology with natural user-facing language. Candidate questions subsequently undergo automated critic and selection procedures, followed by a final rephrasing stage that preserves the underlying evidence requirement and temporal anchor while reducing repetitive or templated phrasing.

Evidence planning. For each selected question, a deterministic evidence planner maps the query to potentially relevant modalities, temporal regions, and evidence operators. These operators extract a neutral evidence brief directly from the visible trajectory using the question time as a strict cutoff. For temporally anchored queries, the planner resolves local event windows and historical comparison windows when required.

Structured evidence specification. An LLM converts the extracted evidence brief into a structured specification containing key observations, relevant signals, a critical temporal window, primary supported attribution, secondary or unsupported explanations, expected uncertainty, and critical or supporting evidence. The specification is used only by the evaluation pipeline and is never shown to the answering model.

Quality control. Structured specifications are automatically validated for schema completeness, temporal visibility, and compatibility with the fixed evaluation rubric. Benchmark construction additionally includes multi-judge calibration. Medical personnel manually review disputed cases and, according to their professional judgment, revise cases with minor issues and discard those with major issues. Automated checks separately reject invalid scoring outputs, future-data leakage, or missing required artifacts before final export. This review concerns benchmark cases and evidence specifications; it is not an independent medical assessment of all model-generated answers.

A.3 Task Taxonomy

We characterize each benchmark question along two complementary axes. The *health domain* describes the physiological or behavioral information involved, while the *task mode* describes how longitudinal evidence must be accessed and interpreted.

The eight health domains are Cardiovascular, Sleep, Physical Activity, Exercise & Intensity, Recovery & HRV, Stress & Wellness, Body Temperature, and Multimodal Health Context. The seven task modes are Lookup, Personal Baseline Comparison, Anomaly Detection, Trend Analysis, Event Context Attribution, Component/Driver Analysis, and Data Sufficiency.

Table 5 reports the joint distribution of the 3,000 benchmark questions. The benchmark is intentionally heterogeneous rather than uniformly balanced, reflecting differences in signal availability and the number of valid questions that can be grounded in each source trajectory.

Table 5: Distribution of the 3,000 benchmark questions across health domains and task modes. "Baseline", "Event", "Driver", and "Suff." denote Personal Baseline Comparison, Event Context Attribution, Component/Driver Analysis, and Data Sufficiency, respectively.

Health Domain	Lookup	Baseline	Anomaly	Trend	Event	Driver	Suff.
Cardiovascular	48	100	13	–	54	–	–
Sleep	61	335	54	43	40	–	–
Physical Activity	26	74	94	121	107	–	–
Exercise & Intensity	462	17	1	8	–	–	–
Recovery & HRV	168	175	13	83	3	2	173
Stress & Wellness	105	36	2	4	5	53	–
Body Temperature	11	122	4	2	–	–	–
Multimodal Context	214	96	28	39	3	1	–

A.4 Benchmark Statistics

Per-user source records span 29–244 days, with a median of 71 days. These durations describe the retained longitudinal records, not the history required to answer every question or the context supplied in every experimental condition. For each question, only records visible at its query-time cutoff are eligible; the construction scan described in Appendix A.2 uses up to 30 days of preceding context. Dataset-specific question and participant counts are reported in Appendix A.1, and the joint domain–task distribution is given in Table 5.

B Evaluation Details

B.1 Evaluation Rubric

Rubric-based assessment supports evaluation of open-ended health answers (Arora et al., 2025). Our benchmark uses a fixed five-dimensional rubric grounded in query-specific wearable evidence. A response's total score is the sum of its five dimension scores, with a maximum of 20 points.

Table 6 summarizes the five dimensions and reproduces their operational scoring anchors. Judges apply these anchors to the query-specific evidence specification, including relevant signals, temporal windows, supported attribution, uncertainty, and supporting evidence.

Partial credit and evidence status. Dimensions are scored independently: an evidence error does not automatically erase demonstrated capability in every dimension. Judges first assess demonstrated strengths and dimension-specific errors, then apply any relevant score caps (Table 7). Critical evidence is required for high credit in the affected dimensions. The evidence specification is not exhaustive: an unlisted auxiliary value is not penalized merely because it is absent from the scoring context. It remains secondary and cannot replace critical evidence for high credit. Contradicted, impossible, future, or disallowed evidence, and auxiliary evidence that drives the opposite conclusion, are penalized under the corresponding rules. Minor numeric imprecision can retain partial credit; exact wording is not required.

B.2 Scoring Protocol

We use criteria-guided LLM evaluation (Liu et al., 2023), with Kimi-K2.5 as the primary judge and additional judges on the robustness-study response sets. Judges receive the natural-language question, the fixed evaluation rubric, the query-specific evidence specification, a judge-only compressed evidence context, and the candidate answer. Judges do not receive a reference answer or information unavailable at the query-time cutoff. Unless explicitly stated otherwise, reported results use the primary judge alone, not an average across judges.

Coverage and missing scores. All ten answer models are evaluated against the same 3,000-question benchmark under Direct LLM and ReAct Agent. Table 8 lists the combinations with incomplete primary-judge coverage; every other combination has 3,000 valid scores. Missing or

Table 6: Five-dimensional rubric and operational scoring anchors. Anchors guide credit within each dimension; score caps are specified separately in Table 7.

Dimension	Scoring anchors
D1 (4 points) Data Accuracy & Localization	*High credit.* Uses the critical visible data accurately, localizes the correct time window, selects relevant signals, and avoids future/cutoff leakage or unsupported critical evidence. *Partial credit.* Main conclusion is data-grounded but has minor numeric imprecision, omits some noncritical localization detail, or cites unlisted auxiliary values that do not contradict the supplied evidence. *Heavy penalty.* Uses cutoff-after data, asserts clearly contradicted or impossible critical numbers/signals, misidentifies the time window, or presents inference as directly observed data.
D2 (4 points) Multi-Signal Reasoning	*High credit.* Combines the relevant available signals into a coherent observation-to-inference chain and uses cross-signal consistency to support the answer. *Partial credit.* Uses more than one signal but connects them shallowly, or relies mostly on one signal while mentioning other relevant evidence. *Heavy penalty.* Ignores clearly relevant signals, reasons from a single cherry-picked signal, or integrates nonexistent/irrelevant signals as if they were evidence.
D3 (4 points) Attribution Quality	*High credit.* Identifies the primary supported attribution, supports it with matching evidence, and clearly distinguishes any secondary explanation as weaker. *Partial credit.* Gives a plausible but secondary explanation, or mentions the primary explanation without clear evidence-weighted prioritization. *Heavy penalty.* Centers an unsupported or overreaching attribution, asserts unsupported causality, states a diagnosis/intent not supported by the evidence, or lists explanations without priority.
D4 (3 points) Uncertainty Calibration	*High credit.* Expresses confidence at the expected level, separates observed facts from inferred causes, and states meaningful limitations without being evasive. *Partial credit.* Includes some uncertainty language but slightly overstates the evidence or misses an important limitation. *Heavy penalty.* States weakly supported inferences as fact, makes unsupported medical/causal claims, or ignores ambiguity called out by the evidence spec.
D5 (5 points) Context-Aware Reasoning	*High credit.* Directly answers this user's question using the case-specific timing, trends, signals, and baseline/context windows. *Partial credit.* Uses some case facts but remains partly generic, underspecified, or loosely tied to the user's actual question. *Heavy penalty.* Mostly gives generic health advice, ignores the case context, dumps metrics without interpretation, or gives recommendations unsupported by the evidence.

Table 7: Score caps in the evaluation prompt. Caps apply after dimension-specific scoring and do not automatically zero other dimensions.

Condition	Maximum score
Critical signals or values clearly contradicted by supplied evidence, impossible in magnitude, or from a disallowed hidden source	D1: 1; D3: 2
Future or cutoff-after data	D1: 0
Opposite conclusion driven by unlisted auxiliary evidence or a wrong comparison window	D1: 2; D3: 2; D5: 3
Unsupported diagnosis or strong medical/causal claim	D3: 1; D4: 1
Mostly generic boilerplate	D5: 2
Inference presented as direct observation	D1: 2; D4: 2

invalid scores are excluded from averages rather than imputed as zero; valid zero scores are retained. The main causes of incomplete coverage were context-window overflows for Direct LLM and failure to produce a final answer within the prescribed interaction budget for ReAct Agent.

Table 8: Model–setting combinations with incomplete primary-judge coverage. All other combinations have 3,000 valid scores. Missing scores are excluded from averages, not assigned zero.

Answer model	Setting	Valid scores
Qwen3.5-35B-A3B	Direct	2,930
Qwen3.5-122B-A10B	Agent	2,950
Qwen3.5-397B-A17B	Agent	2,950
DeepSeek-V4-Flash	Agent	2,876
DeepSeek-V4-Pro	Agent	2,877

Aggregation and paired checks. Each model–setting mean uses its available valid primary judgments, so the difference between Direct and Agent means is not necessarily a paired effect. On the within-model intersections of valid questions, Agent still scores higher for all ten models. Each intersection has the same size as the smaller of the two valid-score sets. This check does not remove possible selection effects from missing outputs.

B.3 Judge Agreement Analysis

Following the use of agreement checks to characterize LLM judges (Zheng et al., 2023), we assess robustness at both the response-score and aggregated model/system levels. Kimi-K2.5 is the primary judge; DeepSeek-V4-Flash and Qwen3.5-397B-A17B are alternative judges for this main-benchmark analysis. The paired-score statistics below cover the seven open-weight answer models under Direct LLM and ReAct Agent, not the three proprietary answer models or PHM. They are separate from the controlled Master500 study reported later in this appendix.

We match judgments by question, answer model, and answering setting, retaining only responses with valid scores from both judges in each pair. Across the seven open-weight models and two settings, Kimi, DeepSeek, and Qwen provide 41,583, 41,583, and 41,733 valid judgments, respectively. Table 9 reports pooled total-score agreement and dimension-level within-one-point agreement on the corresponding pairwise intersections. Sample counts refer to scored responses, not unique questions: one question can contribute responses from multiple models and settings. These pooled correlations summarize score agreement and do not establish clinical validity.

Table 9: Cross-judge agreement for seven open-weight answer models under Direct LLM and ReAct Agent. Total scores are out of 20; within-one-point rates use the same pairwise intersections and unnormalized dimension scores.

(a) Total-score agreement

Judge pair	Matched responses	Spearman ρ	MAE
Kimi–DeepSeek	41,582	0.793	1.659
Kimi–Qwen	41,583	0.795	2.463
DeepSeek–Qwen	41,582	0.816	2.264

(b) Within-one-point agreement (%)

Judge pair	D1	D2	D3	D4	D5
Kimi–DeepSeek	99.6	99.9	99.2	99.6	97.0
Kimi–Qwen	98.6	97.1	98.2	98.8	92.2
DeepSeek–Qwen	98.8	96.6	98.4	99.1	93.8

Total-score correlations are strong, while MAE shows that absolute score calibration differs across judges. Within-one-point agreement is high for all dimensions, but the tolerance is one raw point and the dimension maxima differ; these percentages should not be interpreted as directly comparable normalized errors.

Across the seven open-weight answer models evaluated under Direct LLM and ReAct Agent, the Agent consistently outperforms its corresponding direct-inference counterpart under each of the three judges. These results support the robustness of this Direct–Agent comparison to evaluator choice within the evaluated coverage; they do not establish cross-judge robustness for PHM.

Table 10: Overall scores assigned by different judges. Kimi, DeepSeek, and Qwen denote Kimi-K2.5, DeepSeek-V4-Flash, and Qwen3.5-397B-A17B, respectively. Kimi-K2.5 is the primary judge used in the main paper. Each entry reports the mean 20-point benchmark score for a given answer model and answering system.

Answering System	Answer Model	Kimi	DeepSeek	Qwen
Direct LLM	Qwen3.5-35B-A3B	5.63	5.23	4.35
	Qwen3.5-122B-A10B	6.00	5.42	4.48
	Qwen3.5-397B-A17B	5.67	5.14	4.16
	DeepSeek-V4-Flash	6.21	6.28	5.10
	DeepSeek-V4-Pro	6.16	6.28	5.03
	Kimi-K2.5	7.02	6.92	6.13
	GLM-5	6.81	6.67	6.09
ReAct Agent	Qwen3.5-35B-A3B	7.91	8.57	7.43
	Qwen3.5-122B-A10B	8.20	8.79	8.27
	Qwen3.5-397B-A17B	8.43	9.04	8.58
	DeepSeek-V4-Flash	7.89	8.15	6.98
	DeepSeek-V4-Pro	8.26	8.53	7.08
	Kimi-K2.5	7.83	8.08	7.30
	GLM-5	8.47	8.75	8.37

B.4 Complete Judge Prompt

Judge input construction. The structured evidence specification and fixed dimension anchors are serialized together with the candidate answer. The visibility/scoring context retains only `question_time`, `target_signal`, and `baseline_credit_policy`, when present. The compressed judge-only evidence context removes dataset/person/case identifiers and internal operator identifiers. The specification defines evidence and attribution constraints, not a reference answer. These evaluation-only inputs are withheld from all answering systems.

The system prompt and user-message template below reproduce the confirmed evaluation prompt. Placeholder values in the user message are JSON-serialized; line wrapping here is for readability. The judge returns a score and reason for each dimension, a total score, and an overall summary.

System prompt.

You are a strict LLM-as-judge for wearable-health benchmark answers.

Grade the answer with the fixed wearable-health rubric using only the user question, structured evidence specification, fixed rubric dimensions, seed context, and model answer supplied by the user message.

Use only the supplied inputs:
- user_question
- structured_evidence_spec
- fixed_rubric_dimensions
- seed_context
- compressed rubric evidence context
- model_answer

Scoring authority:
- The fixed rubric dimensions define what each dimension scores.
- Each dimension's scope, high_credit, partial_credit, and heavy_penalty anchors define the dimension-specific scoring range.
- The structured evidence spec is not a gold answer; it defines observable evidence, supported/unsupported attributions, and expected uncertainty.
- Award credit only for answer content that is present or clearly implied in the model answer.

Capability preservation:
- Preserve partial credit for demonstrated reasoning capabilities even when evidence fidelity is imperfect.
- Do not collapse all dimensions to zero because of one category of error.
- Each dimension should evaluate its own capability independently whenever possible.
- Clearly contradicted, disallowed, or out-of-window evidence claims should strongly affect evidence-related dimensions, but should not automatically collapse unrelated reasoning dimensions.

- If the answer demonstrates coherent reasoning structure, contextual synthesis, or multi-signal integration, preserve partial credit even when some evidence is unsupported.

Evidence-status definitions:
- Critical evidence is the evidence named in structured_evidence_spec. Missing or misusing critical evidence should strongly reduce D1 and D3 and should prevent high credit, but it is not by itself a reason to zero D2 or D4.
- Unlisted auxiliary evidence is any answer-cited metric, value, timestamp, or context that is not listed in critical_evidence. The rubric evidence is not exhaustive, so absence from critical_evidence, structured_evidence_spec, or compressed rubric evidence context is not evidence against the value.
- Treat unlisted auxiliary evidence as allowed and non-critical by default. Do not penalize it merely because the judge cannot verify it in the rubric evidence.
- Penalize unlisted auxiliary evidence only when it is clearly contradicted by supplied evidence, impossible in magnitude, from future/cutoff-after data, from a disallowed hidden source, or used to drive the opposite main conclusion.
- Conclusion mismatch means the answer uses visible auxiliary evidence, an approximate value, or a wrong comparison/window to reach the opposite main conclusion from the structured evidence spec. Penalize affected dimensions, especially D1/D3/D5, but do not automatically zero D2/D4 if reasoning structure or uncertainty awareness is present.
- Severe evidence error requires clear evidence of impossible magnitudes, future/cutoff-after data, disallowed hidden data, or claims contradicted by the supplied evidence. Only these clear evidence errors should trigger evidence caps.

Scoring procedure:
- Start with additive scoring: identify what the answer does well within each dimension, then subtract proportionally for problems in that dimension.
- Apply score caps only after additive scoring. A cap limits the maximum score for affected dimensions; it does not force other dimensions to zero.
- Do not use automatic zeroing or hard-collapse scoring except for empty/nonresponsive answers or dimensions with no demonstrated capability.
- Exact wording is not required, and exact numeric match is not always required.
- Reasonable conclusions with minor numeric or evidence-fidelity issues should receive partial credit, but not full credit.
- Primary supported attribution can receive high Attribution Quality; secondary acceptable attribution can receive medium credit; unsupported attribution can receive low credit only when there is still plausible reasoning structure.
- For usual/typical/normal/baseline/routine claims, require a clearly bounded visible comparison window for high credit; sparse short clusters can still receive partial credit when used cautiously.
- Unlisted auxiliary evidence is allowed even when it is absent from the supplied scoring context; treat it as neutral or secondary unless it contradicts supplied evidence, leaks future/hidden data, is impossible, or drives the opposite main conclusion. It cannot by itself earn high credit or replace critical evidence.
- If unlisted auxiliary evidence, approximate values, or a wrong comparison window drive the opposite conclusion from the structured evidence spec, treat it as conclusion mismatch and subtract heavily in the affected dimensions.
- If an answer is directionally consistent with the primary conclusion but adds too many secondary metrics, penalize redundancy mainly in D5 and lightly in D2/D4; do not treat directionally consistent redundancy as a cross-dimension failure.

Score caps:
- If the answer uses critical signal(s) or cited values clearly contradicted by supplied evidence, impossible in magnitude, or from a disallowed hidden source, D1 maximum score = 1 and D3 maximum score = 2. Preserve D2/D4/D5 partial credit when reasoning structure, uncertainty awareness, or case synthesis is present.
- If the answer uses future/cutoff-after data, D1 maximum score = 0. Other dimensions should still be scored independently unless their reasoning depends on the leaked data.
- If the answer relies on unlisted auxiliary evidence or a wrong window to reach the opposite conclusion, D1 maximum score = 2, D3 maximum score = 2, and D5 maximum score = 3.
- If the answer makes an unsupported diagnosis or strong medical/causal claim, D3 maximum score = 1 and D4 maximum score = 1.
- If the answer is mostly generic boilerplate, D5 maximum score = 2.
- If the answer treats inference as direct observation, D1 maximum score = 2 and D4 maximum score = 2.
- If the answer is empty or nonresponsive, all dimensions may be 0.

Dimension-specific guidance:

- D1 Data Accuracy & Localization: severe evidence errors can cap D1, but preserve credit for correct temporal localization, correct visible values, or correct signal selection that remains present.
- D2 Multi-Signal Reasoning: even if some signals are unsupported, preserve partial credit for coherent observation-to-inference structure, cross-signal synthesis, or multi-step interpretation.
- D3 Attribution Quality: primary attribution > secondary attribution > unsupported attribution. Unsupported attribution should be capped, but plausible prioritization or explanatory structure can retain low-to-medium credit.
- D4 Uncertainty Calibration: reduce scores for overconfidence, but preserve credit for explicit uncertainty awareness, limitations, or distinction between known facts and possible explanations.
- D5 Context-Aware Reasoning: preserve credit when the answer genuinely addresses this user's question, time window, trends, or personal context, even if some cited evidence is inaccurate.

Partial credit mindset:
- Exact wording is not required.
- Evaluate what capability the answer demonstrates before deciding what cap applies.
- Prefer continuous score separation over all-or-nothing judgments.

Return only valid JSON with this exact shape:
{
"Data Accuracy & Localization": {"score": <number 0-4>, "reason": "concise but specific reason"},
"Multi-Signal Reasoning": {"score": <number 0-4>, "reason": "concise but specific reason"},
"Attribution Quality": {"score": <number 0-4>, "reason": "concise but specific reason"},
"Uncertainty Calibration": {"score": <number 0-3>, "reason": "concise but specific reason"},
"Context-Aware Reasoning": {"score": <number 0-5>, "reason": "concise but specific reason"},
"total_score": <number 0-20>,
"overall_summary": "one concise summary of the main scoring rationale"
}

User-message template.

User question:
{user_question}

Structured evidence specification:
{structured_evidence_spec}

Fixed rubric dimensions:
{fixed_rubric_dimensions}

Visibility/scoring context:
{seed_context}

Compressed rubric evidence context:
{rubric_evidence_context}

Model answer:
{model_answer}

C METHOD DETAILS

C.1 MEMORY CONSTRUCTION

Local and event-level memory. PHM organizes each user's visible wearable records into local, event-level, daily, and historical memories. The sequence layer uses 60-minute windows with a 60-minute step and requires at least one sample per window. Window summaries retain statistics such as count, mean, median, standard deviation, and extrema. Sleep and exercise episodes are represented as explicit events, with additional aggregates over event-bounded intervals. Anomaly nodes describe detected local deviations, sustained shifts, volatility, or sleep-related changes, using historical sequence summaries as comparison evidence.

Daily summaries and personal reference. Daily aggregates summarize available observations rather than assuming complete days. Personalized baselines use 7-, 15-, and 30-day windows, with

distributional summaries including counts, central tendency, spread, and selected percentiles. Weekday/weekend conditioning is supported when constructing personal reference distributions. For sparse histories, adaptive baseline construction seeks at least five samples, with a configured expansion factor of two and a maximum horizon of 90 days. Trend memories summarize recent daily changes over seven days when at least three days are available.

Visibility and partial days. Source rows are filtered at the question time before memory construction. A daily aggregate's temporal window is the first and last timestamp of the rows it actually contains, not a fixed midnight-to-midnight interval. For example, a question at 10:00 may use that day's observations through 09:45, but not an eventual full-day aggregate. Historical questions use their own cutoff-specific graph or the incremental state at that cutoff; later updates must not be substituted for that state. Node and edge visibility checks also exclude items whose windows end after the question time.

Construction and updates. The implementation supports both independent per-question construction and sequential incremental ingestion. In the former, each graph is built from records visible at that question's cutoff. In the latter, cases are processed in question-time order: new sequence and event nodes are added, cross-midnight events are reconstructed from accumulated data, and affected daily, baseline, and trend memories are refreshed. Same-day aggregates can replace earlier versions; baseline identifiers include their anchor date, with at most five retained versions per signal and granularity. Incremental embedding processes new entries rather than recomputing the entire memory index.

Types, relations, and optional detail. Node types include observations, events, window aggregates, daily aggregates, baselines, trends, anomaly events, and patterns. Stored temporal relations describe ordering, overlap, and proximity; hierarchical relations connect observations to aggregates; baseline-comparison relations associate measurements with personal references; cross-signal relations encode aligned or lagged context. These are associations, not causal conclusions. None of these edges is traversed or rendered as an edge record in the reported 1-hop configuration. Optional 5-, 15-, and 30-minute drilldown windows can be triggered by anomalies or query-derived event anchors; drilldown is disabled by the default empty-window configuration.

C.2 Query-Conditioned Retrieval

The reported configuration is *1-hop*: direct selection of memory nodes, not traversal of one graph edge. Retrieval cues and evidence are reconstructed from the user question and visible history. In particular, signal families, intent, and temporal anchors are not taken from fields saved during question generation. Neither the fixed evaluation rubric nor the query-specific evidence specification is available to the retrieval pipeline.

Query parsing and temporal constraints. One call to an 8B-class chat model extracts intent, signal families, time windows, and rewritten search text, with keyword-based parsing as a fallback. Parsed windows define the temporal search range. When no window is resolved, the fallback extends from eight days before the inferred event anchor to the question time; if no event anchor is available, the question time is used as the anchor. All access remains bounded by the cutoff-specific visible memory state.

Candidate pools and score normalization. Under hybrid retrieval, embedding-based search returns up to 200 candidates by cosine similarity. Independently, rule scoring scans all nodes in the visible graph and retains positive-scoring candidates. The two pools are merged by node identity. Each channel is divided by its maximum candidate score, with a zero-denominator guard; a node absent from a channel receives zero for that channel:

$$s(v, q) = \lambda_q \widehat{s}_{\mathrm{sem}}(v, q) + (1 - \lambda_q)\widehat{s}_{\mathrm{rule}}(v, q).$$

The default hybrid policy uses $\lambda_q = 0.4$. For temporal-pattern queries, including timing, periodicity, and trend questions, the structured policy skips vector search and uses $\lambda_q = 0$. These are heuristic retrieval weights, not learned parameters.

Node-attribute rule score. Rules use signal type, time window, granularity, node type, statistics, and evidence-role attributes; they do not inspect relation edges. Signal-family matching adds 5 points. Query-dependent sleep rules add 3–6 points, with a 5-point bonus for a sleep session matching a clock anchor and a 3-point bonus for wake/end-time evidence. Heart-rate matching adds 6, anomaly evidence 4, and associated activity 2. Point-event local context within 90 minutes receives 8 points plus an anchor-dependent bonus. Recovery-related matches add 3–4 points, with 3 for activity-component evidence. Daily, baseline, or trend nodes matching temporal-pattern intent receive 3 points; preferred granularities receive 1.5, and a baseline node matching an explicit baseline keyword receives 2. Temporal proximity contributes 5, 3, or 1 point for the three-hour, eight-day, or thirty-day neighborhoods. An anomaly more than six hours from the question is penalized by $\min(4, (h-6)/24)$, where h is its distance in hours. Incompatible point-event candidates are assigned zero and excluded.

Retrieval-stage quotas and coverage. Selection uses `final_top_k=120`, independently of the later prompt limit. Default role-specific quotas are 12% component breakdown, 14% event-bounded statistics, 12% target events, 12% personal baselines, 16% recent comparisons, 12% anomalies (capped at 15), 8% linked/other evidence, and 18% raw signals. These are role-specific allocation rules, not a probability distribution. Unfilled slots are supplemented through overflow and type-diverse selection, whose default per-type cap is 30. The linked-evidence role does not introduce edge records into 1-hop retrieval.

For the layer-balanced temporal policy, allocations are 45% events, 30% daily aggregates, 10% baselines, 10% trends, and 5% other items, with baseline and anomaly caps of 10 and 15. Layer quotas apply when enabled by the selected policy; the rule-only temporal policy and its layer-balanced variant should not be conflated. Coverage checks then seek both actual measurements and comparison evidence for requested signals, supplementing missing roles from the merged candidate pool when available.

Reranking and prompt-stage selection. After quota selection and coverage checks, a roughly 4B reranker scores the question against each selected item's text and reorders the same set. Prompt construction subsequently deduplicates items, applies question-focused ordering, enforces a second set of role budgets, protects requested-signal coverage, and truncates to at most 90 items. Thus reranker order need not equal final prompt order.

Prompt role budgets (default / time-series) are 26/24% target events, 16/34% daily sequences, 14/16% baselines, 6/10% trends, 30/12% local raw evidence, and 10/7% anomalies. Unsupported proxies are capped at three items; other items have no separate role cap but remain subject to the overall limit. These budgets act at the prompt stage, not on the 120-item retrieval selection. Item limits do not imply matched token or computation budgets across answering systems.

Failure handling. If vector retrieval fails, rule-based retrieval remains available. If reranking fails, the selected order is retained. Empty retrieval is explicitly reported as *No retrieved graph items were available after filtering.* The reported 1-hop configuration does not expand graph neighborhoods; the results do not isolate benefits of graph traversal or individual relation families.

C.3 Evidence Context and Answer Generation

Model-visible context. The context contains the question and time, dataset/person/case identifiers, and numbered evidence items. Each item exposes its kind, signal, granularity, actual temporal window, a summary of at most 750 characters, and at most 20 statistical keys. Interpretation notes and question-focused comparison summaries may also be included. Retrieval scores, raw source rows, node identifiers, and technical metadata are not exposed in the primary configuration; raw-source traceability is retained outside the prompt. Personal-baseline contrasts can therefore be visible as facts or summaries without supplying graph edges.

Answer generation and comparison scope. The answer model receives the following system and user templates and generates one response without further file inspection. Single-pass refers to this answering stage, not to the separate query-parser call or retrieval computation. The PHM prompt differs from the Direct LLM prompt, notably in its explicit treatment of missing target signals and

unsupported proxies. Consequently, the comparison evaluates the complete memory, selection, and prompting pipeline rather than an isolated retrieval change under identical prompts.

System prompt.

```
You answer wearable-health questions using only the provided visible
GraphRAG context. Be concrete about dates, times, windows, baselines,
and uncertainty. Do not invent hidden data. When evidence is
incomplete, say what can and cannot be concluded. Do not replace an
unavailable target signal with a proxy; for example, wearable calories
are energy expenditure, not food intake. Do not cite internal
detector labels, z-scores, trigger names, or graph mechanics;
translate retrieved anomaly cues into visible values and
plain-language comparisons.
```

User prompt.

```
User question: {user_question}
Question time: {question_time}
{graphrag_context}

Write a concise but complete answer. Prioritize the exact data
windows and baselines that answer the question; mention alternative
explanations only when supported by the retrieved context. If the
requested data type is not present in the visible context, lead with
that limitation before discussing indirect context.
```

`GraphRAG` is retained here to reproduce the implementation's prompt wording; the reported configuration remains direct node retrieval without graph traversal.

Illustrative rendered evidence. The following abbreviated context uses fictional values, not a benchmark response or an evaluation evidence specification. The question time is 2019-12-09 at 10:00. Only two evidence items are shown; the current day's summary ends at the last visible reading, 09:45, and is explicitly partial-day.

```
User question: Is my resting heart rate today unusual
compared to my normal range?
Question time: 2019-12-09T10:00:00

Actual vs baseline contrast:
Resting HR on 2019-12-09 morning: 68 bpm vs 7-day baseline
median 62 bpm (IQR 58--67), delta +6 bpm, above typical
range.

Evidence 1
Data: daily_aggregate; metric=resting_heart_rate_bpm;
granularity=daily
When: 2019-12-09T06:00:00 to 2019-12-09T09:45:00
Summary: Resting heart rate daily summary for 2019-12-09
(partial day: 14 morning readings): mean=68.0, median=68,
n=14
Values: {"mean":68.0, "median":68, "min":65, "max":72,
"std":2.1, "n":14}

Evidence 2
Data: baseline; metric=resting_heart_rate_bpm;
granularity=rolling_7d_baseline
When: 2019-12-02T06:00:00 to 2019-12-08T23:55:00
Summary: Resting HR 7-day rolling baseline (median: 62
bpm, IQR: 58--67)
Values: {"median":62, "q25":58, "q75":67, "n":7,
"window_days":7}
```

D Additional Experimental Results

D.1 Answering-System Details

We compare three answering systems that operate over the same processed wearable records and obey the same question-time visibility constraint, but differ in how longitudinal evidence is represented and accessed. For all settings, only observations available at or before the question time are eligible for use. The evaluation rubric, query-specific evidence specification, and future observations are never exposed to the answering system.

Direct LLM. The Direct LLM receives a fixed preconstructed representation of the visible wearable history and generates the final answer in a single model call. To control context length while retaining recent temporal detail, high-frequency signals from the most recent 7 days are represented at fine temporal resolution, while older observations within the visible history are compressed into daily summary statistics. Other lower-frequency modalities are retained for up to 30 days when available. For answer models whose context limits cannot accommodate the default representation, the recent fine-grained window is reduced to 5 days while preserving the same preprocessing policy. The controlled study in Sec. 5.3 replaces this default construction with explicitly controlled history-length and representation conditions.

ReAct Agent. The ReAct Agent retains on-demand access to the visible wearable records rather than receiving the full longitudinal representation in a single prompt. It may iteratively inspect relevant time windows, retrieve underlying measurements, compute summary statistics, and perform numerical analysis before producing the final answer. The Agent is implemented with a constrained set of 12 data-access and computation tools and is allowed up to 20 reasoning/tool-use rounds per question. Numerical operations are executed in a sandboxed environment, and all tool calls remain subject to the same question-time cutoff as the other answering settings.

PHM. PHM retrieves memory entries before generating one answer, without iterative tool use during answer generation. The confirmed 1-hop retrieval and prompt-construction stages are detailed in Appendix C.

Comparison scope. The systems share eligible source records and the query-time cutoff, but not identical token-level inputs: Direct receives a fixed representation, ReAct accesses records through tools, and PHM receives retrieved memory. The comparison evaluates these complete answering pipelines. Numerical results and aggregation follow Appendix B.2.

D.2 Controlled Context and Granularity Study

This section provides supporting details for the controlled study in Sec. 5.3. We evaluate input history length and temporal representation on a fixed subset of 500 questions from 115 individuals. The subset contains 100 questions from Fitbit Fitness Tracker Data, 200 from LifeSnaps, and 200 from PMData. It was selected before inspecting model answers or judge scores and is held fixed across every experimental condition. For each of two answer models, the same question is answered under all nine combinations of 3-, 7-, and 30-day history with Coarse, Fine, and Multiscale representations, yielding 9,000 answers. Kimi-K2.5 and Claude Sonnet independently score every answer, yielding complete coverage of 18,000 judgments.

Controlled input representations. We define three deterministic representations over the same query-time-visible history. Fine aggregates high-frequency continuous signals into fixed 5-minute (300-second) bins while preserving discrete events, sleep and exercise sessions, and native daily records. Coarse aggregates continuous signals into fixed 1-hour (3,600-second) bins. Aggregation is signal-specific: continuous physiological signals retain statistics such as mean, median, and range, whereas count- and volume-like signals are summed; missing values are never interpolated. Multiscale retains 5-minute resolution for the most recent 24 hours, represents earlier observations within the selected history window using 1-hour summaries, and appends daily baseline and trend summaries computed from the same visible snapshot. All representations use identical history hori-

Table 11: Complete score matrices for the controlled context-and-granularity study. Each cell averages the same 500 paired questions and is scored on the 20-point rubric.

		Kimi-K2.5			Claude Sonnet		
Answer Model	**History**	**Coarse**	**Fine**	**Multiscale**	**Coarse**	**Fine**	**Multiscale**
Qwen3.5-397B	3d	5.94	5.66	5.88	6.96	6.69	7.13
	7d	5.58	5.38	5.67	6.51	6.48	7.01
	30d	5.08	4.86	5.23	5.93	5.63	6.52
DeepSeek-V4-Flash	3d	6.07	6.11	6.29	7.21	7.07	7.44
	7d	6.03	5.96	6.14	7.16	6.83	7.24
	30d	5.58	5.37	5.72	6.67	6.47	6.77

Table 12: Secondary-judge robustness of the controlled contrasts. Score differences are averaged over the other experimental factor; brackets show 95% person-cluster bootstrap confidence intervals. Bold differences have confidence intervals excluding zero.

Answer Model	**30d − 3d**	**Fine − Coarse**	**Multiscale − Coarse**
Qwen3.5-397B	$-0.90\ [-1.13, -0.65]$	$-0.20\ [-0.34, -0.07]$	$+0.42\ [0.26, 0.57]$
DeepSeek-V4-Flash	$-0.60\ [-0.82, -0.36]$	$-0.22\ [-0.35, -0.09]$	$+0.14\ [-0.01, 0.29]$

zons and strict query-time cutoffs, and neither future observations nor question-specific evidence annotations are exposed during construction.

Paired contrasts and evaluator robustness. History effects are defined as the paired 30d-minus-3d score difference averaged over the three representations. Representation effects are defined relative to Coarse and averaged over the three history lengths. Confidence intervals use person-cluster bootstrap so that repeated questions from the same individual remain in the same resampled cluster. Primary-judge contrasts are summarized in Figure 3; Table 12 reports the same contrasts under the secondary judge.

Although Claude Sonnet assigns scores 0.874 points higher on average, the two judges show substantial question-level agreement (Spearman $\rho = 0.788$) and highly consistent rankings across the 18 answer-model–condition cells (Spearman $\rho = 0.975$). The secondary judge reproduces the direction of every main history-length contrast and every aggregate representation contrast.

Prompt-token usage. Table 13 reports average prompt tokens for the 30-day conditions. Token counts are compared only within an answer model because provider tokenization differs across models. Multiscale reduces prompt tokens by 52.7–54.2% relative to Fine while retaining more local detail than the uniformly Coarse representation.

Table 13: Average prompt tokens for 30-day inputs. Reduction compares Multiscale with Fine within each answer model.

Answer Model	**Coarse**	**Fine**	**Multiscale**	**Reduction**
Qwen3.5-397B	26.6K	69.3K	31.7K	54.2%
DeepSeek-V4-Flash	15.8K	41.0K	19.4K	52.7%