WEARABLEQA: A BENCHMARK FOR HEALTH REASONING OVER REAL-WORLD WEARABLE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in wearable sensing enable continuous monitoring of physiological and behavioral signals, yet the ability of LLMs to reason over longitudinal wearable data remains underexplored. Available benchmarks also largely rely on synthetic or simulated signals, leaving it unclear how well current LLMs can reason over a *real user's* longitudinal wearable measurements. We here introduce **`WearableQA`**, a benchmark comprising 4,084 ten-option multiple-choice questions constructed from the wearable time series, blood biomarkers, and demographics of 200 real users, each with up to about 500 days of daily measurements. `WearableQA` preserves authentic wearable distributions that include device noise and inter-individual variability. To evaluate distinct reasoning capabilities, we introduce 16 question types organized along two complementary axes: data versus health reasoning, which distinguishes computation over longitudinal raw measurements from physiological interpretation; and single- versus cross-signal reasoning, which separates reasoning about individual signals from the integration of multiple signals. To construct reliable questions at scale, we adopt a dual-grounding framework that combines literature-grounded physiological findings with statistically validated population-grounded physiological patterns. This enables the capture of meaningful relationships observed in real-world wearable data. Evaluation of 14 proprietary and open-source LLMs demonstrates that `WearableQA` effectively differentiates model capabilities, with performance ranging from 19.6% to 72.9% against a 10% chance baseline. Moreover, `WearableQA` remains far from solved: most models achieve accuracies below 60%. Overall, `WearableQA` provides a realistic and diagnostic benchmark for evaluating LLM reasoning over real-world wearable data.

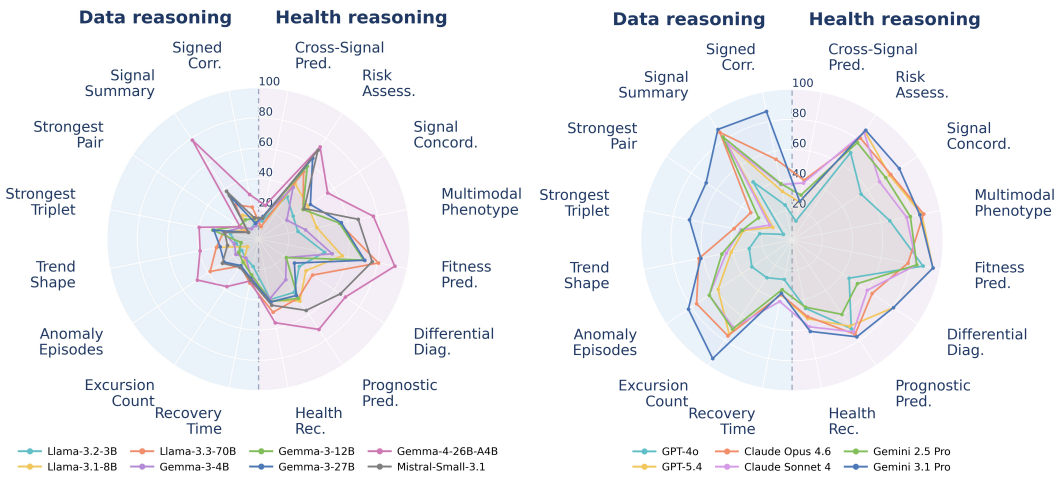


Figure 1: **Radar plot for performance of Open-Source models (left) and proprietary models (right) on our `WearableQA` Benchmark.**

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.

1 INTRODUCTION

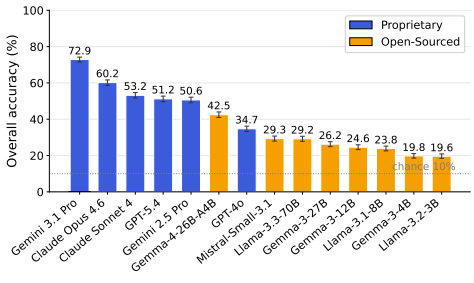


Figure 2: **Overall accuracy (CoT, 95% CI).**

Recent advances in wearable sensing technologies have enabled continuous monitoring of physiological and behavioral signals, including heart rate, sleep, physical activity, and heart rate variability. Combined with the rapid progress of large language models (LLMs) (OpenAI, 2024; Singh et al., 2025; Anthropic, 2026; Google DeepMind, 2026; Comanici et al., 2025; Grattafiori et al., 2024), daily longitudinal multimodal measurements are increasingly used to support personalized health understanding, lifestyle assessment, and health monitoring (Kim et al., 2024; Li et al., 2025; Merrill et al., 2026; Khasentino et al., 2025; Pillai et al., 2025; Narayanswamy et al., 2026). Reasoning with wearable data, however, requires more than simply recalling general health knowledge. The model needs to first identify and compute the information from the noisy longitudinal physiological measurements, and then interpret it in the context of health, which can also be user-specific. Yet, the ability of LLMs to reason over longitudinal wearable data remains underexplored, with only recent studies beginning to investigate this capability (Kim et al., 2024; Merrill et al., 2026; Khasentino et al., 2025). Systematic benchmarks for evaluating this capability are also scarce. Existing ones largely rely on synthetic or simulated signals (Cai et al., 2024; Merrill et al., 2026) or general-domain time series (Jing et al., 2026; Gwiazda et al., 2026) rather than real-world wearable measurements. Consequently, it remains unclear whether current LLMs can reason about the daily longitudinal wearable history of a *real user* and determine what it implies about the user's health.

To this end, we introduce **`WearableQA`**, a benchmark for data and health reasoning on real-user longitudinal wearable records. `WearableQA` contains 4,084 multiple-choice questions constructed from 200 real users' wearable measurements collected over hundreds of days per individual, which also integrates complementary blood biomarkers and demographic information. Each instance naturally captures the inter-individual variability, user-specific baselines across different time spans, and signals. `WearableQA` is structured to evaluate model capabilities along two axes: reasoning type and signal complexity. Reasoning types include data and health reasoning, where *data reasoning* computes trends, anomalies, and relationships from the daily aggregated wearable measurements, while *health reasoning* requires clinical and physiological interpretation. Second, the signal complexity axis distinguishes *single-signal* questions from *cross-signal* questions, where the latter challenge the model to integrate patterns and information from two or more signals. Together, these axes enable a fine-grained diagnosis of whether model failure arises from the computation over longitudinal measurements, health-related interpretation, or the integration of different signals.

Given the nature of the domain, constructing reliable reasoning questions from longitudinal wearable data at scale is challenging. The questions need to reflect physiological patterns that genuinely arise in real-world data, while allowing the available wearable measurements to be deterministically annotated according to those patterns. To address these challenges, we adopt a dual-grounding framework composed of two complementary sources: Peer-reviewed literature-grounded findings and statistically validated population-grounded patterns from a large cohort. Then these objectives are instantiated on each user's measurements through deterministic computations that are refined through careful audit and feedback loops. Overall, `WearableQA` comprises 16 question types, evenly divided between data reasoning and health reasoning questions.

Our experiments on `WearableQA` across 14 proprietary and open-weight models show that the benchmark effectively distinguishes models with widely varying levels of performance, ranging from 72.9% to 19.6% against a 10% chance baseline. `WearableQA` also enables a fine-grained diagnosis of model capabilities across reasoning type and signal complexity. We observe that models generally fall short in data reasoning when it comes to deriving answers directly from the raw measurements. In addition, cross-signal reasoning remains challenging, with stronger models showing clear drops relative to single-signal reasoning questions, while open-source models generally perform substantially lower on both subsets. Overall, we hope `WearableQA` provides a discriminative and diagnostic testbed for advancing LLMs that can reason reliably over real-world wearable data.

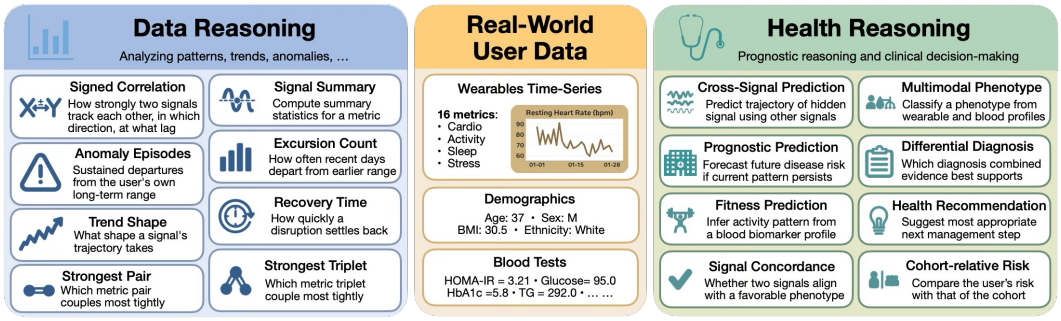


Figure 3: **Overall taxonomy of `WearableQA`.** Taxonomy for 4,084 questions of real-world user data that span blood, demographics, and time-series wearables. The benchmark covers 16 question types of data-reasoning and health-reasoning.

2 Related Work

2.1 Health-related Benchmarks

A growing body of work evaluates language models on medical and health reasoning. Benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MedXpertQA (Zuo et al., 2025) primarily assess medical knowledge through examination-style questions and biomedical text, while more recent benchmarks such as HealthBench (Arora et al., 2025) extend evaluation to more realistic health-related interactions. Despite their importance, these benchmarks are largely text-based and focus on static questions or cases, providing limited insight into whether models can reason over an individual's longitudinal physiological history. ECG-QA (Oh et al., 2023) and subsequent benchmarks over high-frequency or short-term sensor data (Li et al., 2026) move beyond purely textual inputs by incorporating sensor and physiological measurements. However, reasoning over a waveform differs from reasoning over longitudinal wearable measurements, where the models must aggregate measurements across time and integrate multiple behavioral and physiological signals; this setting remains underexplored. More recently, several studies have begun to evaluate LLM reasoning over wearable measurements (Kim et al., 2024; Merrill et al., 2026; Khasentino et al., 2025; Narayanswamy et al., 2026). For instance, PHIA (Merrill et al., 2026) explores questions over wearable data, but its objective evaluation is based on simulated user trajectories and primarily focuses on retrieval and aggregation over the measurements. Existing wearable evaluations also typically focus on wearable signals alone, without jointly incorporating complementary health information such as blood biomarkers.

2.2 Time-Series Reasoning Benchmarks

Recent benchmarks have extended LLM evaluation to time-series reasoning that includes numerical reasoning, forecasting, and question answering over sequential data (Cai et al., 2024; Jing et al., 2026; Merrill et al., 2026; Gwiazda et al., 2026; Merrill et al., 2024; Kong et al., 2025). These works provide systematic evaluation beyond static textual inputs, but largely rely on synthetic or simulated signals (Cai et al., 2024; Merrill et al., 2026) or general-domain time series (Jing et al., 2026; Gwiazda et al., 2026) rather than real-world longitudinal wearable measurements. As a result, they do not capture the complexity of real-world longitudinal wearable measurements or evaluate health interpretation grounded in physiological signals. For example, TimeSeriesExam (Cai et al., 2024) evaluates general time-series understanding through procedurally generated synthetic signals spanning pattern recognition, noise understanding, similarity analysis, anomaly detection, and causality. Consequently, they provide limited evaluation of whether models can both derive information from real-world daily aggregated wearable measurements and interpret those patterns physiologically. `WearableQA` targets this setting using real longitudinal wearable measurements collected from real users, together with blood biomarkers and demographic information. It jointly evaluates data reasoning and health reasoning through a diagnostic taxonomy of 16 question types, with deterministic ground truth constructed from literature-grounded physiological findings and statistically validated population-grounded patterns.

3 THE WEARABLEQA BENCHMARK

3.1 REAL-WORLD USER DATA

WearableQA comprises longitudinal wearable and blood-panel records from 200 real users, sampled from a larger cohort of in-the-wild wearable time-series data with high coverage across taxonomies. These 200 users span a diverse range of demographic characteristics.

Signals. Each user contributes a longitudinal multivariate time series comprising 16 daily wearable metrics across four physiological domains. Cardio-fitness metrics include resting heart rate, heart-rate variability, and $\mathrm{VO_2max}$. Activity and energy metrics include steps, active calorie burn, BMR calories, exercise count, exercise duration, average exercise heart rate, and METs. Sleep metrics include duration, efficiency, deep-sleep percentage, REM (rapid eye movement) percentage, and bedtime regularity. We also include the average stress level. Each question is accompanied by up to about most recent 500 days of longitudinal records, represented as daily aggregated measurements. Each user is also associated with a 17-biomarker blood panel *e.g.*, insulin, HbA1c. Demographic attributes include age, sex, BMI, and ethnicity. To support questions that compare an individual with their peers, we provide cohort-reference percentiles for five core wearable metrics: resting heart rate, steps, heart-rate variability, sleep duration, and sleep efficiency. The reference percentiles are computed over a larger cohort of real-world users.

3.2 TAXONOMY

WearableQA is organized along a primary 2×2 taxonomy that crosses *reasoning type* with *signal complexity*, together with a separate, orthogonal *grounding* axis that records how each question's ground truth is sourced. The 4,084 questions span 16 question types (see Fig. 3), where there exists a total of 2.7k questions of data reasoning and 1.4k questions for health reasoning. Among them, 1.7k questions ask models to ground a single metric, while the remaining 2.4k require multi-signal grounding. Each question is formulated as a 10-option multiple-choice problem, yielding a chance-level accuracy of 10%, with answer positions uniformly balanced across the full benchmark.

Reasoning: Data vs. Health. Data-reasoning questions require computing over the raw sensor measurements, such as anomalies and excursions, with the answer determined by the individual wearable records. Health-reasoning questions instead require clinical and physiological interpretation, such as prognostic framing, layering domain knowledge on top of the signals.

Signal Complexity: single vs. cross-metric. Single-metric questions concern one signal in isolation, whereas cross-metric questions require integrating two or more signals, for example, a coupling between physical activity and resting heart rate, or a multi-signal phenotype.

Grounding: literature vs. population. Orthogonal to the primary axes, each question also carries a label for how its ground truth was derived, either from a literature finding or from a population-grounded pattern (Sec 3.3).

3.3 BENCHMARK CONSTRUCTION

Designing questions for wearable health reasoning must consider two objectives. Questions should (i) probe clinically meaningful reasoning anchored in biomedical evidence, and (ii) reflect physiological patterns that genuinely arise in real-world wearable recordings, rather than synthetic scenarios. WearableQA reconciles the two through dual grounding: every question is grounded either in the biomedical literature or in a statistically validated pattern discovered directly from a large cohort. Under both sources, the answer is re-derived from the underlying measurements rather than asserted a priori, which keeps the benchmark faithful to the target population and resistant to memorization.

Literature-grounded questions. Biomedical literature encodes clinically validated relationships among wearable signals, physiological biomarkers, and health outcomes. We assembled consumer-wearable studies and retained only observational studies of those where wearable exposure and clinical outcome are both computable on our cohort, excluding feasibility and modality-mismatched work. Each retained study was independently retrieved and verified against its primary record *e.g.*, PubMed, and the identifiers were cross-checked across DOI, PMID, and PMCID. Overall, this yields 11 anchor papers, published in venues such as *Nature Medicine*. To guard against spurious or non-

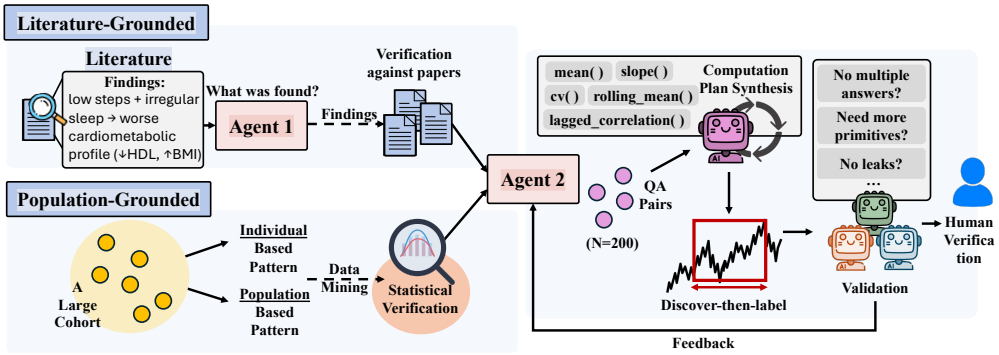


Figure 4: **Framework Overview.** From real users' wearable time series, blood, and demographics, we generate questions under dual grounding (peer-reviewed literature findings and statistically verified population-grounded patterns), compute deterministic ground truth from the raw measurements, organize by a (2×2) reasoning$\times$signal taxonomy, and apply validation loops for quality control.

reproducible findings, we require each retained relationship to be supported by **at least two independent**, same-direction studies, likewise verified against PubMed/PMC full text. Each is recorded as a structured card carrying DOI/PMID, verbatim thresholds, cohort details, and an explicit statement of what the source does *not* establish, which prevents downstream questions from over-claiming beyond the evidence. We deliberately do not import study-specific thresholds or decision boundaries, since such thresholds are cohort-dependent (*e.g.*, defining 'high exercise' as $\geq$ 5,000 steps).

Population-grounded questions. Established literature may not fully cover the diversity of relationships present in large-scale wearable data. We therefore mine physiological patterns directly from a large cohort, surfacing temporal trends, anomalous episodes, recovery dynamics, cross-signal relationships, and higher-order multi-signal interactions, specifically targeting a 28-day window. A candidate pattern is promoted to a question only if it survives a sequence of predefined statistical-consistency gates evaluated across the cohort, spanning three complementary axes.

- **Effect size.** The target must be strong and well separated. A coupling must exceed a minimum correlation magnitude ($|\rho| \geq 0.5$) with a consistent sign in both halves of the window, and for comparative questions, the winning option must beat the runner-up by a fixed margin (≥ 0.15), so that no answer hinges.
- **Robustness.** The labeled outcome must be stable under perturbation. It must persist under bootstrap resampling of days (reproducing in $\geq 80\%$ of 30 resamples), under leave-one-out removal of any single day, and, where applicable, across a range of baseline-window choices, so that no single point or knife-edge boundary drives the answer. Near-boundary cases falling inside an explicit dead-band are rejected.
- **Authenticity.** Reproducibility alone does not imply that a pattern is real. We therefore apply a cross-user null test, recomputing each statistic on deliberately mismatched user pairings and taking the ratio of null to observed prevalence as an empirical false-discovery rate; a pattern is retained only at FDR ≤ 0.20, i.e. positive predictive value $\geq 80\%$.

3.4 Discover-then-label: Labeling user measurements

For labeling measurements, we adopt a discover-then-label approach. Specifically, a reasoning objective is first *discovered* from a literature finding or a population-grounded pattern, and its answer is then *labeled* by running a deterministic program over the individual's real measurements. Representing each instance as an explicit computation graph guarantees that identical inputs always yield identical answers, enabling deterministic and fully reproducible evaluation.

Computation primitives. Given a reasoning objective derived from the dual-grounding framework, we express the required reasoning as a composition of reusable computation primitives. For instance, we instantiate a strongest-pair question with `lagged_correlation(max_lag=3)` for each candidate signal pair over lags of up to ± 3 days, then conduct `argmax_pair()`.

Similarly, a blood-state question applies `threshold_flags(glucose, HbA1c, TG, HDL, HOMA_IR)` to compare each biomarker against predefined clinical cutoffs and aggregates the resulting flags to assign the blood state. We maintain these operations in a shared primitive library, which allows them to be subsequently reused across question types.

Answer construction. The selected primitives are executed on each user's wearable trajectory. They operationalize literature-grounded or population-grounded relationships using user-specific baselines and, when needed, cohort-relative statistics, *e.g.*, resting heart rate elevation will be ≥ 1.5 standard deviations above the user's recent 28-day mean. Instances are retained only when the resulting criterion yields a deterministic answer; undefined or ambiguous cases are discarded.

Distractor construction. Each question presents 10 options; 9 incorrect samples differ by reasoning type, reflecting what can be verified in each. Each data-reasoning template defines a distractor category appropriate to the quantity it asks about, such as the opposite trend direction, an adjacent time window, or an incorrect magnitude drawn from the cohort distribution of the same quantity. Because the answer is produced by an executable computation, we extend that computation to evaluate all ten options and retain an item only when the gold option uniquely satisfies it. Health-reasoning answers involve interpretation rather than a single computed quantity, so their distractors are instead derived from the grounded finding itself: alternative interpretations of the same physiological relationship that are clinically plausible but inconsistent with the individual's measurements.

3.5 Benchmark Validation

`WearableQA` applies two complementary validation procedures to ensure that instances are correct, reliable, and free of trivial shortcuts.

Computational validity. Population-grounded questions are retained only after passing predefined statistical consistency gates spanning effect size (a strong, well-separated target), robustness (stability under bootstrap resampling, leave-one-out day removal, and boundary dead-bands), and authenticity (a cross-user null test that bounds the false-discovery rate) so that answers reflect stable physiological relationships rather than spurious observations or single-user artifacts.

Quality control and shortcut auditing. Finally, we apply benchmark-level quality controls, including duplicate removal, answer-choice balancing, distractor verification, and removal of any context that directly reveals the ground truth. Generated questions are further reviewed by three models from different families, followed by human verification. We also explicitly audit for unintended shortcuts. For example, in an earlier version of the benchmark, a small number of highly predictable metric pairs, such as active burn and steps, frequently appeared as the strongest relationship, allowing models to answer some questions from prior knowledge without using the time series. We also identified cases where particular answer choices were associated with specific ground-truth categories, creating option-based shortcuts. When such issues are detected, we revise the corresponding generation procedure and repeat the verification process.

4 Experiments

In this section, we first describe our experimental setup and present full results for 14 LLMs, proprietary and open-source, on `WearableQA`. We then diagnose model capabilities at a finer grain, breaking performance down by reasoning type and signal complexity. Finally, we present ablation studies, test robustness to input representation, and report an evaluation of agentic models.

4.1 Experimental Setup

We evaluate a diverse set of instruction-tuned LLMs in order to analyze reasoning capability across model families and scales. Proprietary models include `GPT-4o` and `GPT-5.4` (OpenAI), `Gemini-2.5-Pro` and `Gemini-3.1-Pro` (Google), and `Claude-Sonnet-4` and `Claude-Opus-4.6` (Anthropic). For open-weight models we evaluate `Llama-3.2-3B`, `Llama-3.1-8B`, and `Llama-3.3-70B` (Meta), `Gemma-3-4B`, `Gemma-3-12B`, and `Gemma-3-27B`, `Gemma-4-26B-A4B` (Google), and `Mistral-Small-3.1` (Mistral). We evaluate every model under the same protocol. Each question is presented with its associated context, comprising wearable measurements, demographic information, blood biomarkers (when appli-

Table 1: **Comparative Performance of proprietary and open-source models on `WearableQA`.** Best and second-best results are shown in **bold** and underline, respectively.

	Reasoning		Signal Complexity		
Model	**Data**	**Health**	**Single**	**Cross**	**Avg**
Proprietary models					
Gemini 3.1 Pro	**75.4** ±1.6	**67.8** ±2.5	**72.4** ±2.1	**73.2** ±1.8	**72.9** ±1.4
Claude Opus 4.6	57.1 ±1.9	66.3 ±2.5	64.9 ±2.3	56.8 ±2.0	60.2 ±1.5
Claude Sonnet 4	47.1 ±1.9	65.2 ±2.5	59.8 ±2.3	48.5 ±2.0	53.2 ±1.5
GPT-5.4	45.0 ±1.9	63.5 ±2.6	59.1 ±2.3	45.7 ±2.0	51.2 ±1.5
Gemini 2.5 Pro	47.5 ±1.9	56.8 ±2.6	58.1 ±2.4	45.3 ±2.0	50.6 ±1.5
GPT-4o	25.3 ±1.6	53.5 ±2.7	32.3 ±2.2	36.4 ±1.9	34.7 ±1.5
Open-source models					
Gemma-4-26B-A4B	**33.8** ±1.8	**59.8** ±2.6	**41.0** ±2.4	**43.5** ±2.0	**42.5** ±1.5
Mistral-Small-3.1	20.0 ±1.5	47.9 ±2.7	27.9 ±2.1	30.3 ±1.8	29.3 ±1.4
Llama-3.3-70B	23.5 ±1.6	40.7 ±2.6	28.8 ±2.2	29.5 ±1.8	29.2 ±1.4
Gemma-3-27B	19.0 ±1.5	40.7 ±2.6	27.1 ±2.1	25.6 ±1.7	26.2 ±1.3
Gemma-3-12B	17.0 ±1.4	39.9 ±2.6	20.6 ±1.9	27.4 ±1.8	24.6 ±1.3
Llama-3.1-8B	17.4 ±1.4	36.8 ±2.6	21.2 ±2.0	25.7 ±1.7	23.8 ±1.3
Gemma-3-4B	15.0 ±1.3	29.4 ±2.4	19.3 ±1.9	20.2 ±1.6	19.8 ±1.2
Llama-3.2-3B	13.6 ±1.3	31.7 ±2.5	16.2 ±1.8	22.1 ±1.7	19.6 ±1.2

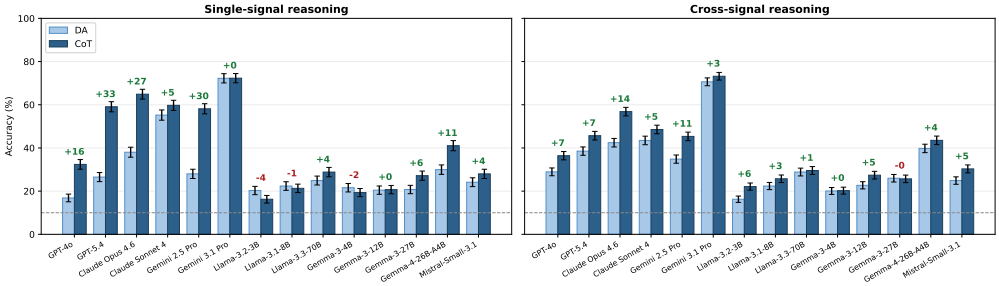


Figure 5: **Direct-answer vs. chain-of-thought across signal complexity.** Left is single-signal reasoning, with right cross-signal reasoning, which shows per-model accuracy (%) under direct answering (DA, light) and chain-of-thought (CoT, dark) with 95% confidence intervals; the number above each pair is $\Delta = \text{CoT} - \text{DA}$ and the dashed line marks the 10% chance level.

cable), and cohort reference statistics, following the benchmark definition in Section 3.2. Unless otherwise specified, each model receives the user's most recent 500 days of wearable history rendered as a row-wise representation. By default, the model reasons step by step (chain of thought) before answering, and we additionally report a direct-answer setting in which the model emits the option immediately in the appendix. We report accuracy with 95% Wilson confidence intervals, and count unparseable or empty responses as incorrect (See Appendix for more details).

4.2 Main Results

Performance spans a wide range across models. Tab. 1 reports overall accuracy for all models. Accuracy ranges from 19.6% for `Llama-3.2-3B`, near the 10% random baseline, to 72.9% for `Gemini-3.1-Pro`, showing substantial variation across model capabilities. `Gemini-3.1-Pro` also leads the next-best model, `Claude-Opus-4.6` (60.2%), by 12.7 points. Among open-source models, `Gemma-4-26B-A4B` performs best at 42.5%, outperforming `GPT-4o` by 7.8 points but remains 30.4 points behind the strongest proprietary model.

Data reasoning remains more challenging than health reasoning for most models. Performance by reasoning type reveals a consistent gap between data and health reasoning (Fig. 6 (left)). With the exception of `Gemini-3.1-Pro`, all models struggle with data reasoning. The gap is particularly pronounced for `GPT-4o` with 53.5% vs. 25.3%, `Gemma-4-26B-A4B` with 59.8% vs. 33.8%, and `Mistral-Small-3.1` with 47.9% vs. 20.0%. `Gemini-3.1-Pro` is the only model to reverse in trend, achieving 75.4% on data reasoning compared with 67.8% on health reasoning. These results suggest that directly deriving answers from real-user longitudinal measurements remains a major challenge for most models.

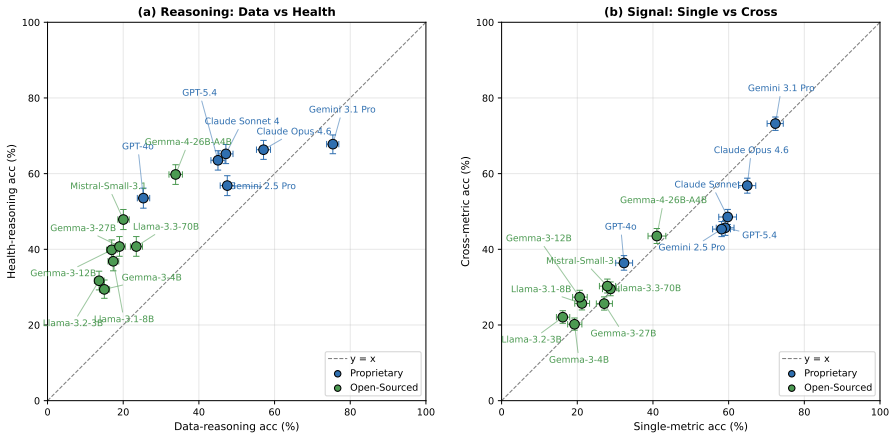


Figure 6: **Accuracy along the two taxonomy axes.** Each point is a model (95% CI); blue = proprietary, green = open-sourced, dashed line is $y = x$. **(a)** Most of the models sit above $y = x$ (health $>$ data): models reason better about clinical health than about underlying data. **(b)** Most proprietary models sit below $y = x$ (single $>$ cross); pooled accuracies mix reasoning types, and within each type nearly all models degrade on cross-signal questions (Appendix, Tab. 8).

Cross-signal reasoning remains challenging across model scales. Along the signal-complexity axis, most proprietary models perform better on single-signal than on cross-signal questions (see Fig. 6 (right)). The largest gaps are observed for `GPT-5.4` with 59.1% vs. 45.7% and `Gemini-2.5-Pro` with 58.1% vs. 45.3%, followed by `Claude-Sonnet-4` with 59.8% vs 48.5%. Open-weight models generally perform worse overall, with the strongest model `Gemma-4-26B-A4B`, reaching 41.0 on single-signal and 43.5 on cross-signal questions. Notably, `Gemini-3.1-Pro` performs strongly on both subsets, with comparable accuracy on single- and cross-signal questions, that is, 72.4% vs. 73.2%. Within each reasoning type, nearly all models, including open-weight models, degrade on cross-signal questions (Appendix, Tab. 8).

Per question-type breakdown. Fig. 1 reports accuracy across all 16 question types. Within data reasoning, performance varies substantially by task. *Strongest pair* and *recovery time* are particularly difficult for most models, whereas *signal summary* and *excursion count* are comparatively easier. For example, `Gemini-3.1-Pro` achieves 68.3% on *strongest pair* and 94.4% on *excursion count*, but only 35.9% on *recovery time*. Open-weight models perform substantially worse on most data-reasoning tasks, with several results near the 10% chance baseline. Performance is generally higher on health-reasoning tasks, with strong proprietary models exceeding 80% on *risk assessment*, *multimodal phenotype*, and *fitness prediction*. A notable exception is *cross-signal prediction*, which remains difficult across both proprietary and open-weight models: accuracy ranges from 8.3% for `Llama-3.1-8B` to only 40.5% for `Claude-Opus-4.6`. Unlike other health-reasoning tasks, cross-signal prediction requires models to infer one signal from another by combining user-specific longitudinal evidence with prior knowledge of physiological relationships.

Chain-of-thought vs. direct answer. Fig. 5 compares chain-of-thought (CoT) and direct-answer (DA) prompting across single- and cross-signal reasoning. CoT improves performance for most models, but the magnitude of the gain varies substantially across model families. Proprietary models benefit most; for example, `Claude-Opus-4.6` improves by up to 27 points on single-signal reasoning and 14 points on cross-signal reasoning. In contrast, open-weight models show more modest average gains of $+2.3$ points on single-signal and $+3.0$ points on cross-signal reasoning. These results suggest that explicit reasoning can substantially improve performance, but its effectiveness depends strongly on the underlying model rather than on signal complexity alone.

4.3 Analysis

Reliance on priors. We conduct a targeted analysis of prior-knowledge reliance using the relationship-strength question type, where definitional and empirical signal pairs can be compared under the same task formulation while controlling for association strength. We compare definitional

Table 2: **Definitional vs. empirical signal pairs.** Definitional pairs are related by construction (same-source signals) and answerable from priors; empirical pairs are physiological couplings that must be read from the data. Both groups are restricted to the strong-sync band, ensuring matched relationship strength.

Model	Definitional	Empirical	Δ
Proprietary models			
GPT-4o	50.9 ±9.5	15.0 ±11.4	+36.0
GPT-5.4	63.2 ±9.2	15.0 ±11.4	+48.2
Claude Opus 4.6	70.8 ±8.6	66.7 ±16.0	+4.0
Claude Sonnet 4	80.2 ±7.5	75.9 ±14.3	+4.4
Gemini 2.5 Pro	87.8 ±6.1	36.3 ±16.4	+51.5
Gemini 3.1 Pro	98.2 ±1.8	94.2 ±5.8	+4.0
Open-source models			
Llama-3.2-3B	18.8 ±7.4	5.8 ±5.8	+13.0
Llama-3.1-8B	17.9 ±7.2	5.8 ±5.8	+12.0
Llama-3.3-70B	40.6 ±9.3	8.9 ±8.3	+31.7
Gemma-3-4B	4.6 ±3.6	8.9 ±8.3	-4.2
Gemma-3-12B	17.9 ±7.2	15.0 ±11.4	+2.9
Gemma-3-27B	19.8 ±7.5	11.9 ±10.0	+7.8
Gemma-4-26B-A4B	69.8 ±8.7	66.7 ±16.0	+3.1
Mistral-Small-3.1	24.5 ±8.1	15.0 ±11.4	+9.5

Table 3: **Effect of input representation and tool use.** Accuracy of `GPT-5.4` under different time-series representations. Measurements are serialized row-wise, column-wise, as Markdown, or as CSV; image-grid and image-grouped denote separate plots for each signal and range-grouped plots, respectively. Stats signifies the statistical summary *e.g.*, min, max. For agentic evaluation, *fixed-repr* provides Python access with the row-wise representation, whereas *multi-view* allows the model to select among the available representations.

Input condition	Overall	Data	Health	Single	Cross
row (baseline)	51.2	45.0	63.5	59.1	45.7
col	50.3	44.0	63.1	57.9	45.0
csv	51.5	46.2	62.2	60.0	45.6
markdown	49.9	43.8	62.2	57.1	44.9
markdown+stats	51.9	46.7	62.4	61.1	45.5
image-grid	36.6	25.9	58.2	32.1	39.8
image-grouped	34.1	24.3	53.9	29.8	37.2
csv+chart	51.0	45.3	62.2	58.5	45.6
Agentic Evaluation					
agentic, fixed-repr	71.3	72.9	68.1	63.8	76.5
agentic, multi-view	69.7	71.9	65.3	60.6	76.1

pairs, such as step count and active energy expenditure, whose relationship can largely be anticipated with prior knowledge, with empirical pairs, such as resting heart rate and stress level, whose association must be inferred from the observed measurements (Tab. 2). Both groups are restricted to the same strong-coupling band. While some models show only small differences, several perform substantially better on definitional pairs, including `Gemini-2.5-Pro` (+51.5 points), `GPT-5.4` (+48.2), and `GPT-4o` (+36.0). Within this controlled setting, the results suggest that familiar signal relationships can provide a substantial advantage over comparably strong, user-specific associations that must be recovered directly from longitudinal measurements.

Effect of input representation and tool use. Tab. 3 compares alternative time-series representations and agentic evaluation with Python access. Across plain-text formats, performance changes little: CSV performs best at 51.5% overall, with only a +1.1 point gain in data reasoning over the row-wise baseline. Adding summary statistics provides slightly larger gains, with `markdown+stats` reaching 51.9% overall and improving data and single-signal reasoning by 1.6 and 2.0 points, respectively. Visual representations reduce performance to 36.6% (`image-grid`) and 34.1% (`image-grouped`), while `csv+chart` offers no benefit over CSV alone. In contrast, Python access substantially improves overall accuracy from 51.2% to 71.3% (+20.1 points). Allowing the agent to additionally select among representations provides no further gain, reaching 69.7%.

Effect of time-series and history context. Removing only history reduces history-dependent types (*e.g.*, anomaly for baseline calculation) accuracy from 75.9% to 39.8%, while leaving window-scoped accuracy nearly unchanged (58.1% to 58.0%). When all wearable time series are withheld, window-scoped accuracy falls to 13.7%, and overall falls from 60.2% to 17.3%. This selective degradation confirms that `WearableQA` requires models to ground their answers in the relevant wearable measurements.

Table 4: **Ablation of inputs.** Results are reported for `Claude-Opus-4.6`.

Time series	History	Hist.-dep.	Window-dep.	Acc.
✓	✓	75.9 ±5.7	58.1 ±2.1	60.2 ±1.5
✓	✗	39.8 ±6.5	58.0 ±2.1	60.2 ±1.5
✗	✗	42.6 ±6.6	13.7 ±1.5	17.3 ±1.2

5 Conclusion

We introduced `WearableQA`, a benchmark of 4,084 multiple-choice questions grounded in the longitudinal wearable records, blood biomarkers, and demographics of 200 real users. The `WearableQA` covers 16 question types, with a 2×2 taxonomy, enabling a fine-grained diagnosis of the model. Overall, we hope `WearableQA` provides a realistic and diagnostic testbed for advancing LLMs that can reason reliably over longitudinal real-world user health data.

AI USE STATEMENT

Generative AI tools were used as research and writing assistants during the preparation of this work, including for language editing and code assistance. All scientific claims, methodological decisions, experimental designs, analyses, and final manuscript contents were reviewed and verified by the authors. The authors remain fully responsible for the accuracy, originality, and integrity of the submitted work.

ETHICS STATEMENT

This work introduces a benchmark for LLM reasoning over longitudinal wearable and health-related data from real users. The data used in this work consist of de-identified longitudinal wearable measurements, blood biomarkers, and demographic attributes collected from real users. All direct personal identifiers were removed prior to use in this study. The data were used and released in accordance with the applicable participant consent, data-use agreements, and research approval procedures, and were cleared by our institution's privacy review. No personally identifiable information is included in the benchmark. We acknowledge that a long per-user trajectory combined with a blood panel remains a quasi-identifier even without direct identifiers, and therefore release the data under CC BY-NC 4.0 with additional terms that prohibit re-identification or attempted re-identification.

REPRODUCIBILITY STATEMENT

We release the benchmark data and source code to support reproducibility. Full implementation details, evaluation settings, prompt templates, and model configurations are provided in the Appendix. We also document the benchmark construction and evaluation procedures in sufficient detail to enable independent reproduction of our results.

REFERENCES

Anthropic. System Card: Claude Opus 4.6, February 2026. URL https://www.anthropic.com/claude-opus-4-6-system-card.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. Timeseriesexam: A time series understanding exam. *arXiv preprint arXiv:2410.14752*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Gemini 3.1 Pro Model Card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Malgorzata Gwiazda, Yifu Cai, Mononito Goswami, Arjun Choudhry, and Artur Dubrawski. Timeseriesexamagent: Creating time series reasoning benchmarks at scale. In *ICLR*, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*, 2019.

Baoyu Jing, Sanhorn Chen, Lecheng Zheng, Boyu Liu, Zihao Li, Jiaru Zou, Tianxin Wei, Zhining Liu, Zhichen Zeng, Ruizhong Qiu, et al. Tsaqa: Time series analysis question and answering benchmark. *arXiv preprint arXiv:2601.23204*, 2026.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 2025.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

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 *ACL*, 2025.

Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. Hearts: Benchmarking llm reasoning on health time series. In *ICML*, 2026.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. Sensorllm: Aligning large language models with motion sensors for human activity recognition. In *EMNLP*, 2025.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *EMNLP Findings*, 2024.

Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 2026.

Girish Narayanswamy, Maxwell A Xu, A Ali Heydari, Samy Abdel-Ghaffar, Marius Guerard, Kara Vaillancourt, Zhihan Zhang, Jake Garrison, Levi Albuquerque, Dimitris Spathis, et al. Towards a general intelligence and interface for wearable health data. *arXiv preprint arXiv:2605.22759*, 2026.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. In *NeurIPS*, 2023.

OpenAI. GPT-4o System Card, August 2024. URL `https://openai.com/index/gpt-4o-system-card/`.

OpenAI. GPT-5.4 Thinking System Card, March 2026. URL `https://openai.com/index/gpt-5-4-thinking-system-card/`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *CHIL*, 2022.

Arvind Pillai, Dimitris Spathis, Subigya Nepal, Amanda C Collins, Daniel M Mackin, Michael V Heinz, Tess Z Griffin, Nicholas C Jacobson, and Andrew Campbell. Time2lang: Bridging time-series foundation models and large language models for health sensing beyond prompting. In *CHIL*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *ICML*, 2025.

A Appendix

A.1 Evaluation Protocol and Prompt

Every item is scored by *exact match*: the model is shown one user's demographics, wearable time series, blood panel, and a cohort reference distribution, and must select one of ten options (A–J). The predicted letter is parsed from the response and compared to the deterministic ground-truth key. All models are queried at temperature 0. Each model is evaluated in its default reasoning configuration. The released history length is chosen so that every prompt fits within the context window of all evaluated models.

Prompt structure. Each prompt is assembled from a fixed system instruction followed by five context blocks (USER PROFILE, SENSOR DATA, BLOOD BIOMARKER PANEL, COHORT REFERENCE, QUESTION/OPTIONS), and finally an answer-format directive that differs between the chain-of-thought(CoT) and direct-answer conditions.

System instruction (identical in both conditions).

```
You are given a user's demographics, wearable health
sensor history, blood biomarker panel, a cohort reference
distribution, and a multiple-choice question.  Analyze the
data carefully and select the single best answer.
```

Answer-format directive. The two evaluation conditions differ *only* in the trailing directive appended after the options; the context blocks are byte-identical, so any accuracy difference is attributable to reasoning elicitation alone.

Table 5: The only prompt difference between conditions is the closing directive.

Condition	Trailing directive (verbatim)

With CoT	`Think step by step.`
	`Show your reasoning, then on the LAST line write ONLY:`
	`Answer:  X`
	`where X is the single letter of your answer.`

Without CoT	`Choose the single best answer.  Respond with ONLY the letter (A-J). Do not explain or output anything else.` `Answer:`

Answer parsing. The response is parsed for `Answer: X` (case-insensitive, optional parentheses); failing that, a final lone letter line, then a leading letter, are accepted as fallbacks. Under the no-CoT (direct-answer) the model emits the bare letter directly; under CoT the parser reads the letter off the final line after the reasoning trace.

A.2 Experimental Details

Models used in construction. Both agents in Fig. 4 (Agent 1 and Agent 2) are run with `Claude-Opus-4.6`, which proposes computation plans; ground-truth labels are always obtained by executing the resulting deterministic programs. Question validation uses three models from different families: `Gemini-3.1-Pro` (Google DeepMind, 2026), `Claude-Opus-4.6` (Anthropic, 2026), and `GPT-5.4` (OpenAI, 2026). Items were filtered on quality only, not on whether any model answered them correctly.

Example of details of primitives in Discover-then-label. We illustrate how primitives are used in the discover-then-label path. A composition of primitives is evaluated on the released window, and its output selects the ground-truth option directly. We take `excursion_count` as the example, over the base window of 28 days. The question asks how often a signal left its own established range, and the composition uses three primitives: a split of the window, a robust baseline estimated on the earlier half, and a counting step over the later half.

```
first_half, second_half = window[:14], window[14:] # (1) split

med = median(first_half) # (2) robust baseline
sigma = 1.4826 * median(|x - med| for x in first_half)
lo, up = med - 1.5*sigma, med + 1.5*sigma

count = |{x in second_half : x < lo or x > up}| # (3) count
answer = bucket(count) # 0 | 1-2 | 3-4 | a large share | almost every day
```

We estimate the baseline with a median and a normalised median absolute deviation rather than a mean and a standard deviation, because a single outlier in the first half inflates an SD-based range and silently undercounts the departures that follow. The constant 1.4826 rescales the MAD so that it estimates σ for normally distributed data, which makes the "1.5 deviations" threshold comparable across metrics with different units. Because the reference range is taken from the first half and the count from the second, the question is answerable from the released window alone, and the stem states the definition in full so that the quantity to be computed is unambiguous.

A.3 Further Experiments

Positional bias. Ground-truth answers of `WearableQA` are balanced across the 10 option positions; an unbiased model should select each about 10% of the time. As shown in Fig. 10, proprietary models are relatively uniform, whereas smaller models exhibit strong positional preferences, especially under direct answering. The most extreme case is `Llama-3.2-3B`, which selects a single option for 51.5% of its direct-answer responses. Milder preferences also appear in proprietary models: `Claude-Opus-4.6` favors earlier options, particularly 'A', whereas `Gemini-2.5-Pro` favors later ones like 'J'. Chain-of-thought substantially reduces these biases, *e.g.*, `Llama-3.2-3B` reduces its maximum option-selection frequency from 51.5% to 13.5%, while `Claude-Opus-4.6` reduces it from 14.5% to 10.6%. This suggests that explicit reasoning helps models overcome positional answer-selection shortcuts, reinforcing that successful performance on `WearableQA` depends on reasoning over the provided evidence rather than exploiting superficial option preferences.

A.4 Time-Series Representations

The same wearable history can be handed to the model in several textual serializations or as a rendered plot. We evaluate four text formats (`row`, `col`, `markdown`, `csv`) produced by a single renderer, plus an image condition. We additionally evaluate two hybrid conditions: `markdown+stats`, which appends summary statistics to the table, and `csv+chart`, which pairs the best-performing text format with a rendered plot. All formats carry identical numeric content and identical column labels (`Steps`, `RHR`, `Sleep(h)`, `HRV`, etc.); they differ only in *layout*, which lets us isolate the effect of representation from the effect of information. The examples below show a 3-day, 4-signal excerpt, except for image representations, where we adopt full history with the window to visualize.

(a) `row` — one line per day (default). Each day is a record; signals are inline `key=value` pairs. Emphasizes same-day cross-signal reading.

```
=== SENSOR DATA (row: one line per day) ===
  2023-04-03: Steps=6120.0, RHR=61.0, Sleep(h)=6.80, HRV=42.0
  2023-04-04: Steps=7010.0, RHR=60.0, Sleep(h)=7.10, HRV=45.0
  2023-04-05: Steps=5480.0, RHR=63.0, Sleep(h)=6.40, HRV=39.0
```

(b) `col` — one block per metric. Each signal is a contiguous dated series. Emphasizes within-signal temporal trend/trajectory reading.

```
=== SENSOR DATA (column: one block per metric) ===
  Steps: 2023-04-03=6120.0, 2023-04-04=7010.0, 2023-04-05=5480.0
  RHR: 2023-04-03=61.0, 2023-04-04=60.0, 2023-04-05=63.0
  Sleep(h): 2023-04-03=6.80, 2023-04-04=7.10, 2023-04-05=6.40
  HRV: 2023-04-03=42.0, 2023-04-04=45.0, 2023-04-05=39.0
```

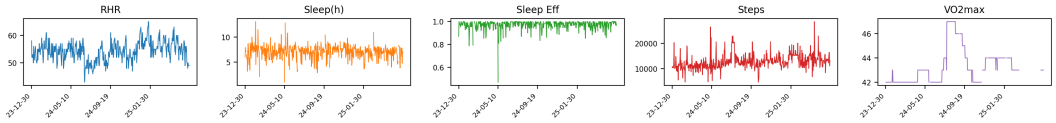
(c) `markdown` — pipe table. Days × signals grid with a header rule.

```
=== SENSOR DATA (markdown table) ===
| date | Steps | RHR | Sleep(h) | HRV |
|---|---|---|---|---|
| 2023-04-03 | 6120.0 | 61.0 | 6.80 | 42.0 |
| 2023-04-04 | 7010.0 | 60.0 | 7.10 | 45.0 |
| 2023-04-05 | 5480.0 | 63.0 | 6.40 | 39.0 |
```

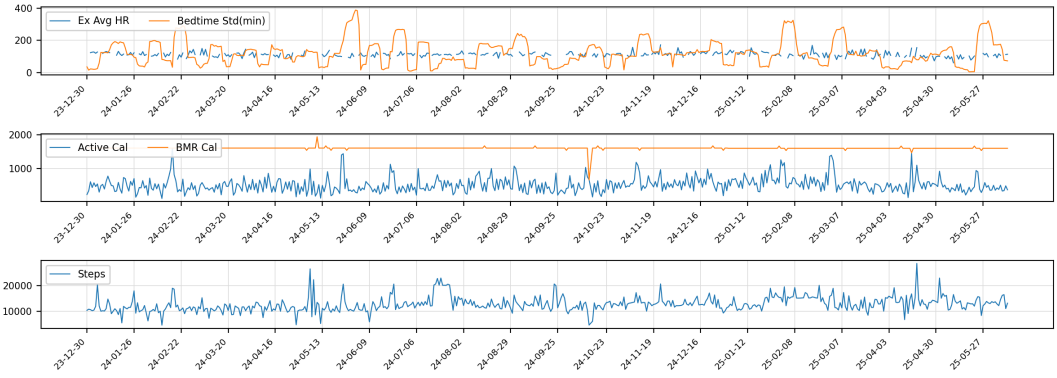
(d) `csv` — comma-separated grid. The same grid as a header row followed by comma-separated value rows.

```
=== SENSOR DATA (csv) ===
date,Steps,RHR,Sleep(h),HRV
2023-04-03,6120.0,61.0,6.80,42.0
2023-04-04,7010.0,60.0,7.10,45.0
2023-04-05,5480.0,63.0,6.40,39.0
```

(e) `image-grid` — separate plots for each signal (5 x N).



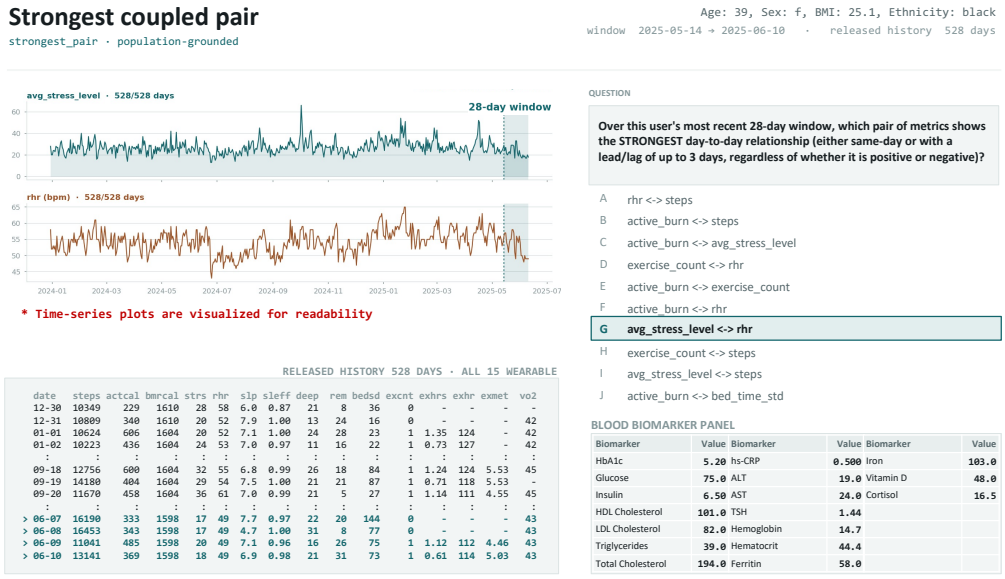
(f) `image-grouped` — range-grouped plots.



Example of evaluation prompt (row format, with CoT)

```
You are given a user's demographics, wearable health sensor history, blood biomarker
panel, a cohort reference distribution, and a multiple-choice question.  Analyze the data
carefully and select the single best answer.


=== USER PROFILE ===
Age:  41, Sex:  F, BMI: 24.8, Ethnicity:  White


=== SENSOR DATA (row:  one line per day) ===
2023-04-03:  Steps=6120.0, RHR=61.0, Sleep(h)=6.80, HRV=42.0, ...
2023-04-04:  Steps=7010.0, RHR=60.0, Sleep(h)=7.10, HRV=45.0, ...


=== BLOOD BIOMARKER PANEL ===
hdl:  58.0, ldl:  121.0, tg:  96.0, glu:  91.0, hba1c:  5.30, crp:  1.20


=== COHORT REFERENCE ===
steps:  p25=5400, p50=7600, p75=10200


=== QUESTION ===
Based on the most recent 28 days of data, how should this user's step volume be framed in
a metabolic-syndrome health summary?


=== OPTIONS ===
A. ...  B. ...  ...  J. ...

Think step by step.
Show your reasoning, then on the LAST line write ONLY:
Answer:  X
where X is the single letter of your answer.
```

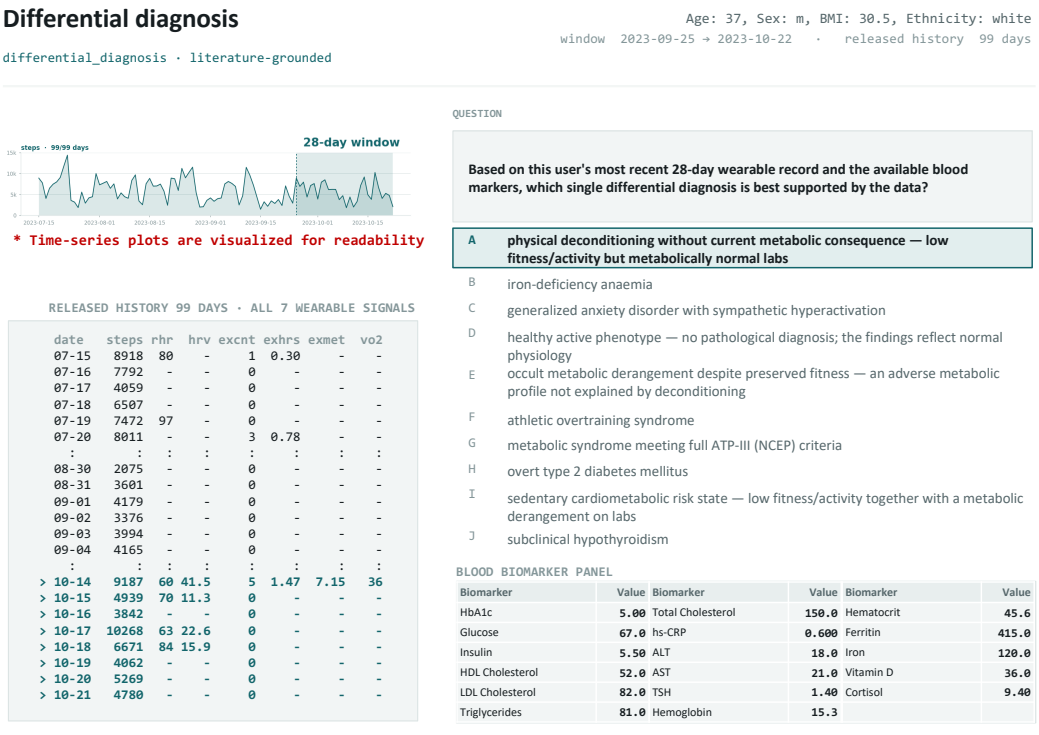
Figure 7: The assembled evaluation prompt. Under the no-CoT (Direct Answer) condition, the final instruction is replaced by "`Choose the single best answer. Respond with ONLY the letter (A--J). Do not explain or output anything else. Answer:`".



date	steps	actcal	bmrcal	strs	rhr	slp	sleff	deep	rem	bedsd	excnt	exhrs	exhr	exmet	vo2
12-30	10349	229	1610	28	58	6.0	0.87	21	8	36	0	-	-	-	-
12-31	10809	340	1610	20	52	7.9	1.00	13	24	16	0	-	-	-	42
01-01	10624	606	1604	20	52	7.1	1.00	24	28	23	1	1.35	124	-	42
01-02	10223	436	1604	24	53	7.0	0.97	11	16	22	1	0.73	127	-	42
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
09-18	12756	600	1604	32	55	6.8	0.99	26	18	84	1	1.24	124	5.53	45
09-19	14180	404	1604	29	54	7.5	1.00	21	21	87	1	0.71	118	5.53	-
09-20	11670	458	1604	36	61	7.0	0.99	21	5	27	1	1.14	111	4.55	45
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
> 06-07	16190	333	1598	17	49	7.7	0.97	22	20	144	0	-	-	-	43
> 06-08	16453	343	1598	17	49	4.7	1.00	31	8	77	0	-	-	-	43
> 06-09	11041	485	1598	20	49	7.1	0.96	16	26	75	1	1.12	112	4.46	43
> 06-10	13141	369	1598	18	49	6.9	0.98	21	31	73	1	0.61	114	5.03	43

Biomarker	Value	Biomarker	Value	Biomarker	Value
HbA1c	5.20	hs-CRP	0.500	Iron	103.0
Glucose	75.0	ALT	19.0	Vitamin D	48.0
Insulin	6.50	AST	24.0	Cortisol	16.5
HDL Cholesterol	101.0	TSH	1.44		
LDL Cholesterol	82.0	Hemoglobin	14.7		
Triglycerides	39.0	Hematocrit	44.4		
Total Cholesterol	194.0	Ferritin	58.0		

Figure 8: **Example question from `WearableQA` for data reasoning.** For readability, the wearable trajectory is visualized as a plot in this figure; models receive the underlying measurements in serialized textual form.



date	steps	rhr	hrv	excnt	exhrs	exmet	vo2
07-15	8918	80	-	1	0.30	-	-
07-16	7792	-	-	0	-	-	-
07-17	4059	-	-	0	-	-	-
07-18	6507	-	-	0	-	-	-
07-19	7472	97	-	0	-	-	-
07-20	8011	-	-	3	0.78	-	-
:	:	:	:	:	:	:	:
08-30	2075	-	-	0	-	-	-
08-31	3601	-	-	0	-	-	-
09-01	4179	-	-	0	-	-	-
09-02	3376	-	-	0	-	-	-
09-03	3994	-	-	0	-	-	-
09-04	4165	-	-	0	-	-	-
:	:	:	:	:	:	:	:
> 10-14	9187	60	41.5	5	1.47	7.15	36
> 10-15	4939	70	11.3	0	-	-	-
> 10-16	3842	-	-	0	-	-	-
> 10-17	10268	63	22.6	0	-	-	-
> 10-18	6671	84	15.9	0	-	-	-
> 10-19	4062	-	-	0	-	-	-
> 10-20	5269	-	-	0	-	-	-
> 10-21	4780	-	-	0	-	-	-

Biomarker	Value	Biomarker	Value	Biomarker	Value
HbA1c	5.00	Total Cholesterol	150.0	Hematocrit	45.6
Glucose	67.0	hs-CRP	0.600	Ferritin	415.0
Insulin	5.50	ALT	18.0	Iron	120.0
HDL Cholesterol	52.0	AST	21.0	Vitamin D	36.0
LDL Cholesterol	82.0	TSH	1.40	Cortisol	9.40
Triglycerides	81.0	Hemoglobin	15.3		

Figure 9: **Example question from `WearableQA` for health reasoning.** For readability, the wearable trajectory is visualized as a plot in this figure; models receive the underlying measurements in serialized textual form.

Table 6: **Per-question-type performance on `WearableQA`.** Accuracy (%) across the 16 question types under the default chain-of-thought prompting protocol. Best and second-best results for each question type are shown in **bold** and underline, respectively.

	Data Reasoning							
Model	**Signed Corr.**	**Signal Summary**	**Strongest Pair**	**Strongest Triplet**	**Trend Shape**	**Anomaly Episodes**	**Excursion Count**	**Recovery Time**
Proprietary Models								
GPT-4o	23.8±3.0	46.4±10.7	6.9±2.9	21.8±7.7	28.9±5.7	31.9±6.2	30.0±3.7	26.6±4.1
GPT-5.4	32.9±3.3	77.4±8.9	15.3±4.2	32.7±8.8	41.0±6.2	58.8±6.6	74.8±3.5	37.0±4.5
Claude Opus 4.6	54.6±3.5	85.8±7.3	32.8±5.6	39.1±9.1	**62.8±6.1**	75.9±5.7	76.3±3.4	36.8±4.5
Claude Sonnet 4	37.5±3.4	77.4±8.9	17.9±4.5	34.5±8.9	44.3±6.3	65.8±6.3	69.7±3.7	**41.5±4.5**
Gemini 2.5 Pro	38.2±3.4	82.2±8.0	26.6±5.2	33.6±8.8	47.3±6.3	65.8±6.3	70.9±3.6	33.5±4.4
Gemini 3.1 Pro	**86.9±2.4**	**88.2±6.7**	**68.3±5.5**	**69.1±8.6**	61.5±6.2	**82.4±5.0**	**94.4±1.8**	35.9±4.4
Open-source Models								
Llama-3.2-3B	11.3±2.2	8.3±5.6	11.7±3.8	19.0±7.3	15.0±4.5	13.4±4.5	15.1±2.8	18.4±3.6
Llama-3.1-8B	14.5±2.5	19.0±8.3	10.9±3.7	23.6±7.9	28.4±5.7	8.8±3.7	17.4±3.0	24.8±4.0
Llama-3.3-70B	21.6±2.9	27.3±9.5	9.1±3.4	23.6±7.9	28.4±5.7	38.4±6.5	21.4±3.3	29.5±4.2
Gemma-3-4B	9.1±2.0	8.3±5.6	13.8±4.1	23.6±7.9	16.3±4.7	18.0±5.1	14.9±2.8	25.5±4.0
Gemma-3-12B	13.9±2.4	15.4±7.6	15.3±4.2	30.9±8.6	11.7±4.0	15.7±4.8	18.4±3.1	23.9±3.9
Gemma-3-27B	10.7±2.2	34.5±10.1	8.0±3.2	30.0±8.5	25.9±5.5	27.3±5.9	21.0±3.2	26.4±4.1
Gemma-4-26B-A4B	30.1±3.2	78.6±8.7	14.6±4.1	40.0±9.2	39.3±6.2	48.6±6.7	37.7±3.9	28.2±4.1
Mistral-Small-3.1	14.4±2.5	38.1±10.4	8.0±3.2	22.7±7.8	20.9±5.1	28.7±6.0	22.3±3.3	27.5±4.1

	Health Reasoning							
Model	**Cross-Signal Pred.**	**Risk Assess.**	**Signal Concord.**	**Multimodal Phenotype**	**Fitness Pred.**	**Differential Diag.**	**Prognostic Pred.**	**Health Rec.**
Proprietary Models								
GPT-4o	12.9±3.6	69.7±8.5	54.9±9.7	65.9±10.4	88.3±5.6	45.4±11.1	70.6±4.2	46.1±8.6
GPT-5.4	25.8±4.7	86.7±6.2	77.5±8.0	87.4±7.1	95.0±3.6	80.6±8.7	68.6±4.3	53.1±8.6
Claude Opus 4.6	**40.5±5.3**	81.3±7.1	78.5±7.9	**88.7±6.7**	78.2±7.4	63.7±10.7	75.1±4.0	51.6±8.7
Claude Sonnet 4	38.6±5.3	85.8±6.4	69.6±8.9	77.3±9.2	84.1±6.5	59.8±11.0	73.1±4.1	58.6±8.5
Gemini 2.5 Pro	30.4±5.0	77.7±7.7	74.5±8.4	79.8±8.7	84.1±6.5	52.0±11.2	59.2±4.5	45.3±8.6
Gemini 3.1 Pro	26.1±4.8	**87.6±6.0**	**85.3±6.7**	86.1±7.4	**95.0±3.6**	**80.6±8.7**	**77.1±3.9**	**61.7±8.4**
Open-source Models								
Llama-3.2-3B	12.3±3.5	33.9±8.8	27.4±8.6	26.5±9.7	44.5±8.9	32.4±10.4	42.3±4.6	40.6±8.5
Llama-3.1-8B	8.3±2.9	43.7±9.2	39.2±9.5	39.2±10.8	56.3±8.9	37.6±10.8	49.2±4.6	41.4±8.5
Llama-3.3-70B	8.6±3.0	53.6±9.2	35.3±9.3	54.4±11.0	80.7±7.0	42.8±11.1	47.0±4.6	49.2±8.7
Gemma-3-4B	15.3±3.9	41.1±9.1	22.5±8.0	31.6±10.2	49.6±9.0	22.0±9.2	32.3±4.3	39.8±8.5
Gemma-3-12B	14.7±3.8	58.0±9.1	35.3±9.3	54.4±11.0	70.6±8.2	22.0±9.2	47.0±4.6	42.2±8.6
Gemma-3-27B	15.3±3.9	64.3±8.9	41.2±9.6	55.7±11.0	71.5±8.1	28.5±10.0	44.8±4.6	42.2±8.6
Gemma-4-26B-A4B	22.7±4.5	73.2±8.2	54.9±9.7	77.3±9.2	91.7±4.8	68.9±10.3	71.7±4.2	56.3±8.6
Mistral-Small-3.1	14.1±3.8	70.6±8.4	36.3±9.3	67.1±10.3	76.5±7.6	65.0±10.6	56.6±4.6	44.5±8.6

Table 7: **Comparison of chain-of-thought (CoT) and direct-answer (DA) prompting on `WearableQA`.** Accuracy (%) is reported for each model, with ± giving a 95% confidence half-width. Gap denotes CoT − DA in percentage points, where positive values indicate an improvement from CoT prompting; as the two conditions are evaluated on the same questions, the gap is a paired difference.

Model	**CoT**	**DA**	**Gap**
Gemini-3.1-Pro	72.9 ±1.4	71.3 ±1.4	+1.6
Claude-Opus-4.6	60.2 ±1.5	40.6 ±1.5	+19.6
Claude-Sonnet-4	53.2 ±1.5	48.3 ±1.5	+4.9
GPT-5.4	51.2 ±1.5	33.5 ±1.4	+17.7
Gemini-2.5-Pro	50.6 ±1.5	32.0 ±1.4	+18.6
Gemma-4-26B-A4B	42.5 ±1.5	35.7 ±1.5	+6.8
GPT-4o	34.7 ±1.5	23.9 ±1.3	+10.8
Llama-3.3-70B	29.2 ±1.4	27.2 ±1.4	+2.0
Mistral-Small-3.1	29.3 ±1.4	24.5 ±1.3	+4.8
Gemma-3-27B	26.2 ±1.3	23.8 ±1.3	+2.4
Gemma-3-12B	24.6 ±1.3	21.7 ±1.3	+2.9
Llama-3.1-8B	23.8 ±1.3	22.3 ±1.3	+1.5
Gemma-3-4B	19.8 ±1.2	20.6 ±1.2	−0.8
Llama-3.2-3B	19.6 ±1.2	17.9 ±1.2	+1.7

Table 8: **Accuracy (%) by reasoning group *and* signal complexity.** Each cell aggregates the question types belonging to it, weighted by the number of questions; ± is a 95% confidence half-width. Bold and underline mark the best and second-best within each model group; cell shading gives the rank across all models in that column. Chance accuracy is 10%. .

	Data reasoning		Health reasoning	
Model	**Single-signal**	**Cross-signal**	**Single-signal**	**Cross-signal**
Proprietary models				
Gemini 3.1 Pro	**70.9** ±2.2	**80.9** ±2.3	**87.6** ±6.2	**65.5** ±2.6
Claude Opus 4.6	63.5 ±2.4	48.1 ±2.9	81.3 ±7.4	64.6 ±2.6
Claude Sonnet 4	57.8 ±2.4	32.6 ±2.7	85.8 ±6.6	63.0 ±2.7
GPT-5.4	57.0 ±2.4	28.7 ±2.6	86.7 ±6.4	61.0 ±2.7
Gemini 2.5 Pro	56.6 ±2.4	35.1 ±2.8	77.7 ±7.9	54.7 ±2.8
GPT-4o	30.0 ±2.3	19.6 ±2.3	69.7 ±8.7	52.0 ±2.8
Open-source models				
Gemma-4-26B-A4B	**38.8** ±2.4	**27.4** ±2.6	**73.2** ±8.4	**58.3** ±2.7
Mistral-Small-3.1	25.2 ±2.1	13.7 ±2.0	70.6 ±8.6	45.8 ±2.8
Llama-3.3-70B	27.3 ±2.2	18.8 ±2.3	53.6 ±9.4	39.7 ±2.7
Gemma-3-27B	24.8 ±2.1	11.8 ±1.9	64.3 ±9.0	38.8 ±2.7
Gemma-3-12B	18.4 ±1.9	15.8 ±2.1	58.0 ±9.3	38.5 ±2.7
Llama-3.1-8B	20.1 ±2.0	14.5 ±2.0	43.7 ±9.4	36.5 ±2.7
Gemma-3-4B	18.2 ±1.9	11.5 ±1.8	41.1 ±9.3	28.9 ±2.5
Llama-3.2-3B	15.4 ±1.8	12.1 ±1.9	33.9 ±8.9	31.9 ±2.6

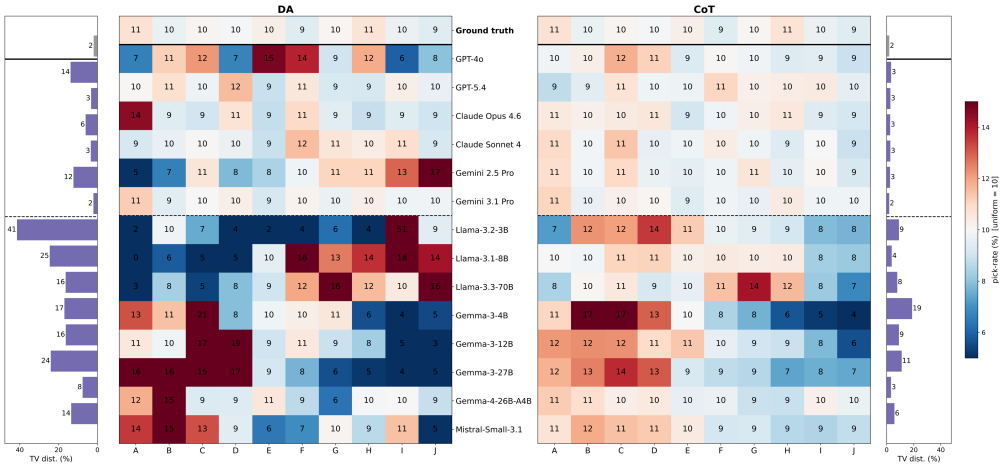


Figure 10: **Answer-letter pick-rate vs. uniform ground truth (DA left, CoT right).** Cells show each model's pick-rate (%) per letter (red = over-picked, blue = under-picked, white = 10); side panels give total-variation distance from uniform. Solid rule = ground-truth row, dashed rule = proprietary vs. open-sourced. Bias grows from CoT to DA and from larger to smaller models; `Llama-3.2-3B` under DA collapses onto one letter (51%), while CoT restores near-uniformity.

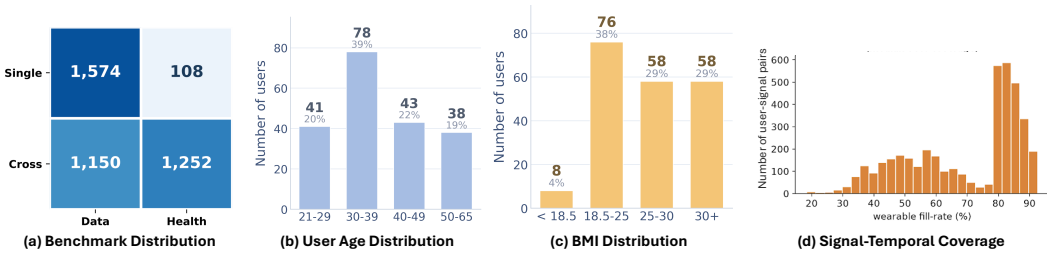


Figure 11: **Overall statistics of 200 real users from `WearableQA`.**