

HiMe: Real-Time Self-Hosted Personal Agent Platform for Health Insights with Wearable Devices

Wei Liu^{*,*}, Siya Qi^{*,*}, Linhai Zhang^{*}, Lorraine Tudor Car^{*}, Yulan He^{*,*✉},

^{*}King's College London, [✉]The Alan Turing Institute,
{wei.4.liu, yulan.he}@kcl.ac.uk

Abstract

Traditional approaches to wearable health signal analysis, such as smartwatches, are constrained by rigid analytical frameworks and limited personalisation. The emergence of LLM agents creates a new opportunity for Personal Health Agentic Analysis, where health insights can be generated adaptively and in context. However, currently there is no open-source locally deployable platform capable of processing personal health data in real time while preserving privacy. We present HiME, a locally deployable, privacy-first agent platform that is fully compatible with real-time health data ecosystems across a wide range of wearable devices. HiME is guided by three design principles. The database is treated as a first-class component. Effectiveness and efficiency are jointly optimised to achieve a low-cost Pareto-optimal balance. Data are processed in real time while the user is modelled over the long term. Together, these principles make it practical for individuals to harness Personal Health Agents for continuous, personalised health monitoring for better wellbeing¹.

1 Introduction

Wearable devices have moved health monitoring out of the clinic and into everyday life. Modern smartwatches and smartphones continuously record signals such as heart rate, blood oxygen, sleep, and workouts into private on-device stores. Over time, these signals form a longitudinal record of an individual's health. But making full use of this stream is hard. Conventional wearable analytics rely on fixed statistics that often ignore both signal dynamics and user preferences, while the long-term raw trace is too large to fit within the context window of a large language model (LLM) directly for intelligent analysis.

^{*}Equal contribution.

¹Available at <https://github.com/thinkwee/HiMe>

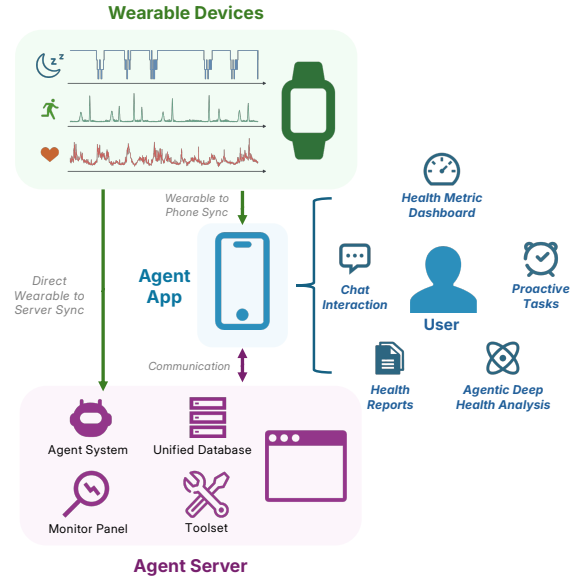


Figure 1: Overview of the HiME system.

LLM agents that reason and act through tools offer a promising alternative (Yao et al., 2023). Such agents can access health data through tools, decide what is worth analysing, and explain their findings in natural language. This promise has mainly been pursued in two settings (Lin et al., 2026), neither of which fully addresses individuals living with their own continuous wearable health data. Clinical multi-agent systems reason over encounter-based records for clinician-facing use cases (Tang et al., 2024; Kim et al., 2024; Schmidgall et al., 2026), making the clinician, not the patient, the primary user. Personal coaching systems such as GPTCoach (Jörke et al., 2025) and MindScape (Nepal et al., 2024) are built around guided dialogue or journal prompts rather than a continuous data stream. General-purpose self-hosted assistants such as OpenClaw (Steinberger, 2026) have also popularised a local-first personal agent reached through messaging channels, but are built for chat and scheduled tasks, not a high-rate

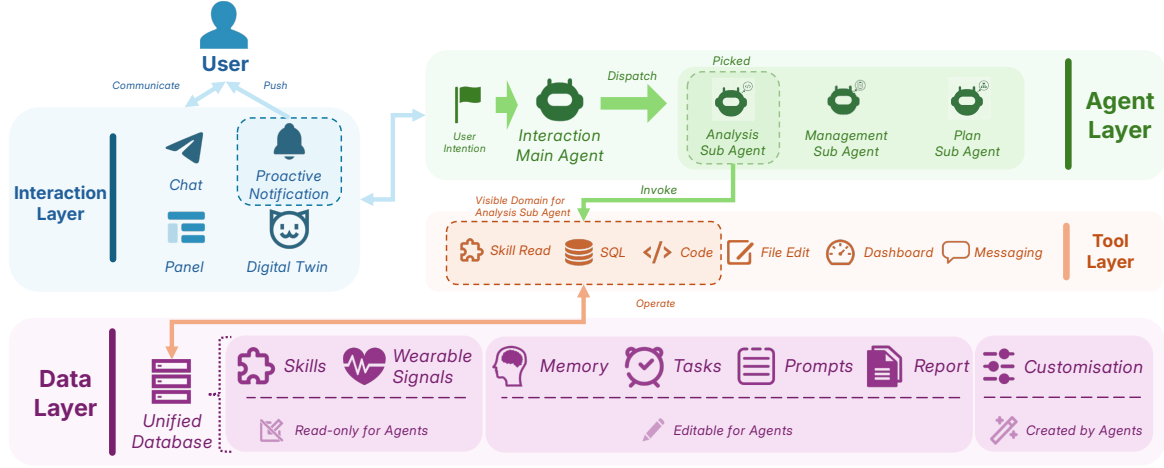


Figure 2: The four layers of the HiME framework.

physiological stream or months-long models of individuals. There is still no open-source platform that lets individuals run, on their own hardware, an agent that ingests wearable data in real time and generates long-term personal health insights.

We present HiME, the **Health Intelligence Management Engine**.² This locally deployable, privacy-first agent platform connects to real-time health data ecosystems from a wide range of wearable devices. HiME is built on three design principles for personal health. **1)** The database is treated as a first-class component. This enables the agent to analyse wearable signals, personal preference and its own memory jointly, while the checkable terminal state of the database provides a direct target for evaluating LLM agents on HiME’s health insight generation tasks. **2)** Effectiveness and efficiency are optimised jointly, seeking a Pareto-optimal trade-off among quality, cost, and latency, allowing always-on agency to remain affordable on consumer hardware and local models. **3)** Real-time data processing is combined with long-term user modelling. This allows the system to capture both momentary anomalies and months-long trends, supporting timely intervention and longitudinal understanding. Each user is effectively a continuous, heterogeneous data stream. Such personal data are an asset fundamentally different from the public data used to train large models. HiME is built to harness LLM agents for this stream, not only to satisfy a momentary request but to help

each user become a healthier self over time. The modular design also supports AI researchers requiring environments for evaluating and training health agents, and medical research applying large models to real-world wearable data.

2 Architecture

2.1 Overall

HiME is a self-hosted platform consisting of an agent server and companion apps for the smartphone and wearable (Figure 1). A single agent sits between the user and a continuously growing store of personal data, as the user never queries the data directly, and the agent reads, analyses and writes it on the user’s behalf, turning a raw physiological stream into personalised insight, with the whole pipeline running on the user’s own hardware. A wearable collects signals such as heart rate, blood oxygen, heart-rate variability, sleep and workouts in real time, buffering readings into batches relayed via the phone to the agent server, where a data adapter normalises any wearable format and deduplicates each sample before storage. All data land in one unified per-user database, a signal store alongside agent-state tables for memory, the user profile, reports, scheduled tasks and monitoring rules, to which the agent may add tables of its own. Live data reaches that adapter along two paths, as Apple Watch and iPhone stream HealthKit samples through the bundled exporter, while an optional self-hosted open-wearables ([Momentum, 2025](#)) sidecar covers the further ecosystems it supports, such as Garmin, Oura, Whoop and Fitbit, behind a single API.

The agent server runs a single loop that inter-

²HiME is a wellness, self-tracking and research platform, not a medical device. It produces no diagnoses, and its LLM-generated interpretations must not be used for clinical decisions or relied on in an emergency; see §2.3 for the full safety and privacy boundary.

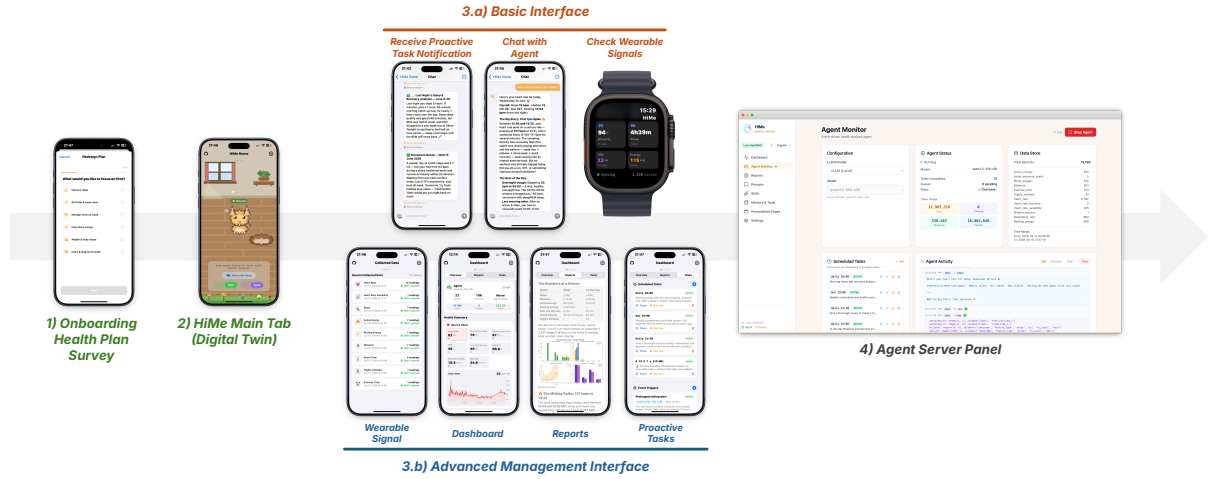


Figure 3: User-facing demonstration flow, from onboarding through a day of use across the in-app chat, wearable device status check, advanced management tabs, and agent server panel for comprehensive monitoring.

leaves user messages, real-time monitor triggers, and scheduled tasks. Each wakes the interaction agent, which routes to specialised sub-agents and acts through a small tool set, reading the signal store over several steps and writing the result back as a report, a memory update or a new task, before replying on the channel the request arrived from. Every reported number is checked against its query evidence. Acting unprompted, a cheap statistical trigger watches the stream at full resolution for free and wakes the expensive analysis only when it fires, pushing the resulting insight to the phone and watch — turning a continuous physiological stream into personalised, auditable insight without the data ever leaving the user’s hardware.

2.2 Layered Design

Figure 2 shows the four layers behind this pipeline.

Data layer. All information the agent holds lives in one unified database, with tables for wearable signals, memory and the user profile, and the agent may create custom tables for whatever else it needs, so heterogeneous data are analysed jointly. Asked for a health-check report, for example, the agent reads established baselines from its own tables and analysis skills from the skill table, queries the signal table over several steps, replies, writes the result to the report table, and may add a morning proactive task. Each agent holds its own table permissions, so every operation stays safe. No workflow is hard-coded. HiME supplies only the tool set and schema design principles rather than rigid constraints, which lets weaker LLMs emit well-formed outputs while leaving stronger ones unconstrained.

Tool layer. With the database as a first-class component, the tool layer stays simple. In principle, the agent only reads and writes operations over the unified database, and a code execution tool is needed, with scenario-specific variants for replying, editing files or generating dashboards sharing the same abstraction. The code tool keeps Jupyter-style state across calls (Kluyver et al., 2016), so the agent builds on earlier steps instead of rewriting from scratch. To curb tool-use hallucination, a fact verifier classifies every outbound message from its tool trace, checks each reported number against the matching query result, and stores a hashed evidence trail behind a show-evidence button.

Agent layer. The interaction agent is the primary, user-facing agent, and at any tool-use step it can invoke specialised sub-agents, including an analysis agent, a management agent for memory updates and proactive tasks, and a plan agent for personalised health plans. Prompts are assembled from modular layers (system instructions, role, soul, responsibilities and experience), maximising reuse and keeping the prefix stable for better key-value (KV) cache utilisation. The backend supports 15 LLM API providers, though we recommend serving open-source models locally through vLLM (Kwon et al., 2023) for privacy.

Interaction layer. Users reach the agent through the built-in chat or integrated messaging platforms such as WeChat and Telegram, and a gateway registry routes each reply back to the channel the message arrived on. As interactions accumulate, the agent keeps updating its memory and the user pro-

file and configures proactive tasks autonomously, all of which stay editable in the agent server panel, either manually or by instructing the agent. Proactive tasks cover time-based schedules and event-based triggers that the agent authors itself, letting it push notifications to the smartphone and watch. A pixel-art cat acts as a digital twin whose animations reflect the user’s current health status in real time. Figure 3 shows a session across these surfaces.³

From principles to design. These three principles, not the field’s defaults, explain why the system takes this shape. Treating the database as first-class makes the agent an interface over persistent state rather than a transient reasoning engine. The user model stays as agent-editable Markdown rather than an opaque vector store (Packer et al., 2024; Chhikara et al., 2025; Xu et al., 2025), every action leaves a checkable terminal state that automates evaluation, and a new capability becomes a new table rather than a new agent. Jointly optimising effectiveness and efficiency removes needless computation, as lightweight filters run before expensive inference, narrow roles carry small prompts routed to the cheapest adequate model, and a static prefix improves KV-cache reuse. Pairing real-time processing with long-term modelling lets cheap components watch the stream at full resolution while the LLM wakes only on meaningful triggers, capturing transient anomalies and long-horizon trends in one framework. §3 analyses these choices.

2.3 Safety, Privacy and Deployment Boundaries

HIME is a wellness and research platform, not a medical device, and does not provide diagnoses or treatment recommendations. Its analysis is limited to descriptive wellness summaries and may flag unusual observations for the user to discuss with a clinician, but it does not claim to determine medical conditions or provide emergency care. Monitoring rules only control when the agent is triggered and cannot expand what the agent is permitted to say. All monitoring and agent-authored rules are explicit, inspectable, editable and removable, with fixed conditions and cooldowns. Privacy depends on the deployment mode. In the recommended local mode, the model, database, agent loop, tools and notifications run on hardware controlled by the user, so health-derived content remains local.

³A screencast walkthrough accompanies the submission.

In hosted-API mode, health-derived content used by the agent, including retrieved memory, profile context and tool outputs, is sent to the configured model provider under its terms, while API calls are recorded in a local audit log. Messaging gateways are disabled by default, raw samples are retained only for a limited period, derived artefacts persist until deleted, and the system provides a command to erase stored data. The server is bound to localhost by default and requires explicit configuration and authentication to be exposed externally. HIME operates without a central server, user accounts or telemetry collection.

3 Evaluation

HIME is a demonstration system rather than a benchmark, yet the first principle makes it measurable. Because every capability leaves a terminal database state, most actions are scored by replaying a public wearable trace, letting the agent act and comparing the resulting state against a gold state computed independently from the same trace. We drive the deployed agent offline over ten randomly sampled participants from each of five public wearable corpora, namely LifeSnaps (Yfantidou et al., 2022), PMData (Thambawita et al., 2020), MMASH (Rossi et al., 2020), CovIdentify (Cho et al., 2024) and GLOBEM (Xu et al., 2022), each through its own data adapter. The sweep covers 22 backbones. Nineteen run locally and span the Qwen2.5 (Yang et al., 2024), Qwen3 (Yang et al., 2025), Qwen3.5 (Qwen Team, 2026a), Qwen3.6 (Qwen Team, 2026b) and Gemma-4 (Google DeepMind, 2026) families from 1.5B to 35B parameters, along with nemotron-3-nano, granite-4.1-8b and glm-4.7-flash. GPT-5-mini, DeepSeek-V4-flash and Gemini-3.5-flash are hosted-API frontier references. On-device execution is the right target for privacy and latency, so the question that matters is how far local small backbones remain from frontier models. We benchmarked latency and show the results in Appendix E. Those measurements are taken on one isolated H200 and are an upper bound rather than a consumer expectation. With a single request in flight, decoding is memory-bandwidth-bound, so against the roughly 4.8 TB/s of an H200 a high-end consumer GPU near 1 TB/s implies a slowdown around five times, an Apple silicon laptop with unified memory six to twelve times, and a CPU-only host one to two orders of magnitude, with four-bit quan-

tisation recovering part of it; these are analytic estimates from the bandwidth bound, not measurements on consumer hardware. For a consumer deployment we recommend Qwen3.5-9b, whose weights need about 18 GB at bf16 and roughly 6 GB at four bits, and latency matters less than the table suggests for most of what HiME does, since triggered analyses, scheduled check-ins and report generation all run in the background and reach the user as a notification, leaving only interactive chat latency-critical.

3.1 Per-role agentic capability

We give every agent in HiME the same backbone and score it against a terminal-state oracle that requires no LLM judge. Results are reported as pass@1 (Chen et al., 2021) over three runs. Each of the five role columns isolates one capability with its own closed-form check. **Route** scores top-level dispatch. The interaction agent must hand a user message to exactly the sub-agents it needs, with every required role present and none spurious. **Analysis** answers a natural-language question over the replayed trace. It succeeds when the reported number $\hat{v}$ matches the value v of an independent reference query, $|\hat{v} - v| \leq \max(0.02|v|, 0.5)$, and is backed by a real query or code run. **Memory** executes a write instruction. It succeeds when the resulting user profile and proactive task state match the request, with the protected file header intact and no collateral edits. **Plan** drives onboarding. It succeeds when the created scheduled tasks carry valid cron expressions, cover every stated health goal and ship with a published report. **Halluc.** captures the failure mode shared across these roles, the rate of asserting a number with no backing query, measured on the analysis tasks. Each role keeps its own oracle, so a failure can be attributed to the role that caused it, a diagnosis unavailable in a single-role agent design.

Per-role results. Table 1 reports results for 12 representative backbones. Three findings stand out. First, local deployment rivals hosted frontier models. The strongest local backbones reach an analysis score of 0.91 (Qwen3.6-27b and Qwen3.5-35b-a3b), matching or surpassing DeepSeek-V4-flash (0.85) and GPT-5-mini (0.76) with no hallucinations; only Gemini-3.5-flash clearly leads (0.94). Second, capability does not scale monotonically with model size. Qwen3-30b-a3b, despite ~ 3 B active of 30B total parameters, trails the smaller

dense Qwen3-8b on analysis (0.36 versus 0.47) and memory (0.33 versus 0.96). Analysis is the most discriminative capability, ranging 0.00–0.94, while routing saturates for the most capable backbones. Third, small models lag substantially. The 1.5B model fails across all capabilities with the highest hallucination rate (0.47), so current small on-device models remain insufficient for personal health agent tasks.

Composite rundown. The per-role scores measure each capability in isolation, which overstates what a user experiences. We therefore run a *composite rundown*, a five-turn persistent chat ordered as a realistic daily interaction, including a weekly check-in plan covering sleep and activity, recalling what was just set up, the average heart rate on the user’s most active day (requiring a grounded number), a durable preference written to memory, and a final turn that compounds the recall, switches to metric units, and asks for the all-time peak heart rate, so both the number and the preference must be correct. With $a_{i,t}$ and $r_{i,t}$ marking whether turn t of session i passes its task and routing checks, over N sessions we report MT (Mean per Turn) score and S5 (Strict all-five-turns pass) score,

$$\text{MT} = \frac{1}{5N} \sum_{i,t} a_{i,t},$$

$$\text{S5} = \frac{1}{N} \sum_i \prod_t r_{i,t} a_{i,t}.$$

We also report s/session, the median wall-clock of one session on a single isolated H200, the latency a user waits per rundown (Appendix E). Performance degrades over long interaction trajectories. Per-turn scores remain strong, with Gemini-3.5-flash at 0.98 and Qwen3.5-35b-a3b at 0.96, yet flawless sessions are rare because per-turn errors compound. The best S5 values are only 0.44 (Gemma-4-26b-a4b) and 0.37 (Gemini-3.5-flash), and most backbones sit at zero. The failure is more specific than uniform degradation. S5 requires the task *and* the routing check to pass on every turn, and for the strongest backbones it is routing consistency, not analysis, that binds: Gemini-3.5-flash answers 0.98 of turns correctly but routes only 0.83 of them, and $0.83^5 \approx 0.40$ almost exactly accounts for its 0.37 strict rate, while DeepSeek-V4-flash pairs a 0.84 per-turn score with 0.60 routing and reaches zero. For mid-tier backbones the loss instead concentrates in the final turns, the only ones that must carry earlier context: Qwen3-32b scores 1.00, 0.70

Backbone	Route	Analysis	Memory	Plan	Halluc.↓	MT	S5	tok/session	s/session↓
Gemini-3.5-flash [†]	0.98	0.94	0.89	0.98	0.01	0.98	0.37	42k	–
DeepSeek-V4-flash [†]	1.00	0.85	0.97	1.00	0.00	0.84	0.00	66k	–
GPT-5-mini [†]	0.88	0.76	0.99	0.88	0.00	0.63	0.00	206k	–
Qwen3.6-27b	1.00	0.91	1.00	0.99	0.00	0.76	0.20	253k	97.3
Qwen3.5-35b-a3b	1.00	0.91	0.98	0.84	0.00	0.96	0.26	169k	21.3
Qwen3-32b	1.00	0.68	0.81	0.80	0.00	0.69	0.00	178k	55.9
Gemma-4-26b-a4b	1.00	0.48	0.93	0.33	0.26	0.89	0.44	236k	32.8
Qwen3-8b	0.61	0.47	0.96	0.57	0.00	0.68	0.00	194k	45.6
Qwen3-4b	1.00	0.43	0.67	0.37	0.00	0.57	0.00	33k	12.3
Qwen3-1.7b	0.50	0.41	0.01	0.99	0.00	0.26	0.00	123k	10.8
Qwen3-30b-a3b	0.97	0.36	0.33	0.99	0.00	0.61	0.00	242k	34.7
Qwen2.5-1.5b	0.37	0.00	0.00	0.00	0.47	0.26	0.00	20k	4.0

Table 1: Per-role capability for 12 representative backbones, 3 hosted ([†]) and 9 local, auto-scored on replayed public traces as pass@1 over three runs. MT is the mean per-turn score and S5 the strict all-five-turns rate of the composite rundown, Halluc. the unbacked-number rate, tok/session the mean tokens per five-turn session, and s/session the median wall-clock seconds of that session on one isolated H200 (Appendix E), shown as – for the hosted APIs, which we do not self-serve. The full 22-backbone sweep is in Table A2.

and 0.86 on turns two to four but 0.13 on the compound fifth turn, and Qwen3-30b-a3b falls from 0.86 to 0.08 at the memory-write turn. The failures are also semantic rather than operational, as crashes and give-ups account for under 3% of all failures for ten of the thirteen audited backbones, the exceptions being Qwen2.5-1.5b (34%), Qwen3-30b-a3b (26%) and Qwen2.5-7b (11%). Reliably chaining competent roles across a session, and in particular holding the router stable while context accumulates, is therefore an open frontier for on-device agents.

Does decomposition help? Splitting the agent into roles is first an engineering choice, as each role owns a bounded slice of the task, easing management and attributing failures rather than tangling tool calls inside a single persona. The ablation shows this split is also nearly free on the metric (Table A3). Most backbones that can run the multi-role protocol gain or break even, led by +0.65 for Qwen3-8b and +0.23 for Qwen3-32b, with only Qwen3.6-27b (−0.36) and Gemma-4-e4b (−0.71) paying a substantial cost. In deployment, where a monolithic persona must hold every role in one long context, the advantage of factored responsibilities should only widen.

3.2 Data-to-report evaluation

When asked for a personalised health report, the agent must read a participant’s multivariate history

and commit to what it means. This cross-modal task, from time series to a natural-language verdict, probes understanding rather than retrieval (Liu et al., 2026). A fixed LLM judge (Liu et al., 2023) scores the wearable-only summary against the participant’s held-out self-report on a rubric covering fatigue, readiness and stress. Figure 4 shows the task is hard. The best backbone reaches only 46% (DeepSeek-V4-flash), and Gemini-3.5-flash leads single-turn analysis at 0.94 yet scores 14.6% here, while nemotron-3-nano at 0.85 analysis reaches 18.7%. Narrating subjective state is therefore a competence separate from retrieving numbers, the one a personal agent adds on top of question answering. It is also the reason HIME presents such narration as a descriptive wellness summary rather than an assessment, and constrains it accordingly (§2.3).

3.3 Real-time triggering analysis

An always-on health agent must act unprompted, monitoring a continuous wearable stream without paying for an LLM call at every reading. HIME lets the agent compile a user’s intent into a trigger, a cheap statistical condition that runs continuously for free and wakes the expensive multi-turn analysis only when it fires (§2). To test whether this gating preserves detection while cutting cost, we compare it against a polling baseline that runs the identical analysis on a fixed schedule, measuring how well each discovers injected anomalies in the

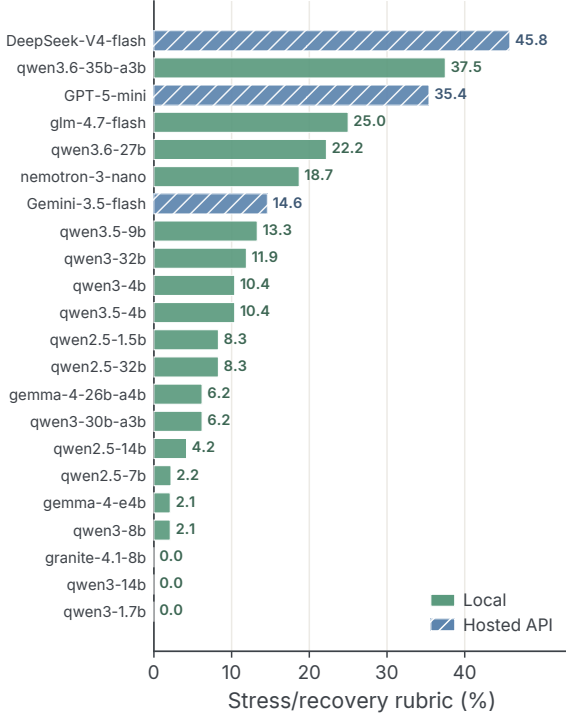


Figure 4: Cross-modal report understanding, the stress/recovery rubric score (0 to 100%) of a wearable-only summary judged against the participant’s held-out self-report. The breakdown is in Table A1.

Backbone	Polling recall	Trigger recall	Tokens saved
Qwen3-1.7b	0.00	0.29	121×
Qwen3-8b	0.69	0.89	76×
Gemma-4-26b-a4b	0.57	0.43	83×
Qwen3-30b-a3b	0.58	0.52	49×

Table 2: Detection on injected anomalies across PM-Data, LifeSnaps and MMASH, five participants per corpus and 47 events per backbone.

stream (Appendix D gives the full protocol). Table 2 reports conditional recall given trigger activation for four backbones. Agent-authored triggers identify more events on average than polling, and the gain is larger for smaller models, which a fixed schedule otherwise overwhelms with normal signals so that critical readings are missed. For larger models the trigger instead buys an order-of-magnitude cut in token usage at only a modest cost in recall.

3.4 Longitudinal Real-User Study

Whether HIME helps users is a question only real use can answer, and the aim is not momentary satisfaction but a healthier self over time. We conducted

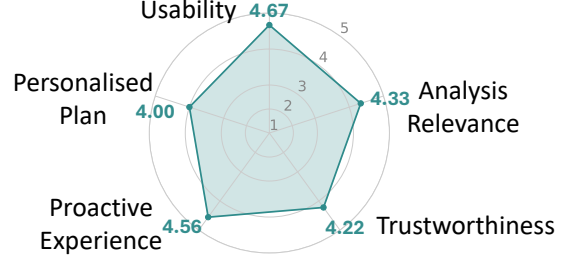


Figure 5: Mean participant ratings of HIME across five dimensions on a five-point Likert scale.

a two-month field study with nine non-technical participants using HIME on their own devices, collecting end-of-study feedback on a five-point Likert scale (Figure 5; survey instrument and further breakdown in Appendix F). Ratings were positive but uneven across dimensions (4.00–4.67). Usability (4.67) and proactive experience (4.56) scored highest, with participants reporting that proactive sleep and recovery reports helped them re-engage with previously overlooked health data. Personalised plan fit was the weakest dimension (4.00), and a few participants found scheduled plans slow to adapt to a changed routine. Compared with prior tools, participants credited HIME mainly with gains in proactivity and personalisation, and less often with gains in raw accuracy or speed (Appendix F). Reported failures were infrequent and mostly minor, so they did not outweigh the perceived benefit of continuous, evidence-backed monitoring. The study is evidence of usability and perceived usefulness only. With nine self-selected participants, no control condition and no clinical endpoint, it establishes neither health benefit nor sustained behaviour change, and we make no such claim.

4 Conclusion

We presented HIME, a self-hosted agent platform that turns continuous wearable streams into personal health insights, with three design principles making its architecture measurable, auditable and extensible. Across five corpora and 22 backbones, strong local models already rival hosted APIs, making privacy-preserving personal health agents increasingly practical, though multi-turn reliability and data-to-report narration remain open. We hope HIME supports research toward trustworthy personalised health agents.

References

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. [Mem0: Building production-ready AI agents with scalable long-term memory](#). In *European Conference on Artificial Intelligence (ECAI)*.
- Peter Cho, Md Mobashir Hasan Shandhi, Ali Roghanizad, and Jessilyn Dunn. 2024. [CovIdentify dataset](#). PhysioNet. Version 1.0.0.
- Google DeepMind. 2026. Gemma 4: Byte for byte, the most capable open models. <https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/>.
- Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. [GPTCoach: Towards LLM-based physical activity coaching](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. [MDAgents: An adaptive collaboration of LLMs for medical decision-making](#). In *Advances in Neural Information Processing Systems*.
- Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. 2016. [Jupyter notebooks – a publishing format for reproducible computational workflows](#). In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626. ACM.
- Georgianna Lin, Rencong Jiang, Noémie Elhadad, and Xuhai ‘Orson’ Xu. 2026. A framework for longitudinal health ai agents. *Nature Health*, pages 1–10.
- Wei Liu, Peijie Yu, Michele Orini, Yali Du, and Yulan He. 2026. [Hunt instead of wait: Evaluating deep data research on large language models](#). *arXiv preprint arXiv:2602.02039*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522. Association for Computational Linguistics.
- The Momentum. 2025. Open wearables: Self-hosted platform to unify wearable health data through one ai-ready api. <https://github.com/the-momentum/open-wearables>.
- Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas C. Jacobson, Mary P. Czerwinski, Eric L. Granholm, and Andrew T. Campbell. 2024. [Mind-Scape study: Integrating LLM and behavioral sensing for personalized AI-driven journaling experiences](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 8(4).
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. [MemGPT: Towards LLMs as operating systems](#). In *Conference on Language Modeling (COLM)*.
- Qwen Team. 2026a. Qwen3.5: Towards native multi-modal agents. <https://qwen.ai/blog?id=qwen3.5>.
- Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model. <https://qwen.ai/blog?id=qwen3.6-27b>.
- Alessio Rossi, Eleonora Da Pozzo, Dario Menicagli, Chiara Tremolanti, Corrado Priami, Alina Sirbu, David Clifton, Claudia Martini, and David Morelli. 2020. [Multilevel monitoring of activity and sleep in healthy people](#). PhysioNet. Version 1.0.0.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [AgentClinic: A multimodal benchmark for tool-using clinical AI agents](#). *npj Digital Medicine*.
- Peter Steinberger. 2026. OpenClaw: Your own personal AI assistant. Any OS. Any platform. <https://github.com/openclaw/openclaw>. Open-source software project, first released as Clawdbot in November 2025. Transfer to the non-profit OpenClaw Foundation announced February 2026.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. [MedAgents: Large language models as collaborators for zero-shot medical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

- 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, Simon Nordvang, Sigurd Pedersen, Anders Gjerdrum, Tor-Morten Grønli, Per Morten Fredriksen, Ragnhild Eg, Kjeld Hansen, Siri Fagernes, Christine Claudi, and 7 others. 2020. [PMData: A sports logging dataset](#). In *Proceedings of the 11th ACM Multimedia Systems Conference (MMSys)*, pages 231–236.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. [A-MEM: Agentic memory for LLM agents](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. 2022. [GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization](#). In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*.
- An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Sofia Yfantidou, Christina Karagianni, Stefanos Efsthathiou, Athena Vakali, Joao Palotti, Dimitrios Pan-teleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. [LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild](#). *Scientific Data*, 9(1):663.

A Report Rubric Breakdown

This appendix reports the complete result tables behind §3, using the same replayed-trace protocol and terminal-state oracles as the main text. Table A1 breaks the data-to-report rubric of §3.2 into its three dimensions on 16 PMData participants.

Backbone	Overall%	Fatigue%	Readi.%	Stress%
DeepSeek-V4-flash [†]	45.8	62.5	31.2	43.8
Qwen3.5-35b-a3b	37.5	25.0	18.8	68.8
GPT-5-mini [†]	35.4	37.5	18.8	50.0
glm-4.7-flash	25.0	37.5	25.0	12.5
Qwen3.6-27b	22.2	20.0	20.0	26.7
nemotron-3-nano	18.7	37.5	6.2	12.5
Gemini-3.5-flash [†]	14.6	6.2	18.8	18.8
Qwen3.5-9b	13.3	20.0	0.0	20.0
Qwen3-32b	11.9	7.1	14.3	14.3
Qwen3-4b	10.4	6.2	18.8	6.2
Qwen3.5-4b	10.4	12.5	6.2	12.5
Qwen2.5-1.5b	8.3	6.2	12.5	6.2
Qwen2.5-32b	8.3	18.8	0.0	6.2
Gemma-4-26b-a4b	6.2	6.2	6.2	6.2
Qwen3-30b-a3b	6.2	18.8	0.0	0.0
Qwen2.5-14b	4.2	6.2	6.2	0.0
Qwen2.5-7b	2.2	0.0	0.0	6.7
Gemma-4-e4b	2.1	0.0	0.0	6.2
Qwen3-8b	2.1	0.0	0.0	6.2
granite-4.1-8b	0.0	0.0	0.0	0.0
Qwen3-14b	0.0	0.0	0.0	0.0
Qwen3-1.7b	0.0	0.0	0.0	0.0

Table A1: Full cross-modal report results on 16 PMData participants: overall rubric percentage and per-dimension hit rate. [†] hosted API.

B Full Per-Role Sweep

Table A2 extends Table 1 of §3.1 to all 22 backbones, sorted by analysis score, and sharpens the main findings. The newest local generations close most of the gap to hosted APIs, with Qwen3.5-9b the cheapest competent configuration, reaching 0.87 analysis with perfect memory and plan scores at only 44k tokens. Single-turn skill does not guarantee session-level reliability. Nemotron-3-nano pairs 0.85 analysis with a collapsed multi-turn score of 0.02, and routing saturates near 1.00 for over half the field yet falls to 0.37 at the 1.5B floor.

C Decomposition Ablation

Table A3 reports the single-turn composed task of §3.1 for the 13 local backbones of the ablation run (the other nine postdate this run). The full list bears out the main text. Six backbones gain from the split, led by Qwen3-8b at +0.65 and Qwen3-32b at +0.23, and five more sit within 0.09 of break-even. Only Qwen3.6-27b at −0.36 and Gemma-4-e4b at −0.71 pay a real cost, while the two smallest fail under either protocol. Factoring into roles therefore costs little even on a single turn, the setting least favourable to it.

Backbone	Rt	An	Mem	Pl	H↓	MT	S5	tok
Gemini-3.5-flash [†]	0.98	0.94	0.89	0.98	0.01	0.98	0.37	42k
Qwen3.6-27b	1.00	0.91	1.00	0.99	0.00	0.76	0.20	253k
Qwen3.5-35b-a3b	1.00	0.91	0.98	0.84	0.00	0.96	0.26	169k
Qwen3.5-9b	1.00	0.87	1.00	1.00	0.00	0.91	0.01	44k
nemotron-3-nano	0.99	0.85	0.93	0.66	0.00	0.02	0.00	52k
DeepSeek-V4-flash [†]	1.00	0.85	0.97	1.00	0.00	0.84	0.00	66k
Qwen3.5-4b	1.00	0.77	0.84	0.18	0.00	0.66	0.00	50k
GPT-5-mini [†]	0.88	0.76	0.99	0.88	0.00	0.63	0.00	206k
Qwen2.5-32b	1.00	0.72	0.66	0.07	0.00	0.74	0.22	159k
Qwen3-32b	1.00	0.68	0.81	0.80	0.00	0.69	0.00	178k
granite-4.1-8b	0.60	0.64	0.93	0.80	0.00	0.00	0.00	34k
Qwen3-14b	1.00	0.52	0.79	0.00	0.00	0.44	0.00	162k
Gemma-4-26b-a4b	1.00	0.48	0.93	0.33	0.26	0.89	0.44	236k
Qwen3-8b	0.61	0.47	0.96	0.57	0.00	0.68	0.00	194k
Qwen3-4b	1.00	0.43	0.67	0.37	0.00	0.57	0.00	33k
Qwen3-1.7b	0.50	0.41	0.01	0.99	0.00	0.26	0.00	123k
Qwen2.5-7b	0.60	0.38	0.66	0.91	0.12	0.31	0.00	38k
Qwen3-30b-a3b	0.97	0.36	0.33	0.99	0.00	0.61	0.00	242k
glm-4.7-flash	1.00	0.27	0.83	0.92	0.00	0.00	0.00	33k
Qwen2.5-14b	0.74	0.15	0.63	0.22	0.22	0.62	0.00	102k
Gemma-4-e4b	1.00	0.02	0.67	0.00	0.00	0.37	0.00	144k
Qwen2.5-1.5b	0.37	0.00	0.00	0.00	0.47	0.26	0.00	20k

Table A2: Full per-role capability over 22 backbones (pass@1, three runs), sorted by analysis; abbreviated columns as in Table 1. [†] hosted API.

Backbone	Dec	Mono	Δ
Qwen3-8b	0.98	0.33	+0.65
Qwen3-32b	0.52	0.29	+0.23
Qwen2.5-7b	0.21	0.09	+0.12
Qwen3-14b	0.77	0.66	+0.11
Qwen3-30b-a3b	0.29	0.22	+0.07
Gemma-4-26b-a4b	0.69	0.63	+0.06
Qwen3-1.7b	0.00	0.00	0.00
Qwen2.5-1.5b	0.00	0.00	0.00
Qwen3.5-35b-a3b	0.90	0.91	−0.01
Qwen2.5-32b	0.07	0.11	−0.04
Qwen2.5-14b	0.80	0.89	−0.09
Qwen3.6-27b	0.63	0.99	−0.36
Gemma-4-e4b	0.01	0.72	−0.71

Table A3: Decomposition ablation, decomposed (Dec) versus monolithic (Mono), for the 13 local backbones in the ablation run, sorted by Δ.

D Trigger Detection Protocol

This appendix details the detection experiment of §3.3. The *same* agent invoked two ways, **polling** (full multi-turn analysis on a fixed five-minute schedule) versus **trigger** (identical analysis only when an agent-authored condition fires). The backbone authors its own rules from one plain-language request covering spikes, sustained elevation, slow heart rate, low blood oxygen and low heart-rate variability, each choosing a signal, a threshold or standard-deviation condition, a window and a cooldown. We replay PMData, LifeSnaps and MMASH, the three intraday heart-rate corpora, under a frozen clock that ingests batches in order so a rule only sees samples up to the simulated present. Since no corpus labels real anomalies, we inject a fixed, diverse set of events as ground truth. Heart-rate spikes and drops (baseline mean $\pm$ several standard deviations, floored or ceilinged at a fixed bpm),

Backbone	Params	Rundown↓ (s)
Qwen2.5-1.5b	1.5B	4.0
glm-4.7-flash	31B [†]	4.3
Qwen2.5-7b	7B	9.4
Qwen3-1.7b	1.7B	10.8
nemotron-3-nano	4B	12.2
Qwen3-4b	4B	12.3
Gemma-4-e4b	8B	12.4
Qwen3.5-35b-a3b	35B [†]	21.3
Qwen3.5-9b	9B	23.8
Qwen2.5-14b	14B	26.7
Qwen3.5-4b	4B	27.9
Qwen3-14b	14B	29.4
Gemma-4-26b-a4b	26B [†]	32.8
Qwen3-30b-a3b	30B [†]	34.7
Qwen2.5-32b	32B	44.6
Qwen3-8b	8B	45.6
Qwen3-32b	32B	55.9
Qwen3.6-27b	27B	97.3
granite-4.1-8b	8B	230.2 [‡]

Table A4: Composite-rundown latency on one NVIDIA H200, median over five PMData participants. [†] mixture-of-experts. [‡] unreliable: loops on malformed tool calls until timeout.

a sustained tachycardic run, a blood-oxygen dip to 86%, and a low heart-rate-variability sample, each replacing the real reading in its batch so it is the latest value the signal has when the gate evaluates. A fire is credited only within a 45-minute grace window after onset, and a fixed LLM judge (Liu et al., 2023) checks, per event, whether the woken analysis surfaced that abnormality, flagging unsupported concerns as false alarms. We report recall on events whose gate fires, over five participants per corpus and 47 events per backbone; the token saving in Table 2 is the ratio of the trigger arm’s few daily analyses to polling’s fixed schedule.

E On-Device Latency Profile

The main text reports token count because it is hardware-independent, but HiME runs on the user’s own machine, so wall-clock latency matters too. Table A4 reports the *rundown* for every local backbone, the end-to-end wall-clock of the five-turn composite session of §3.1 (the session scored there for MT and S5), each served alone on one NVIDIA H200 through vLLM (Kwon et al., 2023) in bfloat16 (bf16) with reasoning backbones thinking, median over five PMData participants. Decode throughput at batch one ranged from 53 tok/s at 32B to 536 tok/s at 1.5B, tracking active rather than total parameters, so the 30B mixture-of-experts decoded as fast as a dense 4B.

F Human Evaluation Survey Design

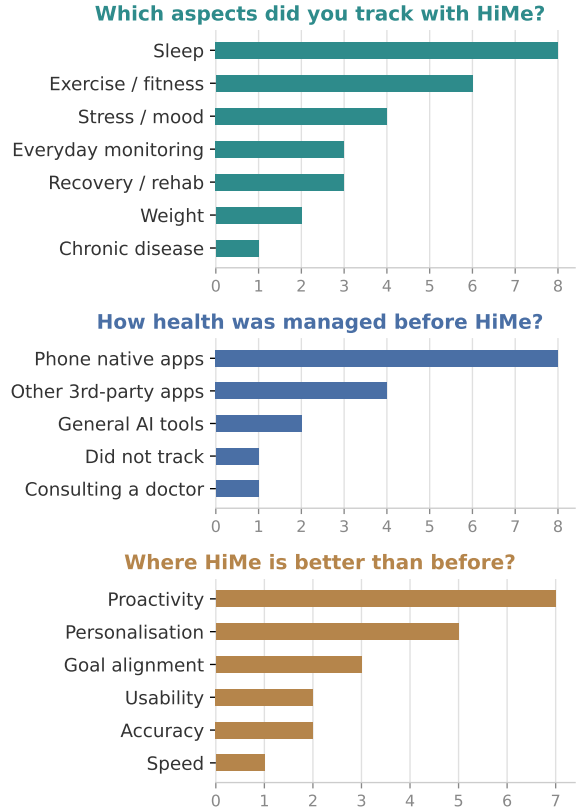


Figure A1: Participants’ multiple-choice responses on HiME usage scenarios, prior health-management methods, and advantages over previous approaches.

This appendix gives the survey instrument behind §3.4. Beyond the Likert questionnaire (Figure 5), participants answered three multiple-choice questions, including which health domains they tracked, how they managed those before HiME, and where HiME improved on their previous approach (Figure A1). Participants adopted HiME as a daily sleep companion, extending to fitness and stress management, having relied mostly on fragmented apps before. They credited HiME with proactivity and personalisation gains, fewer in usability, accuracy or speed (§3.4).