MEDOMNI: BENCHMARKING OMNI-MODALITY INTEGRATION IN MEDICAL AGENTIC SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decisions require reasoning over heterogeneous medical evidence, yet existing benchmarks typically restrict each task to a small set of predefined modalities. This leaves medical agents' ability to select, access, and integrate broad evidence insufficiently evaluated. We introduce MEDOMNI BENCH, a clinically grounded benchmark developed with a multidisciplinary clinical consortium. To our knowledge, under our counting criterion, MedOmni provides the broadest modality coverage per task and is the first benchmark designed specifically for omni-modal medical agent evaluation. Its specialized infrastructure decouples tasks, modalities, and encoders through an extensible modality perception service, allowing new modalities to be incorporated without redesigning existing tasks. A structured assurance mechanism connects benchmark requirements to auditable checks across data, perception, tasks, and evaluation runs. Experiments reveal substantial differences in how agents benefit from and reason over heterogeneous evidence, identifying evidence selection and integration as central challenges for omni-modal medical intelligence. Code and resources are available at `https://anonymous.4open.science/r/MedOmni`.

1 INTRODUCTION

Medical AI has progressed from specialized foundation models and models spanning vision and language (Li et al., 2023; Saab et al., 2024) to agents that access diverse medical data through external perception models (Fallahpour et al., 2025; Ashraf et al., 2026). The shift is from a fixed encoder that reads a few modalities to an agent that can recruit specialist models as tools. That access matters because clinical evidence is assembled per patient. One case may offer only a note and a laboratory panel; another may add CT, pathology, and waveform data that were never collected together in a training corpus. We refer to *omni-modal medical intelligence* as the ability to access, select, and integrate flexible combinations of supported modalities specific to each patient, rather than a fixed pair or a closed catalog.

The first gap is *evidence availability*. Existing benchmarks may report broad modality coverage while exposing only a narrow, fixed evidence set in each item. OmniMedVQA spans twelve imaging modalities but presents almost every question with a single image (Hu et al., 2024). MedXpertQA covers eleven modalities, yet a multimodal item still pairs a clinical narrative with a small set of images (Zuo et al., 2025). ClinicRealm evaluates notes and structured EHR largely as separate unimodal tasks (Zhu et al., 2026a). Table 1 makes the distinction explicit: a large catalog does not imply broad combinations specific to each patient within a task. The second gap is *evidence use*. Access to more evidence does not ensure that an agent will identify what matters or integrate it effectively. Modality value varies by task (Soenksen et al., 2022), while conflicting inputs can induce unimodal shortcuts or unstable fusion (Leng et al., 2026; Zhang et al., 2025; Zou & Wei, 2026). Existing evaluations rarely measure whether agents select relevant modalities, access them through an effective route, and incorporate their evidence into the final decision. The third gap is *infrastructure and validity*. Flexible evidence is useful only if agents can interpret it and the resulting evaluation remains trustworthy. Agents lack extensible infrastructure for recruiting new specialist encoders as modalities appear, leaving evidence outside a model's native interface difficult to use. At the same time, expanding the evidence and tool surface increases the need to scrutinize task construction, encoder reliability, and evaluation integrity. This follows broader calls for transparent reporting and lifecycle assurance in clinical AI (Collins et al., 2024; Lekadir et al., 2025). Without

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

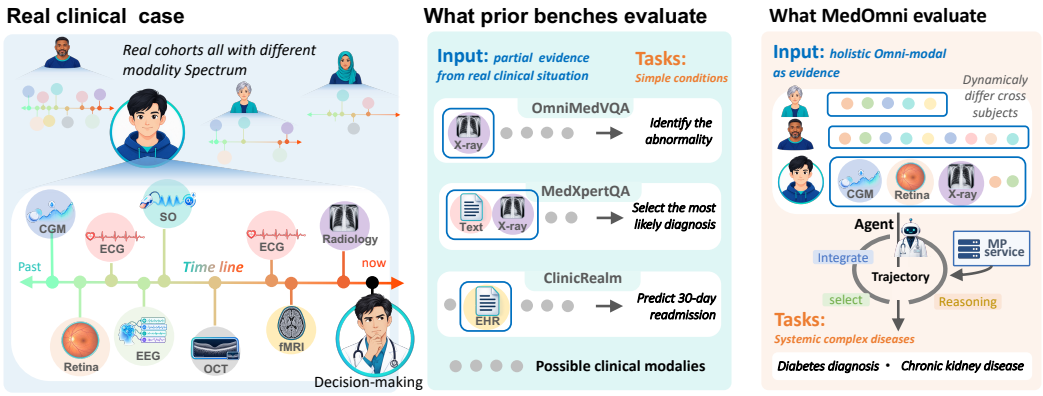

Figure 1: From fixed-modality benchmarks to omni-modal medical agent evaluation. While prior benchmarks expose narrow inputs within each task, MedOmni evaluates agents that select and integrate broad modality combinations within a single task using evidence from real clinical cohorts, with specialist support from the MP service.

both perception infrastructure and benchmark assurance, broader evidence can remain inaccessible or introduce untracked failure.

Together, these gaps leave unclear whether agents can benefit from broad evidence specific to each patient, select what matters, and recruit specialized perception when native capabilities are insufficient. *These observations motivate a concrete question: can medical agents dynamically access, select, and integrate evidence from arbitrary combinations of supported modalities?*

We introduce MEDOMNI BENCH, an omni-modal medical agent evaluation platform that stages evidence specific to each patient under standardized input contracts, grounds tasks in clinical questions, exposes specialized encoders through the modality perception (MP) service, and audits the resulting assets and runs through an assurance mechanism (Figure 1). Under the item-level criterion in Table 1, it is, to our knowledge, the most heterogeneous medical benchmark, exposing about 13 modalities per task on average. The evaluation measures both decision correctness and evidence use. Across six agents and 30 task types, controlled evidence expansion produces substantial gains that vary by model and task. The MP service provides its clearest gains on raw physiological signals, while consulting broader evidence does not by itself ensure better decisions.

This paper makes three contributions.

- A clinically grounded benchmark developed with a multidisciplinary clinical consortium. To our knowledge, it provides the broadest modality coverage per task under our counting criterion and is the first designed to directly evaluate agents over flexible, patient-specific modality combinations.
- Specialized infrastructure that decouples tasks, modalities, and encoders, allowing new or currently unsupported modalities to be added without redesigning existing tasks.
- New evidence that modality combinations, interactions, and conflicts shape performance and produce distinct behaviors across agents.

2 RELATED WORK

Medical multimodal models and benchmarks. Medical foundation models and vision–language models have expanded beyond text to clinical images and records (Li et al., 2023; Saab et al., 2024). Their evaluations, however, generally retain a fixed input interface. OmniMedVQA covers diverse medical imaging but presents almost every item as a single image (Hu et al., 2024); MedXpertQA combines expert questions with a small number of images (Zuo et al., 2025); and ClinicRealm and X-PCR broaden coverage without requiring flexible, patient-specific evidence combinations (Zhu et al., 2026a; Wang et al., 2026a). This distinction matters because multimodal evidence is not uni-

formly additive: its contribution varies by task, and conflicting inputs can induce unimodal shortcuts or hallucinations (Soenksen et al., 2022; Leng et al., 2026; Zhang et al., 2025; Zou & Wei, 2026).

Medical agents and perception tools. Tool-using systems shift the interface from a fixed native encoder to an agent that can invoke specialized components (Karpas et al., 2022; Schick et al., 2023). Medical evaluations have begun to study this setting. ChestAgentBench evaluates an agent with chest-radiograph tools (Fallahpour et al., 2025), MedCTA provides executable tools for image-based clinical tasks (Ashraf et al., 2026), and MedAgentBoard spans several healthcare task families (Zhu et al., 2026b). CHI-Bench instead emphasizes operational workflows across healthcare applications (Chen et al., 2026). These benchmarks establish the value of agentic execution, but they do not jointly evaluate dynamic selection and integration over broad, heterogeneous clinical evidence.

Positioning of MEDOMNI BENCH. MEDOMNI BENCH complements this literature by treating the available modality families as patient-specific, exposing extensible perception through the MP service, and grounding task construction in clinician scrutiny and lifecycle assurance. Appendix A.1 documents the item-level counting assumptions behind our comparison, including average modality families per task, executable perception access, and licensed-clinician involvement.

3 MEDOMNI: BENCHMARK DESIGN AND INFRASTRUCTURE

To evaluate the omni-modal capability in Section 1, MEDOMNI BENCH is built as a platform rather than a fixed multimodal corpus (Figure 2). It stages patient-specific evidence under shared contracts, lets clinicians define decisions, provides specialized perception, and keeps runs auditable through three layers: clinical data and tasks, an MP service, and an assurance mechanism. These layers structure the section.

3.1 CLINICAL DATA AND TASK CONSTRUCTION

What a clinician can use is whatever a given patient actually has. Omni-modal evaluation therefore cannot assume preset inputs: each task presents a patient-specific, clinically admissible set of modality families (Section 1). We place complementary evidence from three sources under unified contracts and let clinicians define the questions. Figure 2 summarizes data, task construction, and execution, with details in Appendix B.

Large-scale clinical data spanning extensive modalities. No single dataset reflects the full clinical evidence landscape. The upstream releases contain 364,627 MIMIC-IV patients (Johnson et al., 2023), 2,280 deeply phenotyped AI-READI participants (Owsley et al., 2025), and 90,166 BDSP patients with 119,234 sleep recordings (Li et al., 2026). The benchmark instantiates 314 unique patients from more than 457,000 represented upstream, leaving over 1,000-fold capacity for expansion subject to eligibility, modality availability, and assurance. Their complementary evidence spans longitudinal acute-care records, imaging, ECGs, and bedside waveforms; metabolic, retinal, cardiac, and wearable phenotyping; and dense overnight physiology linked to clinical records. This supports tasks across acute care, chronic disease assessment, and sleep medicine. Table 2 and Appendix B.1 detail scale, roles, and filtering.

Modality families. A modality family is a versioned contract for inputs decodable into the same normalized representation without changing acquisition or clinical semantics. It specifies format, dimensionality, channel or image-plane semantics, sampling rate, and units. This is stricter than a broad clinical modality: diagnostic 12-lead ECG, single-lead sleep ECG, and continuous bedside ECG form distinct families because their acquisition and channel semantics differ. Defining families independently of encoder availability keeps unsupported evidence visible and permits new encoders without changing tasks. The data pool supplies 34 families spanning physiological waveforms, medical images and volumes, wearable time series, and clinical tables; Appendix B.2 reports the registry.

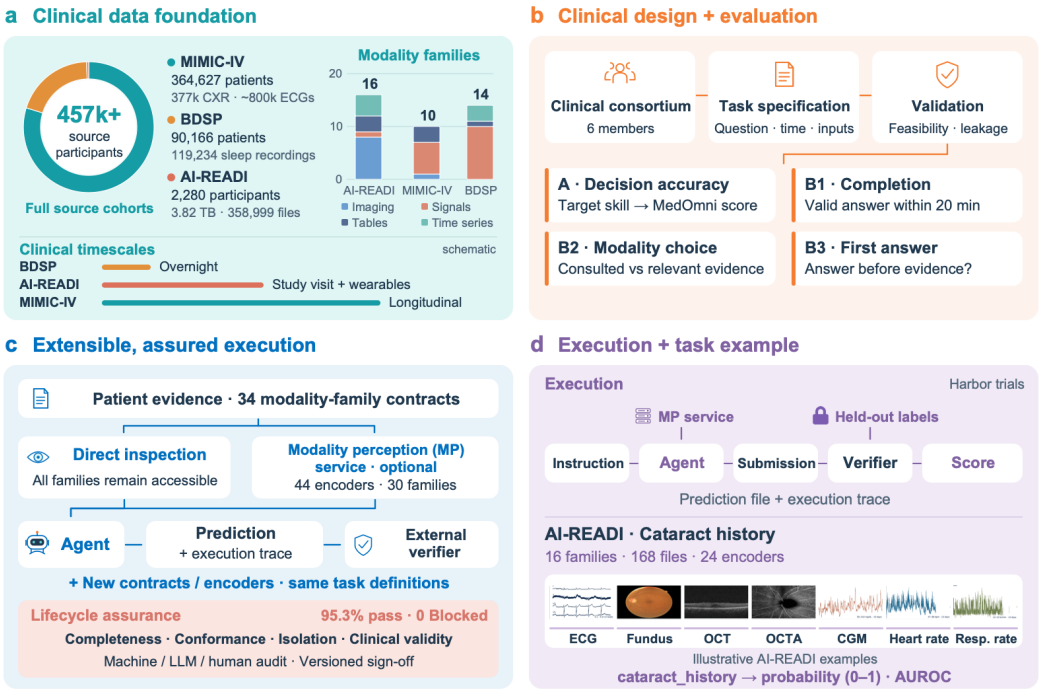

Figure 2: **MedOmni platform overview.** **(a)** Three complementary clinical sources provide broad evidence coverage across modality families and timescales. **(b)** A clinical consortium specifies tasks that undergo feasibility and leakage checks, with evaluation spanning decision accuracy and behavior. **(c)** Modality-family contracts support direct inspection or optional MP service access, while lifecycle assurance covers the full execution path. **(d)** Harbor execution and an AI-READI cataract task illustrate how one patient-level instance exposes broad evidence for agent-directed selection and prediction.

Clinician-grounded task construction. A large multimodal corpus does not, by itself, constitute a clinically meaningful benchmark. Clinical questions, not extractability, must determine what is predicted, when, and from which evidence.

Clinical consortium. Our six-member multidisciplinary consortium includes licensed clinicians and academic researchers from hospital and university sites and is designed to grow. Members propose tasks and review more than 1,000 candidate labels with provenance, prevalence, and timing on the MEDOMNI Task Proposal website. Of 34 proposals, 28 passed pre-construction assurance and yielded 29 task types because one MIMIC-IV heart-failure proposal was instantiated in two evidence configurations; six proposals were filtered. One additional clinically reviewed CKD-staging task was constructed outside the proposal intake, producing the complete suite of 30 task types (Table 7).

Each proposal becomes a frozen specification of target derivation, prediction boundary, admissible families, metric, and verifier. The builder checks label and evidence availability, constructs the eligible cohort, and instantiates patient cases. Prediction boundaries remove post-decision information, while target-specific exclusions block circular evidence. These task-side leakage controls complement the assurance checks in Section 3.3; Appendices B.3 and B.4 provide the full procedure.

Task suite and evaluation. Our benchmark couples a clinically diverse task suite with complementary evaluation criteria and a unified, model-agnostic framework that standardizes task packaging, parallel agent execution, and external scoring.

Evaluation criteria. We evaluate agents at two complementary levels: the correctness of their clinical decisions and the reasonableness of their evidence-seeking behavior. *Decision correctness* scores each final prediction with the target-specific metric and verifier appropriate to the clinical task. *Behavioral reasonableness* examines whether the agent consults and integrates relevant evi-

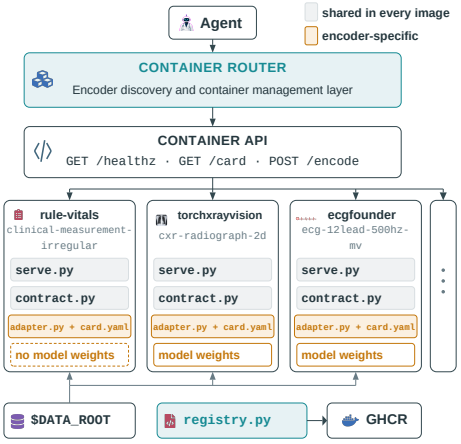

Figure 3: Modality perception service architecture.

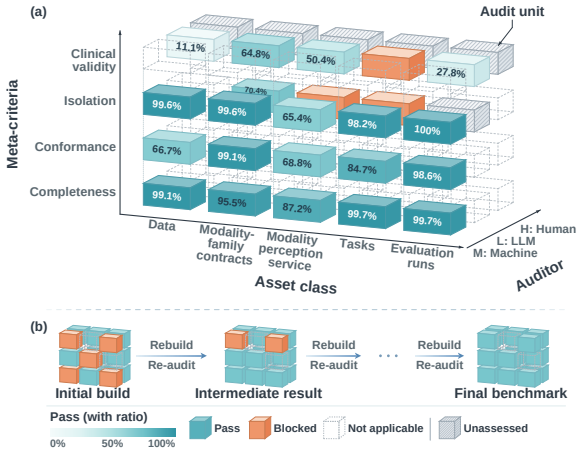

Figure 4: (a) Assurance matrix. (b) Iterative refinement process.

dence, using modality-family access and MP service traces together with efficiency records. These criteria are applied across 600 instances, 30 task types, and 59 scored target slots, while refusals, malformed predictions, and unscorable targets remain explicit outcomes. The complete task census and evaluation criteria are reported in Appendices B.4 and B.5.

Harbor framework. We conduct the full evaluation harness through a reliable, up-to-date agentic evaluation framework rather than a task-specific runner. Harbor, introduced with Terminal-Bench (Merrill et al., 2026), provides this substrate: it runs heterogeneous agents and models on isolated sandbox trials in parallel. We adopt it so that the same clinically specified task can be replayed under different agent scaffolds and model backbones without rewriting its evidence or scoring contract. Each task instance is packaged as one self-contained Harbor trial containing its patient evidence, instructions, available tools, output contract, and verifier. Trials run in isolated containers, while scoring remains external to Harbor and common across runners. This separation keeps execution details from changing the definition or score of a task. Appendix B.6 provides the execution protocol.

3.2 Modality Perception Service

A native encoder cannot cover the combinations required by omni-modal evaluation (Section 1). The MP service therefore packages specialized encoders behind a standardized, containerized interface for modalities an agent cannot natively read. Encoders share input and evidence contracts while retaining their dependencies and, where applicable, weights, and the service handles discovery, execution, routing, and lifecycle through one agent-facing API (Figure 3).

Using the modality-family contracts defined in Section 3.1, the service routes each staged asset by its acquisition and clinical semantics rather than by a broad modality label. Encoders assigned to one modality family consume the same normalized input representation but may produce different evidence: categorical or continuous labels, time-localized events, text, or vector embeddings. This separation decouples data and perception coverage: a modality remains available even without an encoder, while new contracts and encoders can extend the platform without altering existing tasks.

Learned and rule-based encoders. The MP service exposes 44 runnable family-specific encoder entries for 30 of 34 modality families. Entries sharing one of 29 concise encoder names are distinguished by unique model IDs. Learned encoders include representations, classifiers, regressors, and event detectors, while rule-based encoders compute deterministic summaries or measurements without necessarily having model weights. None is trained or fine-tuned on MedOmni data. Both types share the evidence API and undergo the same provenance, compatibility, and runtime checks, with acceptance tracked separately.

The four families without runnable encoders remain available for direct inspection. Appendix C provides implementation details.

Unified agent interface. The MP service presents the encoder registry as one agent-facing tool API. This follows the broader systems view that language models can delegate bounded operations to specialized modules through explicit interfaces (Karpas et al., 2022; Schick et al., 2023). An agent can discover compatible encoders, inspect their declared capabilities and limitations, and request evidence without managing the encoder runtime.

Each encoder card declares compatibility, intended use, limitations, outputs, provenance, reliability evidence, and acceptance status. Adapting model-card principles (Mitchell et al., 2019), weight fields apply only when weights exist, while source revision, behavior, and evidence APIs apply to all encoders. A common output schema lets agents and assurance checks treat heterogeneous implementations consistently.

3.3 Assurance Mechanism

MedOmni Bench combines clinical data, heterogeneous modalities, an MP service, tasks, and evaluation runs, so reliability requires scrutiny of clinical validity, data integrity, and reproducibility (Alaa et al., 2025; Wang et al., 2026c;b). Drawing on structured assurance (Paterson et al., 2025), our assurance matrix links requirements to concrete checks, evidence, and responsible auditors.

Assurance matrix. The matrix crosses the *asset class*, *meta-criterion*, and *auditor*; each intersection defines an audit unit containing atomic checks (Figure 4).[1] Class-level obligations are bound to versioned assets in audit records. *Asset classes* span the evaluation pipeline: Data, including source files, manifests, and provenance; Modality-family contracts, specifying input formats and acquisition semantics; MP service, covering encoders and shared service infrastructure; Tasks, including frozen specifications and patient instances; and Evaluation runs, comprising configurations, predictions, traces, and scoring records. *Meta-criteria* are Completeness, whether required artifacts and evidence are present; Conformance, adherence to declared contracts and reproducibility requirements; Isolation, information and execution boundaries, including leakage; and Clinical validity, whether populations, acquisition semantics, task definitions, and observed behavior support their intended clinical interpretation. *Auditors* are machine auditors (M) that execute checks, LLM auditors (L) that provide advisory review, and human auditors (H) that provide technical or clinical sign-off. Assignments reflect the judgment required: task isolation combines machine and LLM audits, while clinical validity combines LLM advice and human review. *Audit outcomes* are three-valued. For an applicable unit with assessed checks and no critical violation, Pass (with ratio) reports passing / assessed applicable checks; non-critical, non-passing checks remain warnings. Blocked denotes a benchmark-compromising violation, such as leakage, that prevents sign-off regardless of the pass ratio. Not applicable is reserved for combinations assigned no checks. Assessment completion is recorded separately, so missing evidence cannot count as a pass or justify Not applicable.

Sign-off and iterative refinement. Each check has a unique identifier and a fingerprinted record of asset version, evidence, auditor, and result. Machine, LLM, technical, and clinical auditors use a common version-bound sign-off mechanism; LLM advice cannot replace human approval. Clinical consultation becomes formal evidence only when its scope and verdict are bound to the audited version.

Sign-off requires completed checks, documented warning dispositions, and no unresolved Blocked outcome. Findings prompt corrections, while changed fingerprints trigger affected checks and reviews again, keeping evaluation evidence traceable as the benchmark grows. For the reported release, this cycle produced no Blocked audit units and a 95.3% pass rate across assessed applicable checks. Appendix D provides the check catalog, aggregation rules, and a worked example.

[1] The outcomes shown in Figure 4 are synthetic illustrations rather than the final audit state.

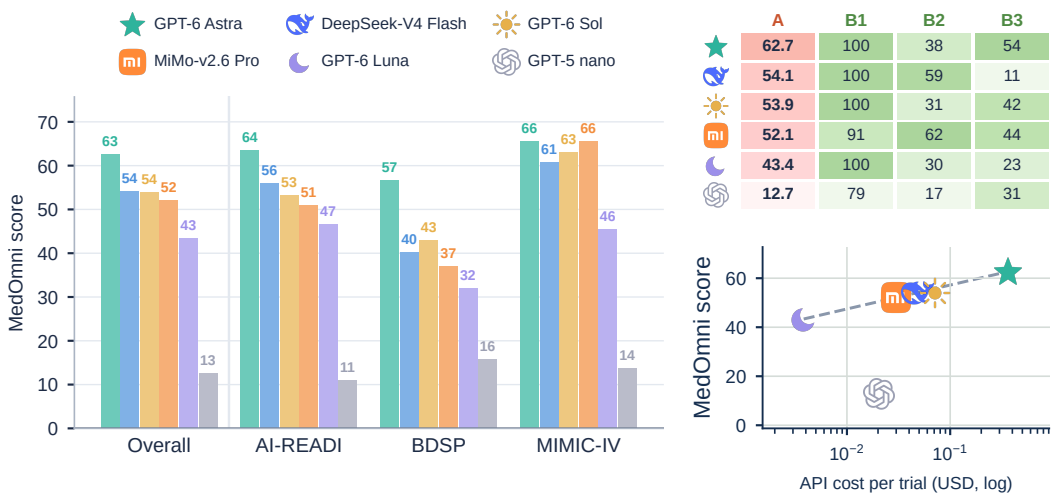

Figure 5: **MedOmni performance. Left:** MedOmni score (A), overall and by source dataset, over 600 instances per model. **Top right:** Overall protocol results, one row per model: A (coral) and B1–B3 (green; B1 and B3 in %, B2 as F1), each column shaded on its own range; CIs and details in Appendix F. **Bottom right:** Score against API cost per trial (log scale), with the cost-score frontier dashed.

4 Benchmarking Omni-modal Medical Agents

4.1 Implementation Details

Agents. We evaluate GPT-6 Astra, GPT-6 Sol, and GPT-6 Luna (OpenAI, 2026), DeepSeek-V4 Flash (Xu et al., 2026), GPT-5 nano (OpenAI, 2025), and MiMo-v2.6 Pro (Xiao et al., 2026) (below: Astra, Sol, Luna, DeepSeek, nano, MiMo) in the same terminus-2 agent scaffold on Harbor. Each trial places one patient instance in an isolated container with a 20-minute budget and at most 300 turns. The agent receives the task instruction, staged patient files, and the modality perception (MP) service, but is neither required nor forbidden to use an encoder. Harness, MP service, and split details are provided in Appendices E.1–E.3.

Evaluation protocol. The split has 30 task types with 20 patients each (600 instances per model) across AI-READI, BDSP, and MIMIC-IV. We report four criteria, defined in full in Appendix E.4. *(A) MedOmni score*: each target is scored from chance (0) to perfect (100) by rescaled AUROC, quadratic weighted kappa, or MAE skill, then averaged over targets and all 30 task types, with 95% bootstrap CIs. *(B1) Completion*: the share of trials that return a valid answer. *(B2) Modality choice*: the F1 agreement between the modalities an agent consults and those a human marked relevant. *(B3) First answer*: the share of evidence-consulting trials that answer before consulting any modality.

4.2 Leaderboard Results

Main leaderboard. GPT-6 Astra leads with 62.7 (Figure 5, left), significantly ahead of DeepSeek-V4 Flash (54.1), GPT-6 Sol (53.9), and MiMo-v2.6 Pro (52.1), which are statistically tied (paired differences $\leq$ 2.0 points). GPT-6 Luna scores 43.4 and GPT-5 nano 12.7, close to chance. Performance also varies by source dataset. BDSP is the hardest source for every model above chance, exposing a shared weakness when decisions depend on raw overnight signals rather than the clinical record. No agent dominates across all 30 task types, and the strongest agent varies by task. Full scores and paired comparisons are reported in Appendix F.1.

Cost and behaviour. Higher performance carries a substantial cost (Figure 5, bottom right). Astra costs $0.36 per trial, approximately five times Sol and eight times DeepSeek for about nine additional points. Behaviour also differs across agents (Figure 5, top right). Completion is near-perfect for the GPT-6 models and DeepSeek but lower for MiMo and GPT-5 nano due to timeouts and unreadable answers. DeepSeek and MiMo consult broader evidence, whereas the GPT-6 models se-

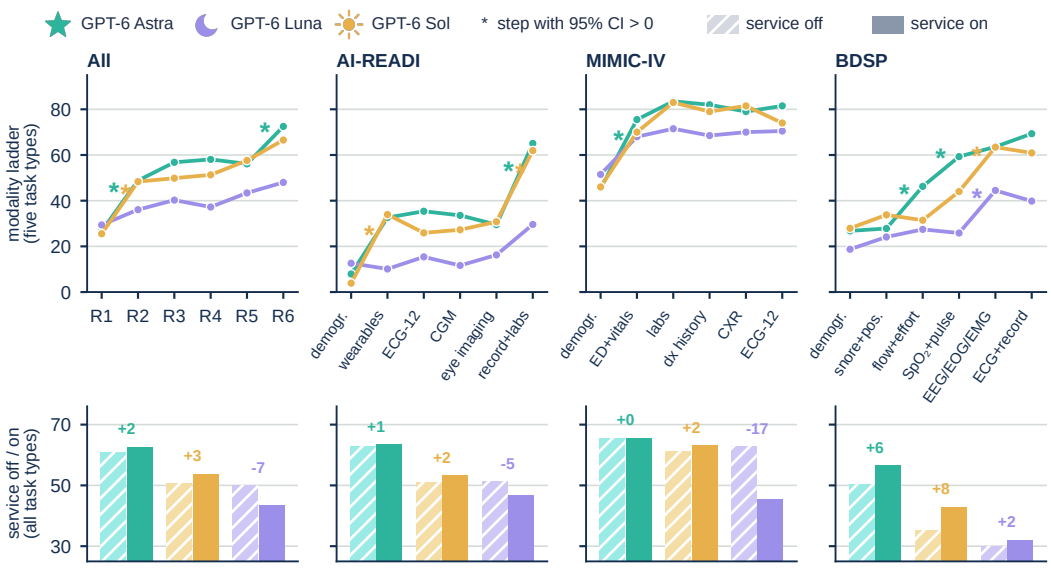

Figure 6: **Testing omni-modal evidence use. Top:** Scores as modality families are added cumulatively from demographics to the full evidence surface. **Bottom:** Scores with the MP service off (hatched) and on (solid) on matched instances; numbers give the paired difference; the axis starts at 25. Asterisks mark ladder steps whose paired 95% confidence intervals exclude zero.

lect fewer modalities. Astra more often drafts an initial answer before consulting modalities, while DeepSeek more often investigates the available evidence first. Detailed behaviour and cost results appear in Appendix F.2.

4.3 Understanding Omni-modal Performance

To understand what drives the omni-modal capability in Section 1 beyond aggregate scores, we conduct three complementary analyses. A controlled modality ladder progressively expands the available evidence, a matched ablation removes access to the MP service, and task-level scores and traces reveal how agents select and use evidence. Together, these analyses examine evidence availability, perception support, and integration behavior.

4.3.1 Evaluating Omni-modal Evidence Integration

The modality ladder directly tests whether medical agents can convert broader heterogeneous evidence into better decisions. Holding patients and tasks fixed, we reveal modality families cumulatively across six rungs (Figure 6, top; Appendix F.3). All three agents improve from demographics alone to the full evidence surface: Astra gains 45.8 points, Sol 41.0, and Luna 18.6. The source-specific trajectories reveal where these gains arise. MIMIC-IV improves sharply when emergency observations and laboratory results become available, then largely plateaus. BDSP continues to benefit as respiratory, oxygenation, and neurophysiological signals are added, while AI-READI gains most when the clinical record and laboratory results complete the evidence surface. Intermediate rungs are not uniformly monotonic, showing that access to more modalities does not itself guarantee better decisions. Agents can therefore integrate heterogeneous evidence, but their ability to identify and exploit each addition remains strongly model- and task-dependent.

4.3.2 Contribution of the MP Service

Matched runs with and without the MP service demonstrate its value as an extension of agents' native modality capabilities (Figure 6, bottom; Appendix F.4). The strongest agent, Astra, preserves its overall performance with a positive difference of $+1.8$ points, while Sol gains $+3.3$. More importantly, all three agents improve on BDSP ($+6.2$, $+7.6$, and $+2.1$), where MP service use is most frequent and decisions depend on raw physiological signals. This consistent gain shows that

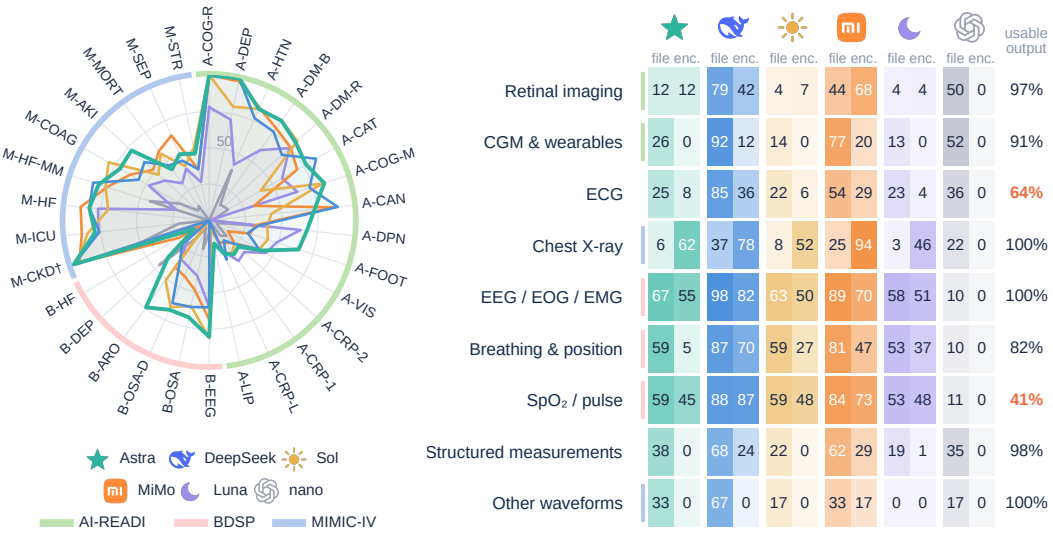

Figure 7: **Task variation and evidence-access routes. Left:** Each spoke is one task type, with higher scores farther from the center; the outer ring identifies its source dataset (abbreviations in Table 12). **Right:** For each modality group and model, the share of trials with direct raw-file access or an MP service call. The marker gives the share of calls returning a healthy, usable output.

specialized perception can unlock useful evidence from modalities that are difficult for agents to interpret directly. Luna's lower aggregate score, concentrated on MIMIC-IV, further indicates that the realized benefit depends on effective evidence integration. Overall, the MP service broadens the usable evidence surface and delivers its clearest gains where specialized modality support matters most.

4.3.3 Evidence Use and Task-Level Variation

Task-level variation. The aggregate leaderboard conceals substantial variation across all 30 tasks (Figure 7, left). Although Astra leads overall, the strongest model changes across task types. Omni-modal capability is therefore not a single, uniform skill. It depends on the clinical question and the evidence required to answer it.

Evidence-access routes. We reconstruct whether each agent reads a modality directly or accesses it through the MP service (Figure 7, right). The GPT-6 models use the MP service for chest radiographs but read breathing, CGM, and wearable signals directly, demonstrating that agents adapt their access route to the modality. Calls return usable outputs across most modality groups, with SpO_2/pulse (41%) and ECG (64%) remaining the main bottlenecks. DeepSeek and MiMo consult the broadest evidence yet trail Astra overall, showing that coverage alone is insufficient. The central challenge is to select task-relevant evidence, access it through an effective route, and integrate it into the final decision. Appendix G.1 provides holistic trace analyses, and Appendix G.2 presents two randomly selected runtime case studies.

5 Conclusion

MedOmni Bench establishes a clinically grounded setting for evaluating omni-modal medical agents. Its consortium-developed tasks, extensible perception infrastructure, and structured assurance mechanism support reliable evaluation while allowing modalities and encoders to evolve independently. Our results show that agents differ substantially in how they use heterogeneous evidence and respond to modality combinations, interactions, and conflicts. These findings demonstrate that omni-modal intelligence requires more than broad evidence access: agents must select, interpret, and integrate the right evidence for each clinical task.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, LaTeX and code drafting, and consistency checks. They were not used to generate clinical labels, determine benchmark ground truth, make clinical judgments, or independently validate experimental findings. The authors reviewed all AI-assisted text and code, verified claims and citations against source materials, and take responsibility for the final content of this work.

REFERENCES

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Tajamul Ashraf, Hyewon Jeong, Fida Mohammad Thoker, and Bernard Ghanem. Medcta: A benchmark for clinical tool agents. *arXiv preprint arXiv:2606.11702*, 2026.

Carl Boettiger. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, 2015. doi: 10.1145/2723872.2723882.

Haolin Chen, Deon Metelski, Leon Qi, Tao Xia, Joonyul Lee, Steve Brown, Kevin Riley, Frank Wang, TY Liu, Hank Capps MD, et al. Chi-bench: Can ai agents automate end-to-end, long-horizon, policy-rich healthcare workflows? *arXiv preprint arXiv:2605.16679*, 2026.

Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15661–15676. PMLR, 13–19 Jul 2025.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tennenholtz. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*, 2022. doi: 10.48550/arXiv.2205.00445.

Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, et al. Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *bmj*, 388, 2025.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *Advances in Neural Information Processing Systems*, 38, 2026.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Qichen Li, Shenghan Wen, Haoqi Sun, Wolfgang Ganglberger, Ayush Tripathi, Niels Turley, Samuel Waters, Arnav Gupta, Aditya Gupta, Manohar Ghanta, et al. The human sleep project: A multicenter clinical polysomnography dataset across the human lifespan. *Sleep*, pp. zsag215, 2026.

Mike Merrill, Alexander Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, volume 2026, pp. 40903–40986, 2026.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596.

OpenAI. GPT-5 System Card, 2025. URL `https://cdn.openai.com/gpt-5-system-card.pdf`.

OpenAI. GPT-6 Astra System Card, 2026. URL `https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf`.

Cynthia Owsley, Dawn S Matthies, Gerald McGwin, Jeffrey C Edberg, Sally L Baxter, Linda M Zangwill, Julia P Owen, Cecilia S Lee, and AI-READI Consortium. Cross-sectional design and protocol for artificial intelligence ready and equitable atlas for diabetes insights (ai-readi). *BMJ open*, 15(2):e097449, 2025.

Colin Paterson, Richard Hawkins, Chiara Picardi, Yan Jia, Radu Calinescu, and Ibrahim Habli. Safety assurance of machine learning for autonomous systems. *Reliability Engineering & System Safety*, 264:111311, 2025.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Gui Wang, Zehao Zhong, YongSong Zhou, Yudong Li, Ende Wu, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. X-pcr: A benchmark for cross-modality progressive clinical reasoning in ophthalmic diagnosis. *arXiv preprint arXiv:2604.20350*, 2026a.

Wenxuan Wang, Zizhan Ma, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. Beyond the leaderboard: Rethinking medical benchmarks for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43078–43123, 2026b.

Xiaoda Wang, Ching Chang, Defu Cao, Kaiqiao Han, Fang Sun, Yue Huang, Minxiao Wang, Chang Xu, Xiao Luo, Runze Yan, et al. Position: Beyond prediction: Toward verifiable physiological waveform reasoning with foundation models and agentic llms. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026c.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77233–77253. PMLR, 13–19 Jul 2025.

Yinghao Zhu, Junyi Gao, Zixiang Wang, Weibin Liao, Xiaochen Zheng, Lifang Liang, Miguel O Bernabeu, Yasha Wang, Lequan Yu, Chengwei Pan, et al. Clinicrealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. *npj Digital Medicine*, 2026a.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Jiyao Wang, Junyi Gao, Liantao Ma, and Lequan Yu. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *Advances in Neural Information Processing Systems*, 38, 2026b.

Shihao Zou and Wei Wei. Corim: Conflict-driven risk minimization for dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37821–37830, 2026.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990, 2025.

A Benchmark Comparison Criteria

Table 1 adds three columns beyond coverage and evidence type. We record the criteria here so that the comparison is reproducible.

Average modalities per task. We use the contract-based definition in Section 3.1: a *modality family* contains inputs that can be decoded into the same normalized representation without changing their acquisition or clinical semantics. We count distinct modality families in each task instance and then average over tasks, rather than counting raw files or broad modality labels. Multiple inputs satisfying one contract count once. Benchmarks without a stable input-contract set are marked as not applicable.

Tools or perception models. We mark ✓ only when the benchmark itself ships an executable tool surface or modality-specific perception models that an agent can call at evaluation time. Direct LLM or VLM prompting and generic code execution inside a third-party agent framework do not qualify.

Licensed-clinician scrutiny. We mark ✓ only when licensed clinicians scrutinize *task configuration*: the queries, cases, options, or reference trajectories that define what the model is asked to do. We do not count clinician review of model outputs, expertise of the original source-dataset authors, unspecified "experts" or students, or review that does not state a professional license. Partial credit is used when licensed clinicians configure only a subset of tasks.

A.1 Per-Benchmark Counting

OmniMedVQA (Hu et al., 2024). Each item is a single-image VQA pair, so the average is 1.0. No tools are provided. QA stems and distractors are generated from existing dataset labels with GPT and a human double-check; licensed-clinician review of task configuration is not reported.

MedXpertQA (Zuo et al., 2025). The Text subset, containing 2,455 questions, is one modality family. The MM subset, containing 2,005 questions and 2,839 images, pairs clinical text with imaging, giving two modality families in a typical item. The task-weighted mean is

$$(2{,}455 \cdot 1 + 2{,}005 \cdot 2)/4{,}460 \approx 1.45,$$

reported as $\sim$1.5. No tools are provided. Medical experts with physician licenses review every question and option set.

ClinicRealm (Zhu et al., 2026a). The released prediction tasks use either unstructured notes or structured EHR, so the average is 1.0. The paper's multimodal fusion experiments are an analysis extension rather than the main task set. No agent tools or perception service are provided. Five clinicians later rated model reasoning on 100 cases; they did not configure the tasks.

X-PCR (Wang et al., 2026a). The corpus contains 177,868 VQA pairs over 26,415 images. The six-stage progressive-reasoning items are almost entirely single-modality, while cross-modal integration and 58 hospital cases use two to six modality families. Because the single-modality items dominate the average, we report $\sim$1.0. No tools are provided. Attending ophthalmologists authored the Stage-6 clinical-decision items and independently reviewed 20% of the automatically generated VQAs, so clinician scrutiny is partial.

ChestAgentBench (Fallahpour et al., 2025). Each query is built from a Eurorad case with a chest radiograph and clinical documentation, so the average is 2.0. The companion MedRAX agent provides CXR perception tools. The 2,500 six-choice questions are generated and filtered with GPT-4o; licensed clinicians do not scrutinize the resulting task configurations.

MedAgentBoard (Zhu et al., 2026b). The four task families use text, images, or structured EHR separately, so the average is 1.0. The benchmark evaluates existing agent frameworks and does not ship a tool or perception service. An expert panel of PhD and MD students rated workflow outputs; licensed clinicians did not configure the tasks.

CHI-Bench (Chen et al., 2026). Tasks are policy-rich operational workflows rather than diagnostic-modality integration, so modality counts are not applicable. The environment exposes 87 MCP tools over 20 healthcare applications. Outsourced licensed clinicians review and edit each surviving task candidate.

MedCTA (Ashraf et al., 2026). The release contains 107 tasks paired with 107 images, so the average is 1.0. Five executable tools are provided. Queries and trajectories are clinician-verified, but a professional license is not stated, so the clinician column is ✗ under our criterion.

MEDOMNI BENCH. Task configurations provide a mean of approximately 13 modality families per task over 34 supplied modality families. The benchmark provides the MP service that agents can call during evaluation. Licensed clinicians in the clinical consortium scrutinize task configuration, while the assurance mechanism reports its audits and human sign-off status separately.

B DATA AND TASK CONSTRUCTION

B.1 SOURCE DATASETS AND COHORT CONSTRUCTION

MEDOMNI BENCH draws from AI-READI, MIMIC-IV, and BDSP. The three sources were selected because their evidence surfaces are complementary: AI-READI supplies deep metabolic and retinal phenotyping, MIMIC-IV supplies longitudinal hospital data, and BDSP supplies dense overnight sleep physiology. Table 2 summarizes the role of each source. Counts in this appendix distinguish the source pool from the patients instantiated in benchmark tasks.

MIMIC-IV. The MIMIC pool is built after collecting the task requirements. Every proposed target, prediction boundary, and required modality family is converted into a demand record. Patients already bound to task labels and patients with indexed waveforms form the initial set; remaining

Table 1: Comparison of medical AI benchmarks. Counting criteria are detailed in Appendix A.

Benchmark	Paradigm	Mods	Avg./task	Evidence	Tools	Clin.
OmniMedVQA CVPR, 2024	VLM	12	1.0	Single-modal	✗	✗
MedXpertQA ICML, 2025	LLM/VLM	11	∼1.5	Multimodal	✗	✓
ClinicRealm npj Digital Medicine, 2026	LLM	2	1.0	Single-modal	✗	✗
X-PCR arXiv, 2026	VLM	6	∼1.0	Multimodal	✗	Partial
ChestAgentBench ICML, 2025	Agent	2	2.0	Multimodal	✓	✗
MedAgentBoard NeurIPS, 2026	Agent	3	1.0	Single-modal	✗	✗
CHI-Bench arXiv, 2026	Agent	—	—	—	✓	✓
MedCTA arXiv, 2026	Agent	16	1.0	Multimodal	✓	✗
MEDOMNI BENCH Ours	**Agent**	**34**	**∼13**	**Extensible**	✓	✓

Column attributes: Mods: modality families under MedOmni's criterion. Avg./task: mean number of those families in one task. Evidence: modalities combined within one task. Tools: whether the benchmark provides executable tools or modality-specific perception models. Clin.: whether licensed clinicians scrutinize *task configuration*. — indicates no stable diagnostic-modality set.

Table 2: Source data used to construct MEDOMNI BENCH. Release scale describes the upstream collections; benchmark pools are filtered before task-specific selection.

Source	Upstream scale	Clinical context	Principal evidence	Benchmark pool
AI-READI	2,280 participants; 358,999 files (3.82 TB)	Metabolic and retinal assessment	Retinal imaging, 12-lead ECG, clinical measurements, CGM, and wearable series	Flagship v3.0 pool followed by a fixed cohort filter
MIMIC-IV	364,627 patients; linked collections include 377,110 chest radiographs and ∼800,000 ECGs	Hospital, ED, and ICU	Structured EHR, chest radiographs, ECG, and bedside waveforms	Demand-driven pool of 550 patients
BDSP	90,166 patients; 119,234 sleep recordings across five sites	Overnight sleep medicine	PSG channels and linked clinical records	1,100 disease-stratified plus 300 random patients

patients are added to fill the sparsest demand cells. The resulting 550-patient pool is shared across MIMIC task types rather than redrawn for each task. A replay gate checks that every task and modality-family requirement remains feasible. The raw ECG mirror is incomplete relative to the upstream release, so ECG availability is treated as an acquisition limitation rather than as a cohort-selection decision.

AI-READI. The complete flagship v3.0 release is treated as the source pool. A fixed cohort filter is then applied before task-specific eligibility checks and patient selection. This separation prevents the source corpus from being conflated with the smaller set of patients instantiated in a benchmark campaign. AI-READI is cross-sectional, so its tasks assess status at the study visit rather than future onset.

BDSP. BDSP is scoped to site I0006 and to the raw BIDS layer of its PSG data. The standing pool contains 1,400 patients: 1,100 selected to broaden coverage across organ-system diagnoses and 300 selected randomly with a fixed seed, disjoint from the stratified set. The stratified component is intentionally enriched for comorbidity and is not prevalence representative. Prevalence-sensitive analyses therefore use the random component or report the sampling design explicitly.

B.2 Modality-Family Registry

A modality family is the unit of input compatibility. Two inputs belong to one modality family only when the same reader and encoder interface can interpret them without silently changing their clinical meaning. A modality-family contract records the accepted format, dimensionality, channel semantics, sampling rate, physical units, and permitted supply aliases. This is stricter than grouping data by a broad name such as ECG or retinal imaging.

The benchmark data pool supplies 34 modality families. Fifteen are physiological waveforms, six are numeric or wearable time series, nine are medical images or volumes, three are clinical tables, and one is a categorical body-position series. Table 3 lists the complete supplied registry, while Table 4 summarizes its distribution across sources. Only five modality families occur in more than one source, which makes the three data surfaces largely complementary.

Modality-family definitions are independent of encoder availability. This prevents the registry from hiding a data stream merely because no current model supports it. Conversely, an encoder is exposed to a task only when the corresponding modality family is present in that task's dataset and the modality-family contract is satisfied. Each generated task records the encoder-manifest fingerprint used during construction, making later registry changes detectable.

B.3 Clinical Task Collection

The MedOmni Task Proposal website is a guided instrument for the multidisciplinary clinical consortium. It presents more than 1,000 candidate labels together with dataset-specific evidence streams. Every candidate carries provenance and prevalence information where available. The website also states source-specific timing and circularity risks before a proposal is submitted.

The website operationalizes a six-question workflow. QA selects the dataset and therefore the available population and evidence catalog. QB records the task name and whether evidence should be clinically curated or broadly included. QC fixes the prediction time. QD and QE assign candidate labels and evidence streams to the Predict and Input trays. QF asks the proposer to state the resulting clinical question in one sentence before submission. Dataset summaries, collection timelines, and circularity warnings remain visible during this process.

The timing choices reflect the structure of each source. AI-READI exposes the study visit because the release is cross-sectional. MIMIC-IV exposes encounter-relative boundaries for emergency, admission, and hospital events. BDSP uses the end of the overnight recording. Candidate labels and evidence streams can be assigned to either the prediction or input side, allowing the same measured quantity to serve different roles in different tasks.

The website exports directly to the task builder's intake schema, avoiding manual transcription between clinical proposal and implementation. The consultation produced 34 clinical proposals: 28 were implemented as clinician-proposed task types and six were filtered by pre-construction assurance checks. The complete evaluation suite contains 30 task types. The frozen proposal records retain the task statement, selected targets and evidence, prediction time, and proposal provenance.

B.4 Task Configuration and Benchmark Composition

Task construction proceeds through four stages:

$$\text{proposal} \rightarrow \text{frozen specification} \rightarrow \text{feasibility checks} \rightarrow \text{patient-level instances}.$$

The frozen specification declares the source dataset, target derivations, prediction boundary, admissible modality families, cohort rules, metrics, and verifier. Feasibility checks confirm that labels and required evidence exist before any instance is emitted. Each emitted instance binds this configuration to one patient while retaining the specification and data fingerprints needed for reconstruction.

Table 3: Modality families used in MEDOMNI BENCH.

Modality family	Evidence form	Protocol summary	M	B	A
Arterial blood pressure	Waveform	WFDB, 125 Hz, mmHg	✓		
Diagnostic ECG	Waveform	WFDB, 12 leads, 500 Hz, mV	✓		✓
Single-lead ECG	Waveform	EDF, 200 Hz, μV		✓	
Monitor ECG	Waveform	WFDB, nominally 3 to 5 leads, 250 Hz	✓		
Sleep EEG	Waveform	EDF, 6 channels, 200 Hz, μV		✓	
Chin EMG	Waveform	EDF, 200 Hz, μV		✓	
Limb EMG	Waveform	EDF, 200 Hz, μV		✓	
EOG	Waveform	EDF, 2 channels, 200 Hz, μV		✓	
PPG waveform	Waveform	EDF or WFDB, uncalibrated, variable rate	✓	✓	
Nonarterial pressure	Waveform	WFDB, 125 Hz, mmHg	✓		
Cardiorespiratory PSG	Waveform	EDF, 4 channels, 200 Hz, mixed units		✓	
Respiratory effort	Waveform	EDF, 2 RIP channels, 50 Hz		✓	
Respiratory flow	Waveform	EDF, variable rate		✓	
Respiratory impedance	Waveform	WFDB, 62.5 Hz, ohm	✓		
Snore audio	Waveform	EDF acoustic signal, variable rate		✓	
Continuous glucose monitor	Numeric series	OMH, 5-minute interval, mg/dL			✓
Pulse rate	Numeric series	EDF, 10 Hz, beats/min		✓	
Oxygen saturation	Numeric series	EDF, 10 Hz, percent		✓	
Wearable heart rate	Numeric series	OMH, 1-minute interval, beats/min			✓
Wearable respiratory rate	Numeric series	OMH, 1-minute interval, breaths/min			✓
Wearable oxygen saturation	Numeric series	OMH, 1-minute interval, percent			✓
Chest radiograph	2D image	Image representation of CXR	✓		
Fundus autofluorescence	2D image	DICOM, monochrome image			✓
Fundus photography	2D image	DICOM color fundus image			✓
OCTA en-face slab	2D image	DICOM, monochrome en-face image			✓
FLIO	3D volume	DICOM, $256\times256\times1024$ bins			✓
OCTA flow cube	3D volume	DICOM flow-cube volume			✓
OCT B-scan set	3D volume	DICOM radial and circular B-scan set			✓
OCT raster volume	3D volume	DICOM raster volume			✓
Retinal height map	3D volume	DICOM floating-point surface maps			✓
Clinical measurement	Clinical table	Irregular measurements with values and units	✓		✓
Clinical record	Clinical table	Structured longitudinal records	✓	✓	✓
Numeric observation	Clinical table	Table or OMH numeric observations	✓		✓
Body position	Categorical series	EDF, 10 Hz position code		✓	

Rows are grouped and colored by evidence form. M, B, and A denote MIMIC-IV, BDSP, and AI-READI; a check mark indicates that the modality family is supplied by that source. Protocol summaries describe normalized modality-family inputs rather than encoder-specific preprocessing.

Table 4: Cross-source distribution of the 34 supplied modality families.

Supply	Count	Modality families
All three	1	Clinical record table
Two sources	4	12-lead ECG, clinical measurement, numeric observation, PPG waveform
MIMIC-IV only	5	Arterial pressure, nonarterial pressure, chest radiograph, monitor ECG, respiratory impedance
AI-READI only	12	CGM, fundus photography, FAF, OCT and OCTA representations, FLIO, retinal height maps, and wearable summaries
BDSP only	12	EEG, EOG, chin and limb EMG, single-lead ECG, respiratory signals, SpO_2, pulse, snore, and body position

Table 5: Information collected for each clinical task proposal.

Field	Purpose
Dataset	Selects the source population and available evidence catalog
Task name and mode	Names the task and records whether evidence is curated
Prediction time	Establishes the temporal boundary for admissible evidence
Targets	Defines what the agent must predict or assess
Evidence streams	Defines what the agent may inspect
Clinical statement	Records the intended clinical question in free text

Leakage controls. Temporal filtering removes events after the declared prediction boundary. Target-specific exclusions address circularity that cannot be solved by time alone. For example, the CKD task excludes the renal measurements used to define the target. Sleep metrics derived from scoring the same PSG night cannot be used as evidence when that metric is the prediction target. Medication fields that nearly determine an AI-READI diabetes group are likewise excluded for that target. Free clinical text remains outside the supplied modality families until a rule can prevent target diagnoses from appearing directly in prose. These construction rules are complemented by the assurance analysis in the main text and its appendix.

Suite composition. The current suite contains 30 task types and 600 patient-level instances, with 20 instances per type. After excluding four held and two degenerate continuous targets, the evaluation scores 59 target slots. A slot denotes one predicted variable attached to one task type, so the same clinical variable in two task types occupies two slots. Table 6 summarizes the distribution across sources.

Assurance filtering. Six proposals were filtered before task construction for the following reasons.

Post-ROSC ECG for prognostication. ROSC time is not directly recorded, so ECG timing cannot be aligned reliably. ICU and procedure engines were unmapped, and discharge-text staging was deferred.

Multimodal prediction of ED reattendance. The proposal requires 22 streams. Several staging engines were deferred or unmapped, and derivation of the 24, 48, and 72-hour return targets was unfinished.

Potassium level from 12-lead ECG. The task requires dedicated ECG-to-blood-draw time alignment and an MAE-based regression scoring path.

Hospital outcomes from demographics, diagnoses, and laboratory results. Diagnosis-history staging and reusable hospital-outcome derivations were unfinished.

Hospital outcomes from discharge diagnoses. Discharge diagnoses are assigned after in-hospital death, ICU admission, and length of stay, making these inputs scientifically circular for the proposed outcomes. Approximately 190 diagnosis categories were also unmapped.

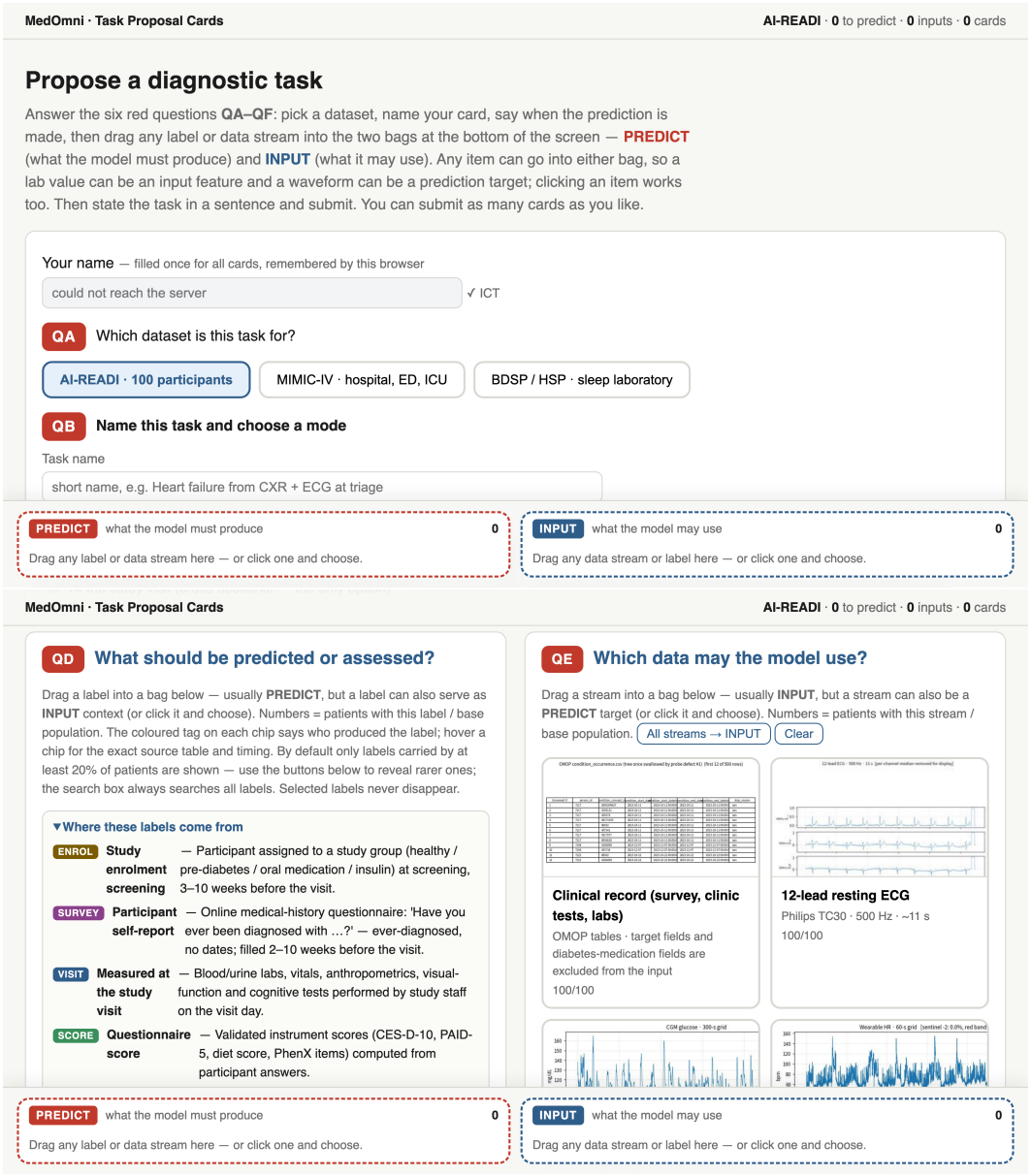

Figure 8: The MEDOMNI Task Proposal website. Top: the proposer selects a source dataset, names the task, and chooses between targeted evidence selection and an all-in mode. Bottom: the website pairs a provenance-aware target catalog with dataset-specific evidence cards; fixed PREDICT and INPUT trays make the intended role of every selected item explicit. The identity field is anonymized for submission.

Hospital outcomes from radiology reports. Free-text radiology staging was deliberately deferred because reports can explicitly reveal the target. Using CheXpert labels as model inputs also required an unresolved policy decision.

The targets cover cardiovascular and renal disease, hospital outcomes, metabolic status, retinal and visual phenotypes, cognition, neuropathy, psychiatric conditions, demographics, mortality, and sleep physiology. AI-READI and BDSP expose nearly uniform modality-family sets within each source. MIMIC-IV varies more because several waveform modality families are available for only a minority of patients. The modality-family count for a MIMIC task type is therefore the union over its sampled patients, not a different evidence policy for each clinical question.

Table 6: Composition of the current task suite.

Source	Types	Instances	Modality families per type
AI-READI	15	300	16
BDSP	6	120	15
MIMIC-IV	9	180	5–10
Total	30	600	$\sim$13 average

Some task types intentionally overlap in their target set. In the current configuration, two AI-READI neuropathy types share left and right foot-sensation targets, and two BDSP sleep types share six sleep-metric targets. They preserve distinct clinical proposal provenance and can differ in selected evidence, but analyses that count unique clinical questions report this overlap separately.

B.5 Evaluation Criteria

Decision correctness. Decision correctness is evaluated independently for each of the 59 scored target slots using the output-specific skills in Table 8. Predictions are first validated at the instance level and then aggregated by target and task type. This order prevents one valid target from masking an unreadable prediction for another. Failed or unreadable trials remain in the denominator with the target-specific no-information prediction, while infrastructure failures are rerun.

Behavioral reasonableness. Decision correctness alone cannot determine whether an agent integrated heterogeneous evidence in a clinically reasonable way. The execution record therefore supports B1 completion, B2 modality choice, and B3 first answer from file-access and MP service traces. Provider-reported cost is analyzed separately rather than folded into the MedOmni score or the behavioral criteria.

B.6 Harbor Execution

Trial packaging. One task instance maps to one self-contained Harbor trial. Harbor provides the general agent-execution substrate associated with Terminal-Bench (Merrill et al., 2026). The trial contains the instruction, patient evidence, available tool manifest, output contract, and verifier. Infrastructure defaults, agent configuration, and campaign configuration are layered separately, so the same task can be paired with different agent scaffolds and model backbones without changing its evidence or scoring contract. Trials run in isolated containers and may execute in parallel.

Scoring is performed outside Harbor by the same scoring implementation used for other runners. Harbor's native reward is retained only as trace metadata. Completed jobs are imported into a ledger with the requested model, served model, task fingerprint, configuration, result, and failure state. Results from different runners are not pooled unless a same-model parity comparison has been completed; otherwise, runner identity remains an explicit evaluation axis.

Table 7: Complete clinical proposal census.

Audit passed (28 proposals)	
Task proposal	**Proposed output**
AI-READI	15 proposals
Cancer history from metabolic and lifestyle data	Cancer history
Cataract status from ocular imaging and clinical data	Cataract history
Cognitive function from retinal and metabolic data	MoCA total score
Cognitive score from metabolic and behavioral data	MoCA total score and subscores
CRP and WBC from lifestyle parameters	hs-CRP and white blood cell count
CRP and WBC from biopsychosocial, socioeconomic, and demographic data	hs-CRP and white blood cell count
CRP and WBC from metabolic syndrome parameters	hs-CRP and white blood cell count
Depressive symptom severity from lifestyle and wearable data	CES-D-10 depression score
Diabetes diagnosis and severity from biopsychosocial data	Diabetes status and severity indicators
Diabetes status from retinal imaging	Healthy, pre-diabetic, or type 2 diabetes status
Diabetic peripheral neuropathy from metabolic markers	Bilateral monofilament foot sensation
Dyslipidemia diagnosis and severity from biopsychosocial data	Total, LDL, and HDL cholesterol; triglycerides
Foot sensation from metabolic complication phenotype	Bilateral monofilament foot sensation
Prior hypertension from cardiovascular and end-organ phenotype	History of high blood pressure
Visual function from biopsychosocial, socioeconomic, and demographic data	Visual acuity, contrast sensitivity, refraction, and retinal phenotypes
BDSP	6 proposals
Arousal parameters from overnight physiology	Arousal index and annotations
Demographic data from multimodal sleep physiology	Sex, age, race, and mortality status
Depressive disorders from overnight sleep physiology	Depressive disorders
Heart failure from overnight sleep physiology	Heart failure
Snoring-derived parameters using demographic and physiological data	OSA severity and sleep-architecture metrics
Snoring-derived metrics from sleep physiology	OSA severity and sleep-architecture metrics
MIMIC-IV	7 proposals
Acute renal failure from early hospital data	Acute and unspecified renal failure
Coagulation and hemorrhagic disorder recognition	Coagulation and hemorrhagic disorders
Early recognition of heart failure	Heart failure
ICU stay from radiology reports	ICU stay
Hospital outcomes from microbiology cultures	Mortality, ICU admission, readmission, ED revisit, and length of stay
Septicemia from early hospital data	Septicemia
Cerebral infarction at ED triage	Cerebral infarction

Filtered by assurance (6 proposals)	
Task proposal	**Proposed output**
Post-ROSC ECG for prognostication	Post-ROSC prognosis
Multimodal prediction of ED reattendance	ED return within 24, 48, and 72 hours
Potassium level from 12-lead ECG	Serum potassium level and normal range
Hospital outcomes from demographics, diagnoses, and laboratory results	Mortality, ICU admission, readmission, ED revisit, and length of stay
Hospital outcomes from discharge diagnoses	Mortality, ICU admission, readmission, ED revisit, and length of stay
Hospital outcomes from radiology reports	Mortality, ICU admission, readmission, ED revisit, and length of stay

Based on the clinician intake records. Task titles and outputs are shortened only for table layout. Engineering-only smoke-test configurations are excluded.

Table 8: Evaluation criteria for MEDOMNI BENCH. Full definitions and aggregation rules appear in Appendix E.4.

Decision correctness		
Item	**Metric**	**Reporting rule**
Binary target	Rescaled AUROC skill: $\max\{0, 2\,\mathrm{AUROC} - 1\}$	Score from chance or worse (0) to perfect (1); ties receive half credit.
Ordinal target	Quadratic weighted kappa skill: $\max\{0, \kappa_w\}$	Round predictions to the nearest class before computing κ_w.
Continuous target	MAE skill: $\min\{1, \max\{0, 1 - \mathrm{MAE}/\mathrm{MAE}_0\}\}$	Define MAE_0 using the population-pool median, not the scored cohort.
Aggregation	Mean over scorable targets, then equal mean over task types	Keep multiple outputs as separate target slots; report source-level scores over their task types.
Failures and unscorable targets	No-information prediction or exclusion under declared rules	Keep failed or unreadable trials in place; rerun infrastructure failures rather than imputing them.

Behavioral reasonableness		
Item	**Criterion**	**Reporting rule**
B1 Completion	Valid answer returned within the 20-minute budget	Percentage of all trials.
B2 Modality choice	Per-trial F1 between consulted and clinician-marked relevant modality families	Average over applicable trials after restricting both sets to staged families; clinical records and structured measurements do not count as modalities.
B3 First answer	Whether the first answer write precedes the first modality evidence	Percentage of trials that consult a modality and write an answer.
Cost	Provider-reported API cost per trial	Report separately rather than combining it with A or B1–B3.

C Modality Perception Service Implementation

C.1 Registry Coverage

The current MP service registry contains 44 runnable encoders, providing executable coverage for 30 of the benchmark's 34 modality families (Table 9). Of these encoders, 37 return labels or text only, five return labels or text together with an embedding, and two return embeddings only. The registry combines pretrained medical models with deterministic rule extractors; none is trained or fine-tuned on MedOmni data. The interface does not privilege either type: compatibility is determined by the modality family contract, and the response is validated against the encoder card.

Four benchmark modality families currently have no runnable encoder. Clinical record tables and numeric observations are read directly by the agent. Arterial-pressure and wearable respiratory-rate data also remain directly accessible because their candidate rule extractors do not currently pass registry acceptance. Keeping these families visible separates the benchmark's evidence surface from the MP service's current executable coverage.

C.2 Encoder Encapsulation and Cards

Each encoder is encapsulated in a separate Docker image containing its software dependencies and, when applicable, its model weights. Isolated environments allow research code with incompatible dependency trees to coexist and improve computational reproducibility (Boettiger, 2015). A shared server and modality family reader are included in every image, while an encoder-specific adapter implements rules or model loading, preprocessing, and inference.

Each encoder card records the encoder identifier, compatible modality family, intended use, known limitations, accepted options, and output API. For label-producing encoders, the card defines the complete label vocabulary and score scale; for embedding encoders, it specifies the vector dimension. The card also records the upstream source and revision, software license, container image reference, reliability evidence, and human acceptance status. A model-weight hash is required only for an encoder that uses model weights.

Modality family decoding is separated from encoder-specific processing. The shared reader interprets serialized clinical data and enforces the modality family API; the adapter applies transformations required by a learned model or computes the declared rule-based evidence. This separation reduces duplicated parsing logic and ensures that compatible encoders receive a consistent representation.

C.3 Container API

Every encoder container exposes the same three internal HTTP endpoints: `GET /healthz`, `GET /card`, and `POST /encode`. The health endpoint reports readiness, the card endpoint returns the structured encoder metadata, and the encode endpoint accepts a modality family identifier, a path relative to a read-only data root, and optional encoder-specific parameters. The response contains the encoder and modality family identifiers, a health indicator, one or more output blocks, runtime metadata, and notes describing material preprocessing decisions. The common response schema supports labels, events, text, and embeddings.

The input path is resolved beneath the configured data root, which is mounted read-only. The container rejects traversal, symbolic-link escape, modality family mismatch, and undeclared request fields. Inputs that are validly addressed but clinically unusable produce explicit degraded evidence rather than an apparently successful output.

C.4 Agent-Facing Routing

The MP service hides container lifecycle management behind one gateway. At startup, encoder cards are loaded into a searchable catalog. Discovery can filter by modality family, output type, label, or free-text query and does not start a container. Only an encoding request reaches the runtime pool: the gateway validates the requested path, selects the named compatible encoder, reuses a warm instance when available, or starts its container on demand.

Table 9: Encoder atlas used throughout this paper. Modality-family display names and evidence-form groups match Table 3.

Evidence form	Modality family	Runnable MP encoders
Waveform	Arterial blood pressure	—
Waveform	Diagnostic ECG	`ecgfounder`
Waveform	Single-lead ECG	`ecgfounder`, `neurokit2`
Waveform	Monitor ECG	`ecgfounder`
Waveform	Sleep EEG	`yasa`
Waveform	Chin EMG	`rule-amplitude`
Waveform	Limb EMG	`rule-amplitude`
Waveform	EOG	`wav2sleep-eog`
Waveform	PPG waveform	`bayesbeat`, `neurokit2`
Waveform	Nonarterial pressure	`rule-summary`
Waveform	Cardiorespiratory PSG	`wav2sleep`
Waveform	Respiratory effort	`neurokit2`
Waveform	Respiratory flow	`rule-amplitude`
Waveform	Respiratory impedance	`rule-amplitude`
Waveform	Snore audio	`rule-amplitude`
Numeric series	Continuous glucose monitor	`pyagata`
Numeric series	Pulse rate	`rule-summary`
Numeric series	Oxygen saturation	`pobm`
Numeric series	Wearable heart rate	`rule-summary`
Numeric series	Wearable respiratory rate	—
Numeric series	Wearable oxygen saturation	`rule-summary`
2D image	Chest radiograph	`biomedclip`, `chexzero`, `torchxrayvision`
2D image	Fundus autofluorescence	`retinograd`, `rule-intensity`
2D image	Fundus photography	`dr-grade`, `retfound`, `retfound-dr`, `retfound-jsiec`
2D image	OCTA en-face slab	`rule-intensity`, `vessel-density`
3D volume	FLIO	`rule-photons`
3D volume	OCTA flow cube	`flow-metrics`, `rule-intensity`
3D volume	OCT B-scan set	`mirage-onh`, `rule-intensity`
3D volume	OCT raster volume	`kdoct`, `mirage`, `rule-intensity`
3D volume	Retinal height map	`layer-thickness`, `rule-surfaces`
Clinical table	Clinical measurement	`labs-summary`
Clinical table	Clinical record	—
Clinical table	Numeric observation	—
Categorical series	Body position	`rule-occupancy`

The paper uses these concise display aliases; the canonical modality-family contract identifiers are longer and unique. A dash marks a modality family that remains directly available to the agent but has no runnable encoder. The atlas contains 44 runnable encoders and omits encoders not retained in the runnable registry.

Operators configure the maximum number of active containers and per-encoder concurrency. When the pool reaches capacity, the gateway removes the least recently used idle container before admitting another. Requests and responses are logged at the service boundary, and repeatable evidence can be cached using fingerprints over the encoder, runtime image, input, and options. These controls are intentionally absent from the agent interface.

C.5 Validation and Publication

Registry tooling validates card schemas and modality family compatibility, builds each isolated image, compares responses with frozen expected outputs, and exercises published modality family samples without network access. A frozen expectation is a regression baseline, not clinical validation; reliability and acceptance remain separate card fields. Image publication is gated on these checks, pinned sources and applicable weights, license review, and assurance sign-off.

D ASSURANCE MECHANISM DETAILS

D.1 ASSURANCE MATRIX SPECIFICATION

Each *audit unit* is an asset class × meta-criterion × auditor combination. Table 10 assigns 29 of 60 positions; the remaining 31 are Not applicable. The 67 *atomic checks* below have stable identifiers; records distinguish check revisions, auditors, and versioned asset instances within each unit's declared inventory. COMP, CONF, ISOL, and CLIN denote the four meta-criteria in table order. Assigned auditors assess each check within that inventory using evidence appropriate to their role. New requirements receive new identifiers or check revisions.

Table 10: Auditor assignments. M: machine; L: LLM, advisory; H: human sign-off (technical/clinical expertise as required). Absent roles are Not applicable.

Asset class	Completeness	Conformance	Isolation	Clinical validity
Data	M	M	M	L, H
Modality-family contracts	M	M	M, L	L, H
MP service	M	M	M, L	L, H
Tasks	M	M	M, L	L, H
Evaluation runs	M	M	M, L	L, H

Data (D). `D-COMP-01`: Required source files and manifests present; `D-COMP-02`: Version, provenance, license, and cohort counts recorded; `D-COMP-03`: Required evidence streams and patient-level availability records present.

`D-CONF-01`: Declared schema, types, units, and hashes satisfied; `D-CONF-02`: Cohort and modality-availability statistics reproduce from manifests; `D-CONF-03`: Source-specific filtering reproduces declared pool.

`D-ISOL-01`: Direct identifiers and prohibited metadata absent; `D-ISOL-02`: Dataset, patient, and evaluation-split boundaries preserved; `D-ISOL-03`: Access restricted to declared roots.

`D-CLIN-01`: Population and acquisition setting support intended benchmark use; `D-CLIN-02`: Missingness, coverage, and known biases clinically characterized; `D-CLIN-03`: Preprocessing preserves clinical meaning.

Modality-family contracts (F). `F-COMP-01`: Applicable format, dimensionality, channel semantics, sampling rate, physical units, and accepted aliases declared; `F-COMP-02`: Representative validation samples available and fingerprinted; `F-COMP-03`: Reader and normalization requirements documented.

`F-CONF-01`: Pinned samples exactly match declared format and dimensionality; `F-CONF-02`: Channels, roles, sampling rates, units, and image encoding match contract; `F-CONF-03`: Readers preserve declared timing and measurement semantics.

`F-ISOL-01`: Readers expose only declared payloads and metadata; `F-ISOL-02`: File names, paths, aliases, and metadata reveal no targets or gold labels; `F-ISOL-03`: Runtime accommodations never silently broaden contracts.

`F-CLIN-01`: Channel, unit, and acquisition semantics match clinical practice; `F-CLIN-02`: Clinically distinct acquisition protocols remain separate families; `F-CLIN-03`: Normalization preserves clinically necessary information.

MP service (P). `P-COMP-01`: Every registered learned or rule-based encoder has card, source revision, and runnable implementation; `P-COMP-02`: Runtime images, rules, applicable model weights, and reliability evidence present; `P-COMP-03`: Endpoints, routing metadata, health checks, and call logging available.

`P-CONF-01`: Behavior matches card, compatible families, and output contract; `P-CONF-02`: Only compatible encoders receive inputs; `P-CONF-03`: Requests, responses, and logs satisfy declared schemas; `P-CONF-04`: Reliability evidence meets declared acceptance requirements.

`P-ISOL-01`: Containers access only permitted read-only roots; `P-ISOL-02`: Reject path traversal and access to test, gold, or undeclared external data; `P-ISOL-03`: Training provenance excludes prohibited evaluation-data overlap; `P-ISOL-04`: Cross-encoder container and dependency isolation preserved.

P-CLIN-01: Intended use, outputs, limitations, and failure modes clinically meaningful; P-CLIN-02: Reliability evidence supports claimed output interpretation; P-CLIN-03: Labels, scores, events, and text preserve declared clinical semantics.

Tasks (T). T-COMP-01: Proposal, frozen specification, target derivation, prediction boundary, evidence policy, cohort rules, metric, and verifier present; T-COMP-02: Every instance contains required evidence and task artifacts; T-COMP-03: Task and instance fingerprints recorded.

T-CONF-01: Frozen specifications agree with proposal records and manifests; T-CONF-02: Rebuilding reproduces artifacts and fingerprints; T-CONF-03: Patient instances satisfy cohort, evidence, target, and output contracts; T-CONF-04: Verifiers implement declared metric and failure policy.

T-ISOL-01: Evidence after prediction boundary excluded; T-ISOL-02: Target-defining, circular, or directly revealing evidence excluded; T-ISOL-03: Agents access neither gold nor verifier internals; T-ISOL-04: Patient evidence never crosses instance boundaries.

T-CLIN-01: Question, target, population, and prediction time form meaningful clinical task; T-CLIN-02: Evidence clinically relevant and admissible at decision time; T-CLIN-03: Target derivation and exclusions support intended clinical interpretation; T-CLIN-04: Clinician review bound to task fingerprint.

Evaluation runs (R). R-COMP-01: Configuration, requested and served model identities, predictions, traces, verifier outputs, and ledger records present; R-COMP-02: Runtime, token, cost, and failure records retained; R-COMP-03: Every reported perception call in service log.

R-CONF-01: Run uses pinned task, service, agent, and model versions; R-CONF-02: Verifier replay reproduces score; R-CONF-03: Predictions satisfy output contract; R-CONF-04: Reported tool use agrees with execution traces and service logs.

R-ISOL-01: Agent and perception calls never access prohibited test, gold, or external patient data; R-ISOL-02: Trial execution and external scoring separated; R-ISOL-03: Parallel trials share no patient evidence, state, or outputs.

R-CLIN-01: Evidence-seeking clinically reasonable; R-CLIN-02: No unjustified systematic neglect of relevant available evidence; R-CLIN-03: Refusals, malformed outputs, timeouts, and unscorable outcomes explicit; R-CLIN-04: Human review bound to run and task fingerprints.

Outcomes. The three outcomes are *Pass (with ratio)*, *Blocked*, and *Not applicable*. Without a blocking finding, Pass reports passing checks divided by assessed applicable checks; non-critical non-passing checks remain *warnings*. Benchmark-compromising findings, including leakage, override the ratio as Blocked; critical conditions are specified before assessment. Not applicable means no checks assigned by design, excluded from denominators. Report coverage as assessed checks divided by assigned applicable checks, alongside pending identifiers. Missing or stale evidence and pending review do not count as assessed; they remain completion metadata, never another outcome or Not applicable. Zero assessed checks yield no evaluated outcome or ratio. Ratios aggregate check executions within the declared inventory. Figure 4 illustrates these outcomes using synthetic values.

We applied the assurance matrix iteratively to the reported benchmark release. Findings were used to correct assets and specifications, after which the affected checks were rerun against updated fingerprints. The final snapshot contains no Blocked audit units and passes 95.3% of assessed applicable checks. The remaining non-passing checks are retained as documented warnings rather than being counted as passes.

D.2 Worked Audit-Unit Example

For one hypothetical contract, the unit *Modality-family contracts* $\times$ *Completeness* $\times$ *M* records F-COMP-01: declarations complete; F-COMP-02: samples fingerprinted; F-COMP-03: non-critical normalization documentation missing despite a pinned, validated reader. The first two pass; the third is an assessed warning. The unit reports *Pass (2/3; 66.7%)* with coverage $3/3$. Gold-data access under P-ISOL-02 instead produces Blocked; Data $\times$ Isolation $\times$ H is Not applicable by design.

D.3 Sign-off and Iterative Refinement Details

The protocol requires fingerprinted records containing check ID and revision, unit, asset instance and version, auditor, time, evidence digests, finding and rationale, and follow-up owner. Machine, LLM, technical, and clinical auditors must use one version-bound submission mechanism: authenticated identity, record digest, reviewed fingerprints, and disposition. Machine records identify the implementation and environment; LLM records identify model and prompt versions; human sign-off specifies expertise and review scope. LLM advice cannot replace clinical sign-off. Service and harness evidence retains distinct provenance from self-report; restricted patient content stays in authorized environments.

Before sign-off, compare fingerprinted inputs with current versions. Changed covered dependencies or check requirements require reassessment; technical review must identify changes outside fingerprint scope and extend it where needed. Correct assets, regenerate evidence, repeat affected checks, and renew approvals. Completion requires full mandated coverage, no unresolved blocking finding, documented warning dispositions, and current technical and clinical sign-offs. Construction-time consultation supplies supporting evidence; formal sign-off requires an authenticated verdict naming its scope and audited fingerprints.

E Implementation details

E.1 Agents and harness

model	API identifier	max turns
GPT-6 Astra	`openai/gpt-6-astra`	300
DeepSeek-V4 Flash	`deepseek/deepseek-v4-flash`	300
GPT-6 Sol	`openai/gpt-6-sol`	300
MiMo-v2.6 Pro	`xiaomi_mimo/mimo-v2.6-pro`	300
GPT-6 Luna	`openai/gpt-6-luna`	300
GPT-5 nano	`openai/gpt-5-nano`	300

Table 11: Evaluated agents. All run in the `terminus-2` scaffold on Harbor.

Each trial is one patient instance in its own container (4 CPUs, 8 GB RAM, 16 GB disk, no GPU) on a closed network that reaches only the MP service. The agent has 20 minutes; the verifier then reads the submission. Cost is the provider-reported API cost of the trial. Every configuration ran once. The "MP off" arms (Section F) are the same instances with the MP service removed from the container and from the instruction.

E.2 Modality perception service

The MP service contains 44 runnable pretrained or rule-based encoders covering 30 of the benchmark's 34 modality families. Each task lists in `encoder_manifest.md` only the encoders for the families staged for that patient. The agent can list encoders (`GET /encoders`), read an encoder's card with its inputs, limits and options (`GET /cards/{id}`), and run one encoder on one staged file (`POST /encode`). An encoder returns named labels, text, an embedding, or a combination, with a health flag; an MP service call counts as a *usable output* when it returns healthy. Using the MP service is optional and never required by the instruction.

E.3 Benchmark split

The split has 30 task types (15 AI-READI, 6 BDSP, 9 MIMIC-IV), 20 patients each, i.e. 600 instances per agent. Their targets are binary (scored by AUROC), ordinal (quadratic weighted κ) or continuous (MAE skill); 4 continuous targets are held and 2 are degenerate (Section E.4), leaving 59 scored target slots. Table 12 lists the task types.

abbreviation	task type	source
A-CAN	aireadi_cancer-history-metabolic	AI-READI
A-CAT	aireadi_cataract-ocular	AI-READI
A-COG-M	aireadi_cognitive-score-metabolic	AI-READI
A-COG-R	aireadi_cognitive-function-retinal	AI-READI
A-CRP-1	aireadi_crp-white-cell-count	AI-READI
A-CRP-2	aireadi_crp-white-cell-count-2	AI-READI
A-CRP-L	aireadi_crp-wcc-lifestyle	AI-READI
A-DEP	aireadi_depression-lifestyle	AI-READI
A-DM-B	aireadi_diabetes-diagnosis-biopsycho	AI-READI
A-DM-R	aireadi_diabetes-retinal	AI-READI
A-DPN	aireadi_dpn-neuropathy-metabolic	AI-READI
A-FOOT	aireadi_foot-sensation-metabolic	AI-READI
A-HTN	aireadi_prior-hypertension-cardiovas	AI-READI
A-LIP	aireadi_dyslipidaemia-diagnosis-biop	AI-READI
A-VIS	aireadi_visual-function-biopsychosoc	AI-READI
B-ARO	bdsp_arousal	BDSP
B-DEP	bdsp_depression-sleep	BDSP
B-EEG	bdsp_demographic-eeg	BDSP
B-HF	bdsp_hf-sleep	BDSP
B-OSA	bdsp_snoring-metrics	BDSP
B-OSA-D	bdsp_snoring-demographic	BDSP
M-AKI	mimic_acute-renal-hospital	MIMIC-IV
M-CKD	mimic_ckd-staging	MIMIC-IV
M-COAG	mimic_coag-hemorrhagic	MIMIC-IV
M-HF	mimic_hf	MIMIC-IV
M-HF-MM	mimic_heart-failure-mm	MIMIC-IV
M-ICU	mimic_icu-stay-radiology	MIMIC-IV
M-MORT	mimic_in-hospital-death-icu	MIMIC-IV
M-SEP	mimic_sepsis-hospital	MIMIC-IV
M-STR	mimic_stroke-infarct-ed	MIMIC-IV

Table 12: Task types and their abbreviations in the per-task radar.

E.4 Evaluation protocol

Per-target skill. For target k of task type t, with gold y_i and prediction $\hat{y}_i$ for patients $i \in \mathcal{P}_t$, the skill score is bounded to $[0, 1]$ (0 = chance or worse, 1 = perfect):

$$\text{binary:} \quad s_k = \max\{0,\, 2\,\mathrm{AUROC}_k - 1\}, \tag{1}$$

$$\mathrm{AUROC}_k = \frac{1}{n_+ n_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \Big(\mathbb{1}[\hat{y}_i > \hat{y}_j] + \tfrac{1}{2}\mathbb{1}[\hat{y}_i = \hat{y}_j] \Big), \tag{2}$$

$$\text{ordinal:} \quad s_k = \max\{0,\, \kappa_k^w\}, \qquad \kappa_k^w = 1 - \frac{\sum_{a,b} w_{ab} O_{ab}}{\sum_{a,b} w_{ab} E_{ab}}, \quad w_{ab} = \frac{(a-b)^2}{(K-1)^2}, \tag{3}$$

$$\text{continuous:} \quad s_k = \min\Big\{1, \max\Big\{0,\, 1 - \frac{\mathrm{MAE}_k}{\mathrm{MAE}_k^0}\Big\}\Big\}, \qquad \mathrm{MAE}_k^0 = \frac{1}{|\mathcal{P}_t|} \sum_i |y_i - m_k|. \tag{4}$$

Binary predictions are clipped to $[0, 1]$. For κ^w, predictions are rounded to the nearest class; O is the observed and E the expected (outer product of marginals) confusion matrix over the K classes spanned by the gold labels and predictions. m_k is the median of target k in the population pool, never in the scored cohort.

Scorable targets. A target is not scored for a patient without a gold value, nor at all when its gold labels have one class, when it is continuous and one value covers $\geq$80% of the pool (2 targets), or when it is held pending a contract query (4 autorefraction targets whose unit and sign convention the instruction leaves open). These target-level conditions do not exclude any task type, leaving $|\mathcal{T}| = 30$ task types in the overall evaluation.

Failures. A trial that times out, or returns no readable answer for a target, keeps its place with the no-information prediction: 0.5 (binary), $\mathrm{round}(m_k)$ (ordinal) or m_k (continuous). Infrastructure failures are resumed and never imputed.

Aggregation and uncertainty. $S_t = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} s_k$ and MedOmni $= \frac{100}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} S_t$; per-source scores average the task types of that source. Because scores below chance count as 0, a model

answering at random scores about 10 rather than 0. The 95% CI is the 2.5th–97.5th percentile over $B = 1000$ two-level bootstrap resamples: draw $|\mathcal{T}|$ task types with replacement, then, within each, the patients with replacement (stratified by the class of the task's primary target), and recompute the score. A paired difference between two configurations applies the same resamples to both ($B = 2000$); it is significant when its CI excludes 0.

Behaviour. Agent steps are classified from keystrokes and the MP service's call log and file-access witness (Section G.1). *B1 completion* is the share of trials that return a valid answer within budget. *B2 modality choice*: for trial j of task t, C_j is the set of modality families the agent consulted (raw file access or MP service call) and H_t the set the task's proposer listed as relevant inputs, both restricted to the families staged for that patient; the clinical record table and structured measurements are not counted as modalities. Then $\mathrm{F1}_j = 2|C_j \cap H_t|/(|C_j| + |H_t|)$, averaged over trials with $H_t \neq \emptyset$. *B3 first answer* is, among trials that consult any modality and write an answer, the share whose first answer write precedes the first modality evidence.

E.5 Task template: an AI-READI example

Every instance follows the same template. For `aireadi_cataract-ocular` and one participant, the container stages 16 families with 168 files and an encoder manifest with 24 encoders. The target is binary (history of cataract), scored by AUROC across the 20 patients; the proposer listed `clinical-measurement-irregular`, `clinical-record-table`, `fundus-photography-2d`, `oct-volume-raster-mono8-3d` as relevant inputs. Figure 9 shows the instruction the agent receives.

aireadi_cataract-ocular

You are given one patient's record. Predict whether this participant has a history of cataract in one or both eyes.

Data

Everything staged for this patient is under `/workspace/data/`. The record is organised by *family*: a family is one named kind of input in one fixed format – for example the six-channel sleep EEG as an EDF file (`eeg-sleep-6ch-200hz-uv`), or coded clinical records as CSV tables (`clinical-record-table`). The encoders described under Tools are indexed by the same family names.

- `cgm-5min-mgdl` – 1 file under `refs/cgm-5min-mgdl/` (listed in `references.jsonl`)
- `clinical-measurement-irregular` – `tables/measurements.jsonl`: laboratory results, vital signs and body measurements as one long table, one JSON object per line (`t`, `measurement`, `value`, `unit`)
- `clinical-record-table` – `tables/condition_occurrence.csv`, `tables/measurement.csv`, `tables/observation.csv`, `tables/procedure_occurrence.csv`, `tables/visit_occurrence.csv`, `demographics.json`
- `ecg-12lead-500hz-mv` – 2 files under `refs/ecg-12lead-500hz-mv/` (listed in `references.jsonl`)
- `flio-tcspc-256x256x1024bin-u16` – 4 files under `refs/flio-tcspc-256x256x1024bin-u16/` (listed in `references.jsonl`)
- `fundus-faf-mono8-2d` – 2 files under `refs/fundus-faf-mono8-2d/` (listed in `references.jsonl`)
- `fundus-photography-2d` – 29 files under `refs/fundus-photography-2d/` (listed in `references.jsonl`)
- `numeric-observation` – 1 file under `refs/numeric-observation/` (listed in `references.jsonl`)
- `oct-bscanset-radial24-circle3-mono8` – 2 files under `refs/oct-bscanset-radial24-circle3-mono8/` (listed in `references.jsonl`)
- `oct-volume-raster-mono8-3d` – 12 files under `refs/oct-volume-raster-mono8-3d/` (listed in `references.jsonl`)
- `octa-enface-slab-mono8-2d` – 77 files under `refs/octa-enface-slab-mono8-2d/` (listed in `references.jsonl`)
- `octa-flowcube-mono8-3d` – 14 files under `refs/octa-flowcube-mono8-3d/` (listed in `references.jsonl`)
- `retinal-heightmap-surfaces-f32` – 14 files under `refs/retinal-heightmap-surfaces-f32/` (listed in `references.jsonl`)
- `wearable-hr-1min-bpm` – 1 file under `refs/wearable-hr-1min-bpm/` (listed in `references.jsonl`)
- `wearable-resprate-1min-brpm` – 1 file under `refs/wearable-resprate-1min-brpm/` (listed in `references.jsonl`)
- `wearable-spo2-1min-pct` – 1 file under `refs/wearable-spo2-1min-pct/` (listed in `references.jsonl`)

This is the patient's whole record up to the prediction time, not a curated subset. Deciding which inputs matter is part of the task.

`references.jsonl` lists this patient's larger signal and image files, each with its `path`, `family` and, where recorded, its time. Every `path` is relative to `/workspace/data/`, where the file is mounted read-only.

Prediction time

- The evidence is the participant's data from their study visit.
- The prediction is made as of that visit; each target is the participant's status as recorded by the study.

Tools

You can read every file directly with the software in this image: `numpy`, `scipy`, `pandas`, `scikit-learn`, and readers for the staged formats (`wfdb`, `pyedflib`, `pydicom`, `h5py`, `pillow`). Nothing else can be installed. Each file format needs its own reader; read each file's own header rather than assuming a layout.

Perception encoders (optional, recommended)

The modality perception (MP) service at `http://perception:8899` runs pretrained medical encoders over the staged signal and image files. It is optional: every file can be read directly. It is recommended because for some modality families – images especially – an encoder may be the only way to get a clinical reading at all. It is not perfectly reliable: a call may come back `healthy: false`, time out, or not apply to a file. Weigh what it returns against what you can check yourself.

`/workspace/data/encoder_manifest.md` lists the encoders available for this task.

- `GET /encoders` lists them; `GET /cards/{id}` gives one encoder's details, limits and options.
- `POST /encode` with `{"encoder_id": "...", "path": "<path relative to /workspace/data>", "options": {}}` runs one encoder on one file.
- 502 or 503: a cold or crashed encoder – retry once or twice. 400 or 422: the request was wrong. 429: this task's call budget is used up.

Submission

Write `/workspace/submission/predictions.csv` as a CSV with the header `target,prediction` and one row per target. Each target's section below says what its `prediction` is: a probability between 0 and 1, a class number, or a numeric value.

```
target,prediction
cataract_history,<probability between 0 and 1>
```

Target `cataract_history`

Your prediction is the probability that `mhoccur_crt` is YES (recorded value 1) for the participant. Higher values MUST mean more likely positive/higher class – do not invert.

Rules

Use only this patient's data in this environment. There is no network access. Time budget: 20 minutes of agent time. A saved submission is scored as-is when time runs out, so write `/workspace/submission/predictions.csv` early and refine it. Before you finish, check that the file exists and has one row per target. Resources: 4 CPUs, 8 GB RAM, no GPU.

Figure 9: The instruction for `aireadi_cataract-ocular` and one participant, verbatim (read left column, then right).

F RESULTS DETAILS

F.1 FULL LEADERBOARD

configuration	A (MedOmni)	AI-READI	BDSP	MIMIC-IV	binary	ordinal	continuous
GPT-6 Astra	62.7 [53.8, 72.9]	63.6 [48.6, 77.3]	56.6 [34.7, 76.1]	65.7 [49.8, 80.8]	65.4	78.3	15.5
GPT-6 Astra · MP off	60.9 [50.5, 71.3]	62.7 [48.6, 76.6]	50.4 [22.4, 79.2]	65.3 [49.7, 79.8]	63.9	74.6	15.1
DeepSeek-V4 Flash	54.1 [44.6, 65.3]	56.0 [39.6, 73.7]	40.4 [18.3, 62.3]	60.9 [44.1, 77.7]	56.4	71.8	9.1
GPT-6 Sol	53.9 [44.0, 66.1]	53.3 [38.1, 70.3]	42.9 [18.1, 70.5]	63.1 [47.5, 77.7]	55.2	73.8	13.7
MiMo-v2.6 Pro	52.1 [42.5, 65.6]	50.9 [34.4, 69.1]	37.1 [19.5, 59.1]	65.6 [48.2, 81.4]	53.6	61.0	0.6
GPT-6 Sol · MP off	50.6 [39.7, 62.1]	51.1 [34.6, 69.0]	35.3 [17.4, 59.6]	61.1 [45.6, 76.9]	51.3	74.2	13.7
GPT-6 Luna · MP off	50.1 [40.8, 62.1]	51.3 [38.6, 67.8]	30.0 [13.8, 54.9]	62.8 [46.4, 77.8]	51.0	66.2	7.9
GPT-6 Luna	43.4 [35.6, 55.0]	46.7 [33.0, 59.7]	32.1 [18.1, 52.7]	45.5 [29.2, 66.1]	46.0	47.4	9.4
GPT-5 nano	12.7 [9.9, 24.0]	11.0 [6.4, 24.0]	15.8 [6.3, 36.7]	13.7 [4.9, 33.3]	14.0	9.1	0.0

Table 13: Accuracy (A) with 95% CIs, overall, by source dataset and by target type. "MP off" denotes the same agent with the MP service unavailable.

comparison	ΔA [95% CI]
GPT-6 Astra − DeepSeek-V4 Flash	$+8.6^{*}$ [+2.3, +15.2]
GPT-6 Astra − GPT-6 Sol	$+8.8^{*}$ [+2.5, +15.5]
GPT-6 Astra − MiMo-v2.6 Pro	$+10.6^{*}$ [+3.0, +17.4]
GPT-6 Astra − GPT-6 Luna	$+19.4^{*}$ [+10.3, +27.2]
GPT-6 Astra − GPT-5 nano	$+50.0^{*}$ [+34.8, +58.0]
DeepSeek-V4 Flash − GPT-6 Sol	$+0.2$ [-7.5, +8.8]
DeepSeek-V4 Flash − MiMo-v2.6 Pro	$+2.0$ [-4.7, +8.8]
DeepSeek-V4 Flash − GPT-6 Luna	$+10.7^{*}$ [+0.3, +20.5]
DeepSeek-V4 Flash − GPT-5 nano	$+41.4^{*}$ [+25.9, +51.1]
GPT-6 Sol − MiMo-v2.6 Pro	$+1.8$ [-7.0, +9.6]
GPT-6 Sol − GPT-6 Luna	$+10.5^{*}$ [+0.2, +18.9]
GPT-6 Sol − GPT-5 nano	$+41.1^{*}$ [+25.4, +50.5]
MiMo-v2.6 Pro − GPT-6 Luna	$+8.8$ [-1.9, +20.0]
MiMo-v2.6 Pro − GPT-5 nano	$+39.4^{*}$ [+23.6, +49.2]
GPT-6 Luna − GPT-5 nano	$+30.6^{*}$ [+17.4, +39.3]

Table 14: Paired differences in A on shared patients (*: CI excludes 0).

F.2 BEHAVIOUR AND COST

model	B1 completion	B2 modality F1	B3 first answer	$ / trial	tokens / trial	steps / trial
GPT-6 Astra	100.0%	37.5	54% (368)	0.364	68k	7
DeepSeek-V4 Flash	99.7%	58.5	11% (531)	0.048	535k	19
GPT-6 Sol	100.0%	31.1	42% (286)	0.071	68k	7
MiMo-v2.6 Pro	90.8%	62.1	44% (564)	0.030	266k	14
GPT-6 Luna	99.8%	29.6	23% (230)	0.004	58k	6
GPT-5 nano	78.7%	17.2	31% (149)	0.021	51k	6

Table 15: Behaviour (B1–B3) and cost. B2 averages 339 trials per agent; consulting every staged modality scores 61.5. B3 in brackets: trials that consult any modality. Tokens and steps are medians.

F.3 MODALITY LADDER

Five task types (A-CRP-1, A-HTN, B-OSA, M-CKD, M-HF; Table 12), the same 20 patients per task, six cumulative rungs; scored with A. The ladder masks every CKD-stating code. Rungs per source: AI-READI: R1 demogr., R2 wearables, R3 ECG-12, R4 CGM, R5 eye imaging, R6 record+labs; MIMIC-IV: R1 demogr., R2 ED+vitals, R3 labs, R4 dx history, R5 CXR, R6 ECG-12 (full); BDSP: R1 demogr., R2 snore+pos., R3 flow+effort, R4 SpO_2+pulse, R5 EEG/EOG/EMG, R6 ECG+record (full).

F.4 MP SERVICE OFF VS. ON

	R1	R2	R3	R4	R5	R6
GPT-6 Astra						
all five	26.7	48.8	56.8	58.1	56.1	72.5
AI-READI	7.9	32.6	35.4	33.6	29.5	65.1
MIMIC-IV	45.5	75.5	83.5	82.0	79.0	81.5
BDSP	26.8	27.8	46.2	59.3	63.6	69.3
GPT-6 Luna						
all five	29.4	36.1	40.2	37.2	43.4	48.0
AI-READI	12.6	10.1	15.4	11.6	16.3	29.6
MIMIC-IV	51.5	68.0	71.5	68.5	70.0	70.5
BDSP	18.7	24.1	27.4	25.8	44.5	39.8
GPT-6 Sol						
all five	25.5	48.3	49.9	51.3	57.6	66.5
AI-READI	3.9	33.9	25.9	27.3	30.8	61.9
MIMIC-IV	46.0	70.0	83.0	79.0	81.5	74.0
BDSP	27.9	33.8	31.4	44.0	63.4	60.9

Table 16: Modality ladder: A per rung.

model	scope	step	ΔA [95% CI]
GPT-6 Astra	all five	R1→R2	$+22.1^*$ [+5.2, +35.3]
GPT-6 Astra	all five	R5→R6	$+16.4^*$ [+2.2, +35.2]
GPT-6 Astra	all five	R1→R6	$+45.8^*$ [+25.2, +61.6]
GPT-6 Astra	all five	R2→R6	$+23.7^*$ [+3.9, +43.5]
GPT-6 Astra	AI-READI	R5→R6	$+35.6^*$ [+11.5, +61.0]
GPT-6 Astra	AI-READI	R1→R6	$+57.1^*$ [+20.3, +83.5]
GPT-6 Astra	AI-READI	R2→R6	$+32.5^*$ [+2.6, +64.5]
GPT-6 Astra	MIMIC-IV	R1→R2	$+30.0^*$ [+8.0, +51.5]
GPT-6 Astra	MIMIC-IV	R1→R6	$+36.0^*$ [+10.5, +61.5]
GPT-6 Astra	BDSP	R2→R3	$+18.4^*$ [+4.3, +29.0]
GPT-6 Astra	BDSP	R3→R4	$+13.0^*$ [+2.2, +22.2]
GPT-6 Astra	BDSP	R1→R6	$+42.5^*$ [+24.2, +58.8]
GPT-6 Astra	BDSP	R2→R6	$+41.5^*$ [+24.2, +54.6]
GPT-6 Luna	all five	R1→R6	$+18.6^*$ [+3.7, +32.6]
GPT-6 Luna	BDSP	R4→R5	$+18.6^*$ [+0.1, +34.5]
GPT-6 Luna	BDSP	R1→R6	$+21.1^*$ [+1.7, +40.2]
GPT-6 Sol	all five	R1→R2	$+22.8^*$ [+5.5, +40.8]
GPT-6 Sol	all five	R1→R6	$+41.0^*$ [+18.1, +63.5]
GPT-6 Sol	all five	R2→R6	$+18.2^*$ [+1.1, +36.0]
GPT-6 Sol	AI-READI	R1→R2	$+30.1^*$ [+2.4, +59.0]
GPT-6 Sol	AI-READI	R5→R6	$+31.1^*$ [+5.0, +54.1]
GPT-6 Sol	AI-READI	R1→R6	$+58.0^*$ [+22.3, +91.0]
GPT-6 Sol	AI-READI	R2→R6	$+28.0^*$ [+2.2, +60.7]
GPT-6 Sol	BDSP	R4→R5	$+19.4^*$ [+5.2, +32.0]
GPT-6 Sol	BDSP	R1→R6	$+33.0^*$ [+16.1, +47.3]
GPT-6 Sol	BDSP	R2→R6	$+27.1^*$ [+10.8, +40.8]

Table 17: Ladder contrasts whose 95% CI excludes 0 (single steps and R1/R2→R6).

model	all	AI-READI	BDSP	MIMIC-IV
GPT-6 Astra	$+1.8$ [-1.8, +5.9]	$+0.9$ [-2.5, +4.3]	$+6.2$ [-7.1, +22.9]	$+0.4$ [-4.4, +5.4]
GPT-6 Luna	-6.7 [-15.1, +2.7]	-4.6 [-17.9, +7.4]	$+2.1$ [-15.2, +19.2]	-17.3 [-33.6, +0.7]
GPT-6 Sol	$+3.3$ [-3.3, +8.9]	$+2.2$ [-6.9, +11.3]	$+7.6$ [-8.3, +21.8]	$+2.0$ [-5.2, +9.7]

Table 18: Paired difference in A, service on $-$ off, same instances (*: CI excludes 0). Each arm ran once; the contrast measures offering the service, not using it.

G Agent Traces and Case Studies

G.1 Holistic Trace Analysis

Each agent step (one scaffold episode) gets one action, the first class it matches in the order *write answer, MP service call, read raw modality file, read clinical record, read encoder cards, orient, reason/other*. Matching uses the step's keystrokes plus the MP service's call log and file-access witness inside the step's time window, which also catches MP service calls and file reads wrapped in Python. Table 19 and Figures 10–11 summarise the trials.

agent	median steps	AI-READI	BDSP	MIMIC-IV
GPT-6 Astra	7	17 / 4 / 27 / 16	30 / 15 / 5 / 17	10 / 6 / 22 / 27
DeepSeek-V4 Flash	19	25 / 11 / 31 / 9	31 / 27 / 8 / 9	15 / 14 / 27 / 11
GPT-6 Sol	7	9 / 3 / 35 / 23	21 / 14 / 9 / 23	10 / 7 / 20 / 34
MiMo-v2.6 Pro	14	21 / 17 / 33 / 15	34 / 17 / 11 / 15	12 / 18 / 31 / 20
GPT-6 Luna	6	6 / 2 / 44 / 21	12 / 25 / 14 / 25	10 / 6 / 27 / 23
GPT-5 nano	6	4 / 1 / 43 / 27	1 / 0 / 32 / 26	1 / 0 / 50 / 26

Table 19: Step mix by source (% of steps: raw file / encoder / record / write); the remainder is reasoning, orientation and encoder cards.



Figure 10: What agents spend their steps on, overall and by source dataset (% of steps per action); the grey number right of each bar is the median number of steps per trial.

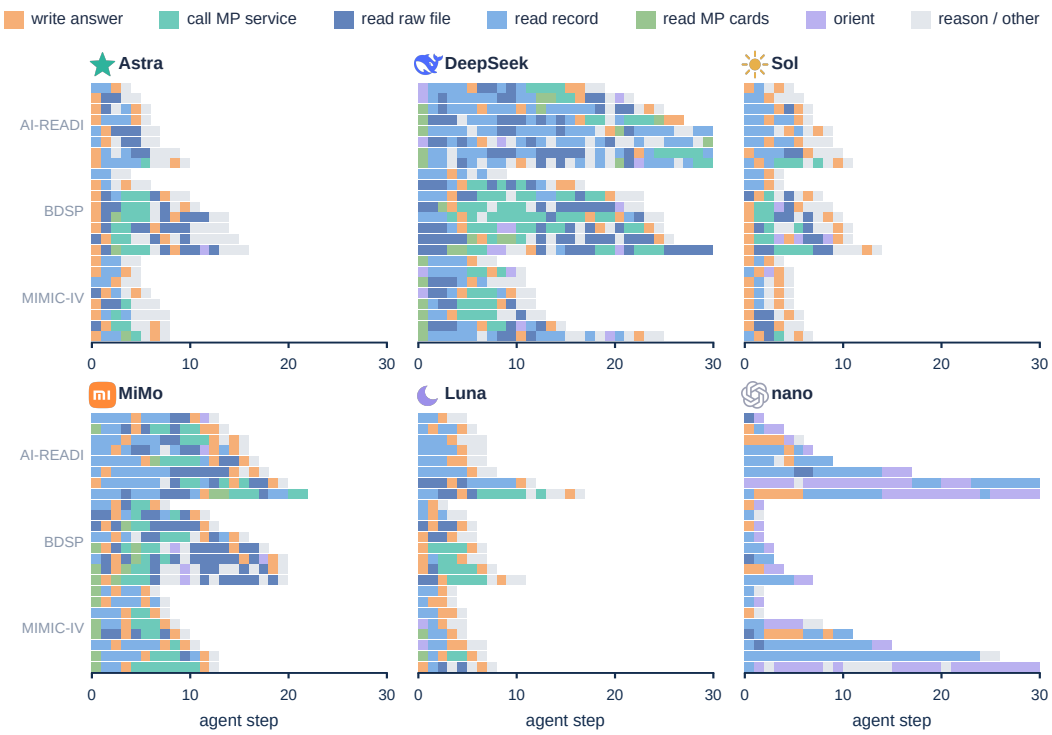

Figure 11: Trace barcodes: eight random trials per agent and source dataset, one cell per step (first 30 steps), coloured by action; trials sorted by length within each source.

G.2 Runtime Case Studies

We randomly select one GPT-6 Astra run from each of two sources. Each page names the task, then shows everything staged for the patient, one compact tile per modality family rendered from the raw files; the mark on each tile shows whether Astra called the MP service on that family, opened one of its files, or did not use it. Below the tiles, Astra's steps are shown in order, coloured by action, next to its answers and the gold labels.

The two runs illustrate a shared evidence-integration workflow across very different clinical settings. In the BDSP sleep study, Astra combines direct inspection of overnight signals with targeted MP service calls on sleep EEG, EOG, and SpO_2, then revises its prediction as new evidence becomes available. In the AI-READI task, it selects retinal imaging from a much broader multimodal record and uses MP outputs before revising its answer. Rather than inspecting every available modality, Astra follows a task-directed route through the evidence and combines raw access with specialized perception. Together, the examples show how the MP service can support selective integration while preserving the agent's control over what evidence to inspect and when to update its decision.

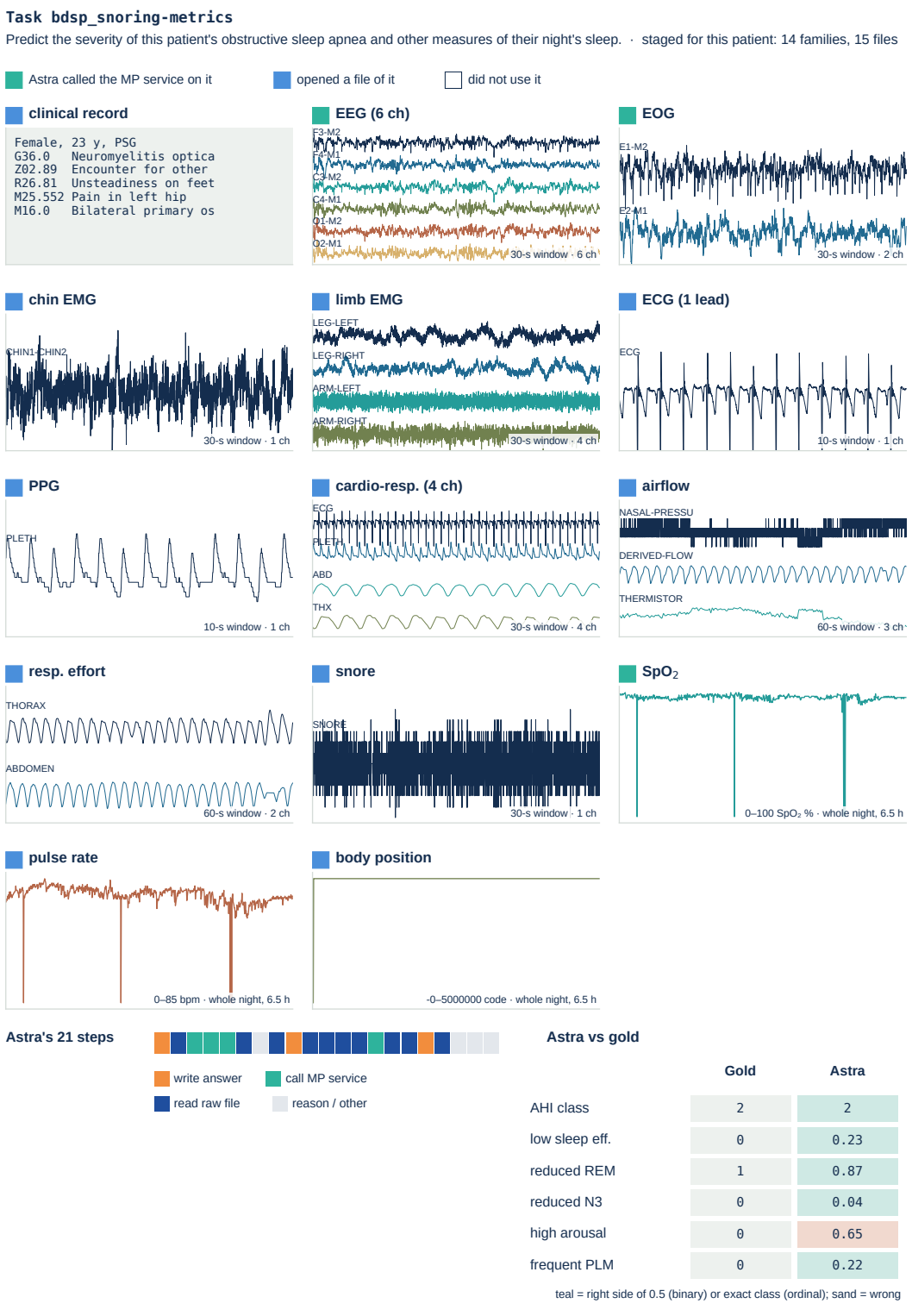

Figure 12: Case 1, `bdsp_snoring-metrics`: one overnight polysomnogram and the diagnosis list. Waveforms show one mid-night window; SpO_2, pulse and body position show the whole night. Astra calls the MP service on sleep EEG, EOG and SpO_2 (one SpO_2 call returns unhealthy), and rewrites its answer twice after those calls.

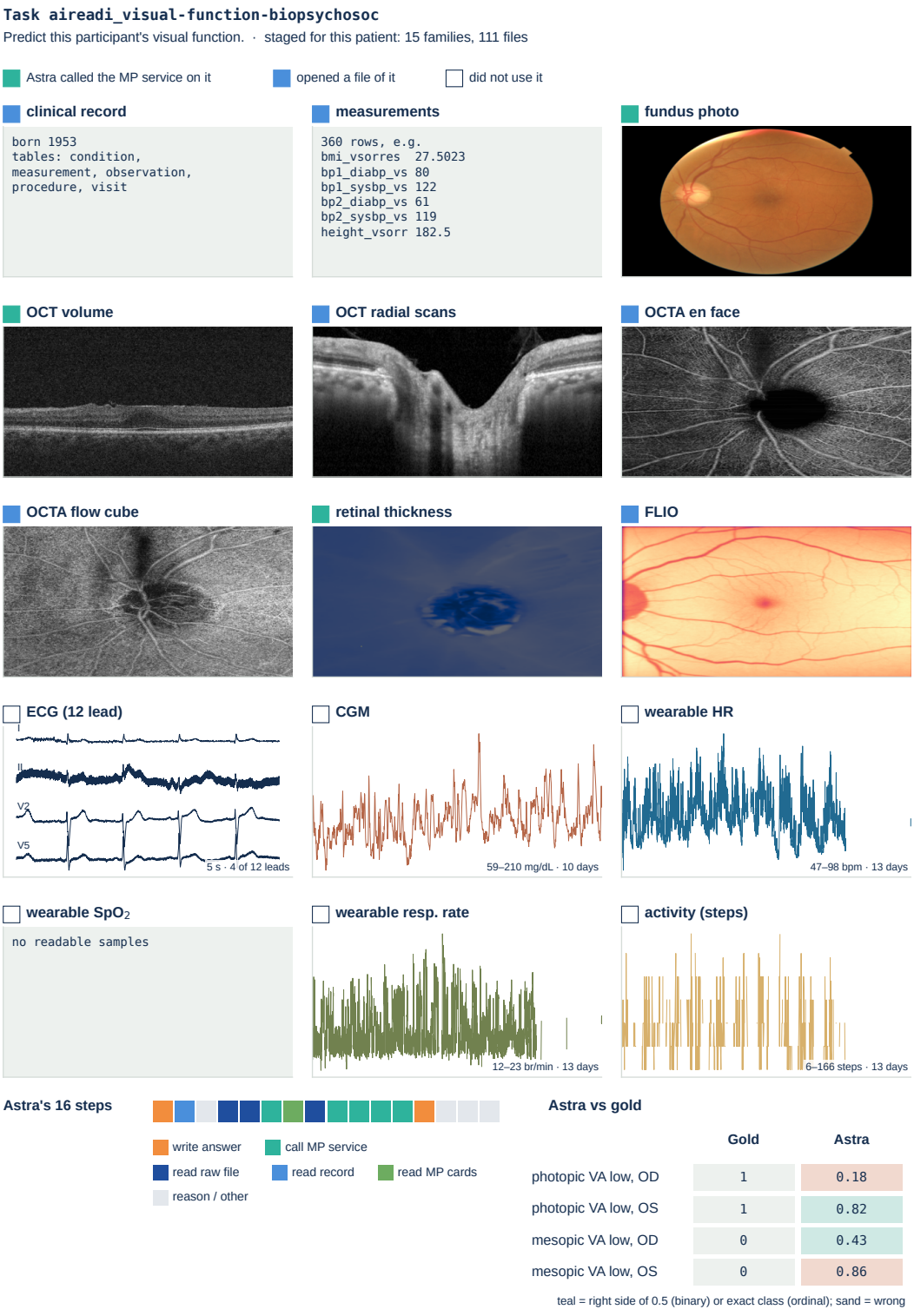

Figure 13: Case 2, `aireadi_visual-function-biopsychosoc`: every staged family, one representative file each (retinal imaging, ECG, CGM, wearables, record and measurements). Astra calls the MP service on fundus photographs, OCT volumes and retinal thickness maps before revising its answer. The four autorefraction targets of this task are held (Section E.4) and not shown.