CliniCARE-Bench: Clinical Calibrated Audit of Medical Reasoning in EHR

Veronica Chatrath[*1], **Bryan Zhu**[*1], **George Pu**[1], **Jingxuan Fan**[1], **Apaar Shanker**[1], **Varun Ursekar**[1], **Anahita Sharma**[1], **Jason Qin**[1], **Keqi Han**[1,2], **Soham Dinesh Tiwari**[1], **Soham Dan**[1], **Vijay Kalmath**[1], **Yuan (Christy) Li**[1], **Daniel Yue Zhang**[1], **Chenguang Wang**[1,4], **Zainab Doctor**[1], **Zhijun Yin**[1,3], **Nigam H. Shah**[5, 6, 7], **Yuan Xue**[1]

[1]Scale AI, [2]Emory University, [3]Vanderbilt University Medical Center, [4]University of California, Santa Cruz, [5]Department of Medicine, Stanford School of Medicine, [6]Technology and Digital Solutions, Stanford Health Care, [7]Clinical Excellence Research Center, Stanford School of Medicine, [*]Co-first authors.

veronica.chatrath@scale.com, yuan.xue@scale.com | https://scale.com/research

Abstract

Large language models perform strongly on medical knowledge benchmarks, but reliable clinical deployment requires agents to conduct defensible investigations over heterogeneous, longitudinal records. Agents must determine what evidence is needed, retrieve and reconcile structured and free-text data, ground conclusions in verifiable evidence, and defer cases that cannot be resolved reliably. We introduce CLINICARE-BENCH (**Clini**cal **C**alibrated **A**udit of Medical **R**easoning in **E**HR), a deployment-oriented benchmark for retrospective clinical audit: 25 clinician-validated scenarios instantiated as 750 patient-specific cases over real-patient-derived MIMIC-IV data. Systems investigate each case through a governed, logged tool environment for record retrieval, computation, and policy access, and return one of four verdicts—*Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*—the last two separating missing evidence from residual medical ambiguity. Beyond verdict accuracy, we score patient-evidence and policy grounding, process adherence, calibrated abstention, reliability, and efficiency against case-level reference verdicts produced by independent multi-model adjudication and calibrated against Clinical Board review. As every retrieval, computation, and report is replayable, the investigation and adjudication trace is itself inspectable and scorable. To our knowledge, CLINICARE-BENCH is the first deployment-oriented clinical-agent benchmark to jointly evaluate real longitudinal EHR investigation, claim-level evidence grounding, governing-policy use, process adherence, and calibrated abstention within a common patient-level adjudication framework. In the evaluation of 16 agentic systems, four-way accuracy spans 65.3–76.1%. However, raw accuracy overstates the quality of the underlying investigation. Defect-free accuracy, which credits a verdict only when it is correct and free of prohibited shortcuts, is 4.8–14.8 percentage points lower and reorders the leaderboard. We will release the scenarios and evaluation code soon.

1 Introduction

Many high-value clinical AI tasks are not self-contained questions or prospective treatment decisions but retrospective investigations of the patient record. Was a protocol followed? Did a clinical event occur within a given window? Does the documentation support a quality or operational conclusion? Answering any of these means working through a heterogeneous, longitudinal electronic health record (EHR), deciding what evidence is needed, retrieving it across structured data and free-text notes, reconstructing the timeline, reconciling conflicting sources, applying the governing clinical standard, and recognizing when the record simply cannot settle the question. This is evidence-grounded clinical audit and adjudication, not medical question answering.

Foundation models now match or exceed expert scores on medical knowledge exams [1, 2], but exam performance is a poor proxy for this investigation and adjudication process. Existing clinical benchmarks each capture part of that process. Knowledge suites probe internalized concepts with no environment to act in [3]. Interactive simulators drive sequential diagnosis on synthetic or partly simulated substrates [4, 5]. EHR-grounded frameworks span heterogeneous cases [6], expert instructions [7], and multi-year timelines [8]. Factuality benchmarks check whether generated statements are supported by the record [9, 10]. Three safety-critical capabilities, however, remain poorly tested: *negative reasoning* (confirming a condition was considered and ruled out), *deep longitudinal synthesis* across multi-year records [11], and *calibrated abstention* that separates missing evidence from genuine medical ambiguity and defers rather than guessing when the record is incomplete or contradictory [12, 13]. More fundamentally, few benchmarks treat the *defensibility* of an adjudication as the unit of evaluation: whether the agent gathered the necessary evidence, grounded its claims, applied the governing standard, and escalated when the record could not support a verdict. Deployment demands more than a correct final answer. It demands an evidence acquisition and adjudication process that is complete, faithful, reproducible, and calibrated.

We introduce **CliniCARE-Bench** (**Clini**cal **C**alibrated **A**udit of Medical **R**easoning in **E**HR), a suite of 25 clinician-authored, evidence-intensive scenarios instantiated as 750 patient cases over MIMIC-IV v3.1 [14] (Figure 1). Rather than handing agents a preassembled context or raw database access, we place them in a reproducible runtime with governed, clinician-verifiable tools for retrieving structured records and notes, running explicit computations, and consulting a fixed corpus of governing policies, isolating query engineering artifacts while preserving the record's heterogeneity. Agents must combine medical knowledge, patient evidence, and authoritative standards such as KDIGO [15], CMS SEP-1 [16], AABB [17], and AHA/ACC guidance [18].

Every case demands one of four verdicts: *Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*. The two indeterminate classes separate cases that more evidence could resolve from those that need expert interpretation even after a full review. Reference verdicts come from a multi-model adjudication ensemble applying the clinician-validated specifications, calibrated against blinded Clinical Board review on a stratified subset. Abstention is thus built into the evaluation standard, not bolted on as a post hoc confidence threshold.

To our knowledge, CLINICARE-BENCH is the first *deployment-oriented* clinical agent benchmark to score real longitudinal EHR investigation, claim-level evidence grounding, governing policy use, process adherence, and calibrated abstention *jointly* within one patient-level adjudication framework. No benchmark in Table 1 combines all of these axes. It builds on the trace-level clinical auditing introduced by Hager et al. [19] but broadens what is scored within the trace. Every retrieval and computation is logged, and every verdict must cite the patient evidence it rests on, together with any governing policy that applies. A defensible conclusion can then be told apart from a lucky guess or an invalid shortcut,

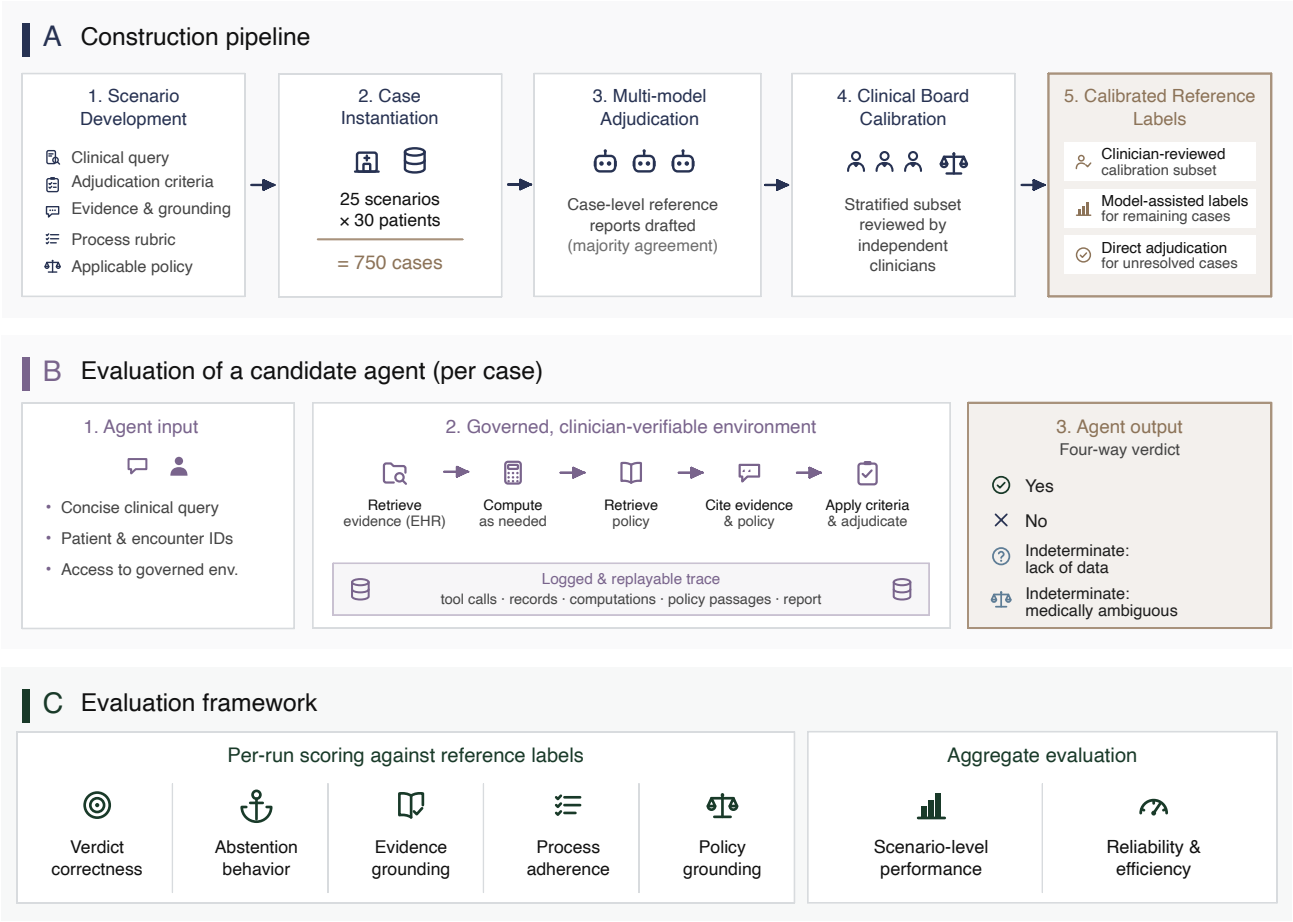

Figure 1. **(A)** Clinician-authored scenarios are instantiated as patient-specific MIMIC-IV cases, independently adjudicated by a multi-model ensemble, and calibrated against Clinical Board review into case-level reference verdicts. **(B)** Per case, a system answers a clinical query within a governed environment for record retrieval, computation, and policy access, producing a cited four-way verdict and a replayable trace. **(C)** Runs are scored on verdict correctness, abstention, evidence grounding, process adherence, and policy grounding, then aggregated into scenario-level performance, reliability, efficiency, and cost.

without ever inspecting the model's private reasoning. Making this feasible at scale is a protocol for reference verdicts, calibrated against clinician review, that replaces exhaustive manual labeling.

Throughout the paper, we refer to the task-performing entity as the *agent*, and to the complete evaluated configurations, including the model, agent harness, prompting or scaffolding, and available tools, as the *system*. In an evaluation of 16 systems combining production agent harnesses with frontier models, as well as open-weight models evaluated on a common harness, even the best system resolves only 76.1% of cases correctly. Every system also under-abstains: the over-commitment rate on cases whose reference verdict requires deferral exceeds the over-abstention rate on cases with definitive reference verdicts. These results suggest substantial remaining gaps toward reliable autonomous clinical audit.

We make four methodological contributions:

1. **A skill- and specialty-spanning scenario suite.** 25 clinician-developed, evidence-intensive scenarios instantiated as 750 patient cases over MIMIC-IV, spanning protocol and guideline audits, clinical

event determination, and longitudinal status assessment across 14 medical specialties drawn from the American Board of Medical Specialties (ABMS) taxonomy[1] and ten reasoning capabilities, stress-testing safety-critical reasoning that existing benchmarks rarely evaluate.

2. **A governed, clinician-verifiable tool environment.** A reproducible execution surface over real longitudinal EHR data with patient-scoped retrieval, explicit computation, policy document access, and fully logged interactions, so performance reflects clinical investigation rather than query engineering.
3. **A grounded, process-aware evaluation framework.** A strict four-way verdict paired with claim-level evidence grounding, required evidence coverage, policy citation and adherence, and a weighted MUST-DO/MUST-NOT process rubric rewarding faithful evidence use, appropriate abstention, and sound procedure, not correctness of the final answer alone.
4. **A scalable protocol for clinician-calibrated reference labeling.** Specifications authored and validated by clinicians, combined with independent multi-model case adjudication and blinded Clinical Board calibration, yielding reliable reference verdicts at a scale exhaustive manual review could not reach.

Key empirical finding: correct verdicts can mask defective investigations. Defect-free accuracy credits a verdict only when it is correct and the investigation violates no prohibited MUST-NOT criterion. Across the 16 systems, defect-free accuracy is 4.8–14.8 percentage points below accuracy on the same report-present runs, and this correction changes the ordering of systems. For the most affected system, roughly one in five otherwise-correct verdicts violates at least one prohibited shortcut criterion. This extends the outcome–process gap documented for medical LLMs [20, 21] to agentic investigation over real longitudinal EHRs, and captures precisely the failure mode that CliniCARE-Bench is designed to expose.

Together these pieces form an open benchmark for whether clinical agents have the safety-critical capabilities that real-world audit and review work demands. Though instantiated in medicine, our design, including governed, auditable tool use, explicit adjudication criteria, traceable evidence, calibrated abstention, and scoring of the observable adjudication process, transfers to other high-stakes domains such as finance, cybersecurity, and law, where decisions must be transparent, reproducible, and verifiable. We will release the benchmark data, evaluation code, and reproducibility resources, for authorized MIMIC-IV users.

2 Related Work

2.1 Static Knowledge-Centric and Context-Provided Clinical Evaluation

Early medical LLM benchmarks primarily assess whether models can apply medical knowledge and reasoning to predefined question-answering tasks. For example, MultiMedQA [1] and related medical-examination benchmarks [2] combine professional examination questions, biomedical research questions, and consumer health queries. Although these benchmarks span diverse clinical domains, the information required to answer each question is generally contained in the prompt or encoded in the model's parameters. More recent rubric-based benchmarks extend evaluation from closed-form questions to open-ended interactions. HealthBench [22] evaluates healthcare conversations using physician-authored criteria, while HealthBench Professional [3] focuses on cases drawn from model–clinician interactions, including clinical consultation, documentation, and medical research. These benchmarks improve the realism of clinical application, but they primarily evaluate how models respond to information

[1] https://www.abms.org/member-boards/specialty-subspecialty-certificates/

made available within a predefined interaction context rather than how they acquire evidence from a longitudinal patient record.

Beyond conversational evaluation, another line of work grounds model assessment in patient-derived EHR data assembled into case-specific inputs or contexts. MedAlign [7] pairs clinician-authored instructions with expert responses grounded in longitudinal EHRs; EHRSHOT [8] evaluates few-shot prediction from structured patient data; and TIMER [11] targets temporal reasoning over longitudinal clinical records. At a broader level, MedHELM [6] provides a clinician-validated framework for organizing evaluation across heterogeneous medical tasks. Collectively, these resources assess important capabilities in instruction following, temporal reasoning, and patient-specific prediction. A shared boundary, however, is that relevant patient information is typically supplied, preprocessed, or retrieved outside the evaluated decision loop. They therefore do not directly evaluate whether an agent can determine what evidence is needed, retrieve that evidence across heterogeneous EHR sources, and assess whether the available record is sufficient to support a defensible conclusion.

2.2 Interactive Clinical Simulation and Prospective Decision-Making

Interactive clinical environments evaluate sequential information acquisition and decision-making. AgentClinic [4] simulates clinical encounters in which agents interact with patients, request information, use tools, and formulate diagnoses. The Clinical Environment Simulator [5] extends this paradigm to digital hospitals by modeling patient-state transitions alongside operational constraints such as bed availability, staff workload, and equipment status. These environments expose failures not captured by static question answering, including inefficient information gathering, inappropriate test selection, and errors arising from sequential interactions. However, their patient states or interaction dynamics are wholly or partly simulated, and they primarily evaluate prospective diagnosis and management.

Motivated by the view that clinical AI should be evaluated within evolving real-world workflows rather than as a static model in isolation [23], recent benchmarks construct interactive evaluation from real clinical data. MIMIC-CDM [19] derives sequential decision-making cases from MIMIC-IV cases, requiring models to request physical examinations, laboratory tests, and imaging before producing diagnoses and treatment plans. Its original evaluation, however, was restricted to open-access models from the Llama 2 generation, leaving the performance of newer frontier agents in this environment unresolved. MIRA [24] constructs prospective-style encounters from MIMIC-IV-derived cases, allowing agents to elicit information, order tests, and develop diagnostic and treatment plans. ClinEnv [25] converts real inpatient admissions into sequential decision stages and evaluates both clinical decisions and information-acquisition behavior. These benchmarks substantially advance the evaluation of longitudinal, consequential clinical decision-making, but their primary unit of evaluation remains a prospective decision or management trajectory rather than what can be established retrospectively from an existing longitudinal record.

2.3 Agentic Retrieval and Execution over EHR Workflow

Research on agentic EHR interaction has progressed from executable question answering to long-horizon clinical workflows. EHRSQL [26] pairs clinician-derived questions with executable SQL queries over structured EHR databases and includes questions that cannot be answered under the available schema. EHRAgent [27] extends this setting by iteratively generating and executing code for multi-table EHR reasoning. MedAgentBench [28] evaluates physician-authored cases in a FHIR-compliant virtual EHR, while FHIR-AgentBench [29] evaluates retrieval, interaction, and reasoning strategies for question

answering over interoperable FHIR resources. EHR-Complex [30] further scales executable analysis over MIMIC-IV to patient- and population-level cases requiring SQL or Python.

More recent benchmarks evaluate longer and more composite clinical workflows. PhysicianBench [31] requires agents to retrieve information across encounters, reason over heterogeneous records, perform clinical actions, and produce documentation, with performance assessed through execution-grounded checkpoints. LongMedBench [32] evaluates fact retrieval, temporal reasoning, and clinical decision-making across repeated admissions and extended patient histories. Collectively, these benchmarks advance the evaluation of EHR retrieval, executable reasoning, case completion, and long-horizon workflow performance, with evaluation centered primarily on answer correctness, execution success, and completion of predefined clinical tasks.

2.4 Evidence Grounding, Policy Adherence, and Abstention in Clinical AI

Complementing benchmarks focused on interaction and case completion, another line of work evaluates the reliability of clinical model outputs beyond aggregate answer accuracy. VeriFact-BHC [9] and FactEHR [10] assess the factual consistency of clinical text against source records, while ArchEHR-QA [33] evaluates evidence-grounded question answering over clinical notes. Related work also shows that performance on medical multiple-choice benchmarks can decline substantially when familiar answer options are replaced with a none-of-the-above choice, suggesting that high accuracy may partly reflect response-format and answer-pattern cues rather than robust evidence-based reasoning [20]. These benchmarks highlight the importance of determining whether generated claims are supported by patient records and whether model conclusions are grounded in the available evidence.

Reliability also depends on whether an agent applies the appropriate clinical standard and follows case-specific investigation procedures. MIMIC-CDM [19] evaluates guideline concordance using encoded criteria for recommended testing and treatment rather than requiring models to retrieve, cite, or interpret the guidelines. Prior work has also incorporated guideline knowledge into longitudinal EHR decision support [34] and evaluated intermediate workflow and information-acquisition behavior [25, 31]. These studies motivate evaluation of whether agents identify and apply relevant standards, interpret thresholds, timing rules, and exceptions correctly, and follow required investigation steps without relying on invalid evidentiary shortcuts [20, 21].

Prior work has further examined when clinical systems should abstain from a definitive answer. EHRSQL [26, 35] evaluates whether questions can be answered under the available database schema, MedAbstain [13] studies uncertainty-sensitive abstention in medical question answering, and ClinDet-Bench [12] assesses whether incomplete clinical descriptions contain sufficient information to apply a clinical criterion. Together, these studies illustrate different sources of indeterminacy, including missing information and uncertainty that remains after the available evidence has been considered.

2.5 The CliniCARE-Bench Distinction

Existing clinical benchmarks have evaluated medical knowledge [1, 2], temporal reasoning [11], factuality [9, 10], abstention [12, 13, 26], and agentic EHR interaction [28–30], but these capabilities are typically assessed in separate settings. CliniCARE-Bench brings them together in an end-to-end evaluation of retrospective clinical audit over real-patient-derived longitudinal EHR data [14]—to our knowledge the most comprehensive combination of these capabilities in a single clinical agent benchmark to date (Table 1). Given a concise patient-level audit question, the agent must independently plan and execute retrieval across structured records and free-text notes through governed, auditable interfaces; identify

and apply the governing clinical policy or adjudication criteria; and produce a cited verdict supported by a complete and traceable evidence trail.

CliniCARE-Bench is designed as a deployment-oriented first exam for selective autonomy rather than solely as a frontier stress test [36]. It focuses on representative, evidence-intensive workflows that recur in clinical quality, protocol, and documentation review. This distinction is also motivated by the optimization paradox observed in clinical multi-agent systems. Using the MIMIC-CDM environment of Hager et al., Bedi et al. [37] found that improvements in component-level measures did not reliably predict end-to-end system accuracy, including across diagnostic outcomes, process adherence, and cost-related measures. CliniCARE-Bench therefore reports these dimensions separately and evaluates the complete investigation-and-adjudication trajectory rather than treating any component metric as a proxy for clinical correctness. It assesses whether an agent can reliably resolve cases within its capabilities while distinguishing cases that lack necessary evidence from those that remain medically ambiguous after review. Accordingly, performance extends beyond final-verdict accuracy to the adequacy of evidence retrieval, faithfulness of cited claims, completeness of decisive findings, adherence to applicable rules and required investigation procedures, and appropriate escalation. This design is intended to inform human-supervised deployment by measuring both the share of work an agent can resolve autonomously and whether those decisions are sufficiently transparent, reproducible, and auditable for operational use.

More broadly, the need for such auditable agents is not specific to medicine. Recent work argues for formalizing LLM agent security in terms of explicit, verifiable properties, such as authorized objectives, action alignment, source authorization, and data isolation, rather than treating agent trustworthiness as implicit [38]. CliniCARE-Bench provides a clinical instantiation of this broader agenda by combining governed, logged tool access with process-aware evaluation of evidence grounding, policy adherence, and reproducibility. This framework provides a general recipe for auditable agent environments that transfers to other high-stakes domains, such as finance, cybersecurity, and law, where agent decisions must be transparent and verifiable.

CLINICARE-BENCH is designed to complement existing efforts. CLINICARE-BENCH will be integrated into the MedHELM framework [6] so that its process-aware, grounding- and abstention-scored cases can be evaluated alongside benchmarks such as PhysicianBench [31] and HealthAdminBench [39] within a shared, openly accessible evaluation ecosystem.

3 The Benchmark Environment

3.1 Data Substrate

CLINICARE-BENCH is constructed on top of three datasets from the MIMIC-IV database: the core MIMIC-IV v3.1, MIMIC-IV-Note v2.2, and MIMIC-IV-ED v2.2 [14, 40, 41]. Together, these datasets provide retrospective, de-identified records from the inpatient, intensive-care, and emergency departments of a single major academic medical center. They include both structured clinical events and unstructured discharge summaries and radiology reports, organized into four modules comprising 41 tables:

- **The `hosp` module (22 tables):** Contains hospital-wide encounter data, including admissions, transfers, coded diagnoses and procedures (ICD-9/10), medication ordering and administration (prescriptions/eMAR), laboratory results, microbiology, and code dictionaries.
- **The `icu` module (9 tables):** Captures high-frequency granular data from intensive care units, including chart events, fluid input/output balances, and procedural records.

Table 1. Positioning of CliniCARE-Bench against representative clinical-AI benchmarks. We compare evaluation frameworks across their technical substrates and core evaluation capabilities. Checkmarks (✓) denote full support, tildes (∼) denote partial support, and em-dashes (—) indicate absence.

Benchmark	Substrate	Agentic Env.	Real EHR	Grounding Scored	Calibrated Abstention
AgentClinic [4]	Simulated patients	✓	—	—	—
HealthBench Professional [3]	Model–clinician chat cases	—	—	—	—
HealthAdminBench [39]	Admin / claims	∼	—	—	—
MedHELM [6]	Mixed EHR cases	∼	✓	∼	—
VeriFact-BHC [9]	MIMIC-III notes	—	✓	✓	—
FactEHR [10]	Multi-source notes	—	✓	✓	—
ArchEHR-QA [33]	MIMIC-III/IV note excerpts	—	✓	✓	—
TIMER [11]	Longitudinal notes	—	✓	∼	—
CliCARE [34]	Cancer EHR + guidelines	—	✓	—	—
MedAlign [7]	Stanford EHR	—	✓	∼	—
EHRSHOT [8]	Stanford EHR (structured)	—	✓	—	—
EHRSQL [35]	MIMIC-IV (text-to-SQL)	—	✓	—	∼
MedAbstain [13]	Medical MCQ	—	—	—	∼
ClinDet-Bench [12]	Clinical scoring cases	—	—	—	∼
EHRAgent [27]	MIMIC-III / eICU (tabular)	✓	✓	—	—
EHR-Complex [30]	MIMIC-IV (SQL/code)	✓	✓	—	—
MedAgentBench [28]	Virtual FHIR EHR	✓	∼	—	—
FHIR-AgentBench [29]	MIMIC-IV-FHIR	✓	✓	—	—
MIRA [24]	Sim. EHR (MIMIC-IV)	✓	∼	—	—
ClinEnv [25]	Real inpatient admissions	✓	✓	—	—
PhysicianBench [31]	Real-world EHR workflows	✓	✓	—	—
LongMedBench [32]	Longitudinal MIMIC-IV	∼	✓	—	—
CliniCARE-Bench (Ours)	**MIMIC-IV (all modules)**	✓	✓	✓	✓

Note: "Agentic Env." implies the model autonomously plans multi-step retrieval and executes tool calls in a closed loop. "Grounding Scored" indicates that model assertions are directly evaluated against granular, cited record spans. "Calibrated Abstention" denotes that the benchmark explicitly scores and rewards principled abstention, distinguishing insufficient evidence (Indeterminate: Lack of Data) from genuine medical ambiguity (Indeterminate: Medically Ambiguous), when chart evidence cannot support a definitive verdict.

- **The `ed` module (6 tables):** Preserves emergency department timelines, tracking triage vitals, initial acuity levels, and medication reconciliation upon arrival.
- **The `note` module (4 tables):** Provides the full unstructured text of discharge summaries and radiology reports alongside structured note-level metadata (e.g. author, exam name, CPT code).

In total, the database comprises 364,627 unique patients, 546,028 hospital admissions, 94,458 ICU stays, approximately 331,000 discharge summaries, 2.3 million radiology reports, and around 900 million structured rows.

3.2 Agentic Tool Environment

While MIMIC-IV provides the underlying data substrate, CLINICARE-BENCH does not expose its physical database schema as the primary interface. Instead, agents interact with patient records through a governed tool layer organized around clinically meaningful entities and retrieval operations. This

design reflects a common production access pattern in which clinical AI systems interact with EHR data through application-layer interfaces, such as Fast Healthcare Interoperability Resources (FHIR)-based interfaces or vendor-specific APIs, rather than through unrestricted access to raw database tables. Our environment is not intended to reproduce a specific interoperability standard. Instead, it captures three operational properties central to evaluation.

First, access through the primary clinical tool surface is scoped to one patient at a time. Rather than receiving a preassembled record, the agent retrieves targeted portions of the patient's chart through bounded, parameterized tools. These tools organize access around clinically meaningful record categories, including demographics and encounters, notes, laboratory results, vital signs, medications, microbiology, procedures, imaging, and orders, rather than around MIMIC-IV's physical table structure. Second, the environment preserves the heterogeneity of the source record. Data are neither pre-joined nor pre-summarized, requiring the agent to reconcile structured events with free-text documentation and reconstruct the relevant clinical timeline. Third, every tool invocation is logged and replayable, enabling the benchmark to inspect which evidence the agent retrieved and whether its final report is supported by that evidence.

Together, these properties make the agent's path through the record, not just its final answer, part of what the benchmark measures. The tool surface comprises the following sets of tools:

- **Clinical EHR Tools:** Parameterized, clinician-verified tools which support common record-retrieval operations across the MIMIC database, including patient and encounter discovery, note retrieval and search, and longitudinal access to laboratory results, vital signs, medication administration, microbiology, procedures, imaging, and provider orders. These tools constitute the primary interface to the clinical record and organize access around clinically meaningful entities rather than the underlying database schema.
- **SQL Fallback:** A capped, read-only SQL interface retained to address coverage gaps in the higher-level clinical tools. Its use is logged separately so that the benchmark can quantify when an agent relies on the underlying database schema rather than the intended clinical interface.
- **Shell Execution:** A sandboxed execution environment with preconfigured analytical libraries, including `pandas`, `numpy`, and `scipy`. Clinical adjudication frequently requires derived quantities, like a computed severity score, an event interval, or a trended value, which are not stored in the record. The sandbox allows the agent to compute these quantities explicitly and reproducibly, with commands, outputs, and generated artifacts retained as part of the run trace for inspection.
- **Policy-Document Tools:** A lightweight file-access interface for listing, searching, and reading documents in the given policy corpus. Retrieved passages retain document and line-level provenance, enabling policy-dependent claims to be traced to the source text. The policy corpus and grounding procedure are described in Section 3.3.

3.3 Policy and Guideline Grounding

Many clinical audit and decision-support cases require context that extends well beyond the patient record. While EHR artifacts capture what occurred during an encounter, evaluating and contextualizing that data requires an external standard. The evaluation may involve determining whether an action was appropriate, timely, or compliant; establishing the diagnostic or laboratory thresholds a determination hinges on; or applying a governing classification, coding, or quality-measure definition. Many CLINICARE-BENCH cases are therefore not purely factual questions about the EHR, but also require

patient-level evidence to be interpreted under an applicable clinical, regulatory, or operational standard.

We refer to this standard as a **governing policy**. A governing policy may be an institutional protocol, a professional-society guideline, a regulatory or payer quality measure, or another formally adopted clinical standard. These policies define the thresholds, timing requirements, exception criteria, and decision rules against which patient-level evidence must be interpreted. For example, a sepsis audit may depend on the inclusion and timing criteria specified by CMS SEP-1 [16], and an anticoagulation-reversal audit may require comparison against the American College of Cardiology's expert-consensus treatment recommendations [42]. Without access to the applicable policy, an agent may retrieve the patient data correctly yet still reach the wrong verdict, applying generic, outdated, or contextually inappropriate reasoning. The governing policy is what supplies the decision criteria the evidence is judged against. Therefore, policy grounding is not an auxiliary retrieval step but part of the adjudication itself, allowing the benchmark to distinguish a policy-supported verdict from one based on unsupported assumptions or coincidentally correct reasoning.

To evaluate this capability, CLINICARE-BENCH includes a policy-grounding component that tests whether an agent can identify, retrieve, cite, and correctly apply authoritative documentary evidence alongside the clinical record. Agents are given a fixed corpus of policy documents curated with input from the Clinical Board. Since the corpus is shared across all cases rather than filtered into case-specific subsets, the agent must itself determine which document governs the question and locate the relevant provisions. For evaluation, each scenario specifies its own governing document or document set. The documents are converted from PDF to Markdown and exposed through the policy-document tools described earlier, which list, search, and read document text and section metadata. Retrieved passages retain document- and line-level provenance, so each policy-dependent claim can be traced back to specific source text.

This makes the incorporation of governing standards into the agent's adjudication both reproducible and clinician-inspectable, and it mirrors a practical requirement of clinical AI deployment, in which an agent must not only retrieve relevant patient information but also identify which policy applies and make its interpretation auditable. Policy grounding therefore complements EHR grounding, connecting observed clinical events to the standards against which those events are judged.

4 Scenario and Case Construction

The design of CLINICARE-BENCH begins with the construction of *clinical scenarios*. A scenario translates a clinically meaningful audit or chart-review question into a standardized evaluation specification: it defines the question presented to the agent, the applicable clinical or policy criteria, the evidence required for adjudication, the conditions associated with each verdict, and the reasoning procedures and shortcuts that should be rewarded or penalized.

Each scenario centers on a clinical proposition that could in principle be affirmed or rejected, but that the benchmark adjudicates with one of four verdicts: *Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*. The first two indicate that the available record supports a definitive adjudication under the scenario criteria. *Indeterminate: Lack of Data* applies when evidence required for adjudication is absent or cannot be reliably established from the record. *Indeterminate: Medically Ambiguous* applies when the relevant evidence is available but does not support a unique, clinically defensible interpretation. This shared output space makes abstention part of the evaluation standard.

Since it is standardized, a scenario specification applies consistently across heterogeneous patient records.

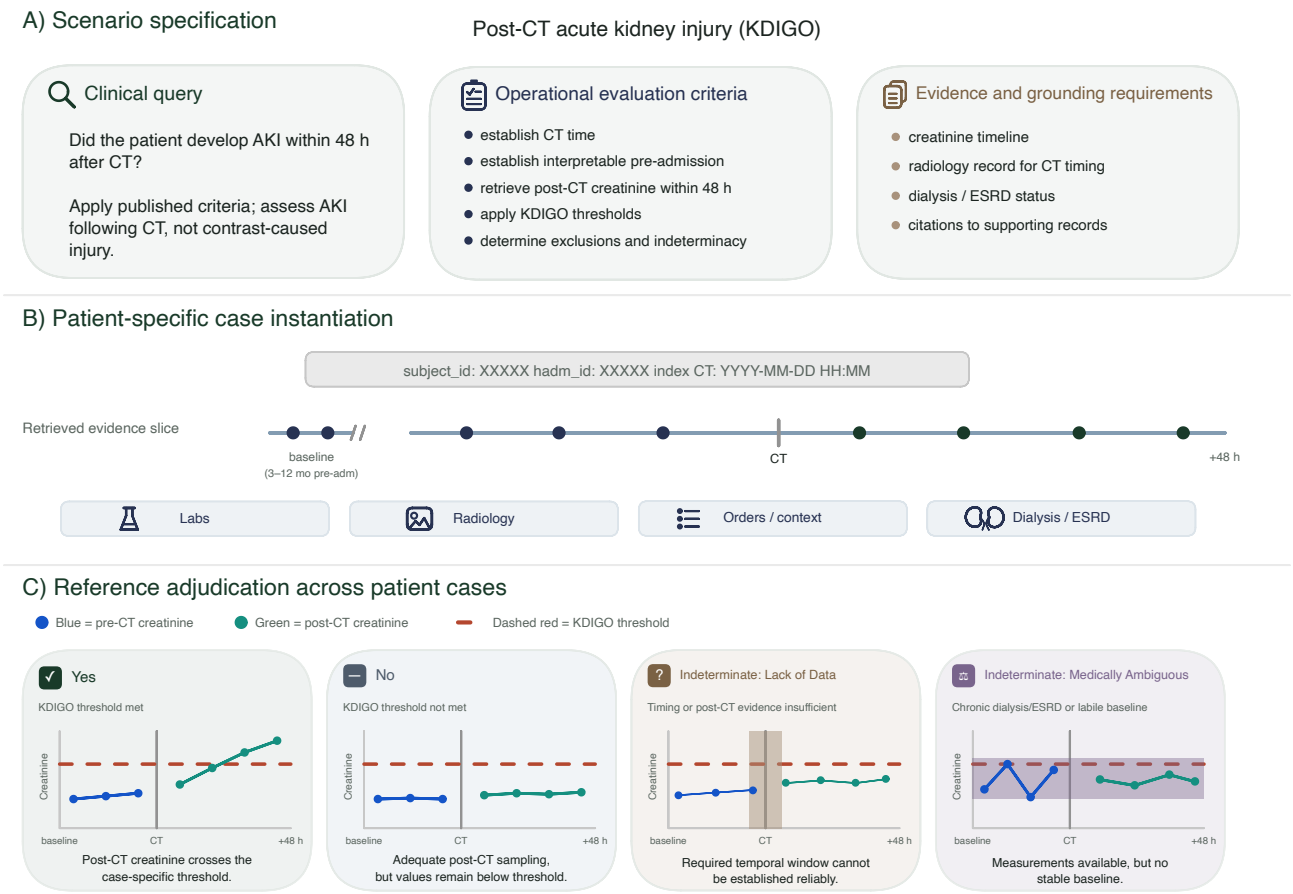

Figure 2. From scenario specification to patient-specific adjudication. (A) The post-CT acute kidney injury (AKI) scenario defines its clinical query, operational evaluation criteria, and evidence and grounding requirements. **(B)** These are instantiated for a specific case (shown schematically) from the encounter, CT event, and retrieved EHR evidence. **(C)** The same KDIGO serum-creatinine criteria yield one of four reference verdicts—*Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*. Blue and green mark pre- and post-CT creatinine; the dashed red line is the case-specific KDIGO threshold ($\geq$0.3 mg/dL or $\geq$1.5$\times$ baseline).

As such, each scenario is instantiated over a cohort of eligible MIMIC-IV records to create *patient-specific cases*. CLINICARE-BENCH contains 30 cases per scenario, for a total of 750 cases. The case-level reference verdicts are produced through the clinician-calibrated labeling procedure described in Section 4.5.

The remainder of this section describes the anatomy of a scenario, presents representative scenarios, explains how scenarios are instantiated as patient cases, and details the clinical review and label-calibration procedure. Terminology and units of evaluation are summarized in Appendix A.

4.1 Scenario Development

Our suite comprises *25 clinician-developed and validated scenarios* spanning recurring clinical audit and review workflows. The suite was designed to provide complementary coverage across both reasoning capabilities and medical specialties. Its reasoning demands span ten capabilities, enumerated in Figure 3, from clinical-score and threshold computation and time-sensitive protocol auditing to cross-modal validation between documentation and structured data, negative and absence reasoning, and longitudinal

synthesis across encounters.

Rather than defining an ad hoc domain grouping, we label each scenario with a single primary clinical specialty drawn from the American Board of Medical Specialties (ABMS) taxonomy, which defines 40 specialties and 87 subspecialties (127 certificate categories in total).[2] The 25 scenarios span 14 specialties: nephrology, infectious disease, critical care medicine, pulmonary disease, cardiovascular disease, neurology (including vascular neurology), endocrinology, gastroenterology and hepatology, hematology, transfusion medicine, medical toxicology, hospice and palliative medicine, general surgery, and radiology. Four scenarios (discharge-audit, readmission-30d, med-reconciliation, and frequent-admitter) are quality, documentation, and care-coordination workflows not owned by a single specialty. We group these as *cross-specialty operations*. Specialty is a single-label axis, where each scenario has exactly one primary specialty, in contrast to the reasoning-skill axis, which is non-exclusive, since a scenario may exercise several reasoning capabilities.

Figure 3 gives the full coverage matrix along both axes. Because specialty is single-label while reasoning skill is non-exclusive, the two views are complementary: the specialty axis conveys clinical breadth, whereas the skill axis shows that the same reasoning capability recurs across specialties rather than being tied to one organ system or workflow. Full specifications for four representative scenarios appear in Appendix B, and Table 13 highlights a representative subset.

4.2 Anatomy of a Scenario

Each scenario is defined by the following components:

- **Clinical Query:** A concise question expressed in the natural language familiar to the intended clinical user. The query specifies the clinical objective while leaving the detailed retrieval and analysis strategy to the agent. At the scenario level, the query contains placeholders such as `subject_id`, `hadm_id`, or `stay_id`, which are populated when the scenario is instantiated as a patient-specific case.
- **Operational Evaluation Criteria:** A detailed specification of the decision rules that map the available patient record to one of four verdicts: *Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*. These criteria define the applicable thresholds, temporal windows, exclusions, exception rules, and conditions under which the record is insufficient or clinically ambiguous.
- **Evidence and Grounding Requirements:** A specification of the evidence needed to support adjudication, including the relevant record sources, clinical variables, temporal windows, and the governing policy documents. These requirements define the decisive findings that the agent must retrieve and cite, or explicitly establish as unavailable, in order to justify its verdict. Three of these requirements are recorded as explicit per-scenario lists, and serve as the reference sets for the grounding metrics of Sections 5.3 and 5.4: the *required record sources*, three to nine MIMIC-IV tables per scenario over a vocabulary of 23 tables; the *required findings*, four to eleven distinct content dimensions per scenario; and the *governing documents*, one to four per scenario drawn from the policy corpus of Section 3.3.

The clinical query is presented to the agent, while the operational criteria are applied through the labeling procedure to produce case-level reference verdicts. The evidence and grounding requirements define the information needed to support adjudication and enable evaluation beyond verdict correctness. Together, these components define a standardized four-way clinical adjudication problem with an explicitly specified evidentiary basis.

[2] https://www.abms.org/member-boards/specialty-subspecialty-certificates/

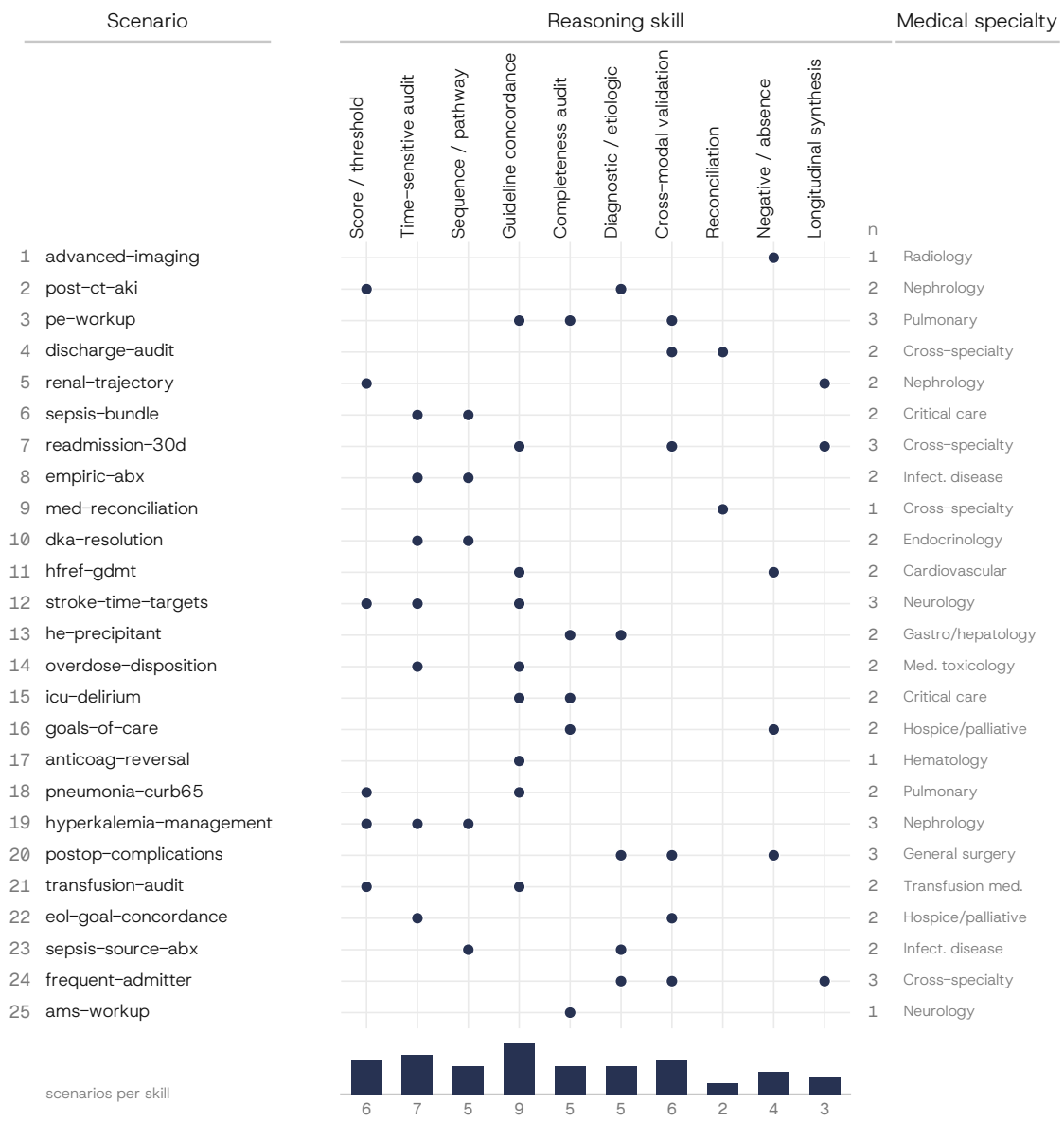

Figure 3. Reasoning-skill and specialty coverage across the 25 CliniCARE-Bench scenarios. Each row is one scenario, and a filled dot marks a reasoning skill that the scenario exercises. Most scenarios probe two or three skills at once (n column = skills per scenario), so the benchmark targets compound clinical reasoning rather than isolated skills. The bars below each column count the scenarios exercising that skill. Decision-tree / guideline concordance (9) and time-sensitive protocol audit (7) are the most widely sampled, whereas discrepancy / reconciliation (2) and longitudinal synthesis (3) are rarer, higher-difficulty probes. The final column gives each scenario's primary ABMS specialty. Because a given reasoning skill recurs across many specialties, the skill and specialty axes are near-orthogonal, so the same capability is not tied to one organ system or workflow.

4.3 Representative Scenarios

To make the scenario structure concrete, we summarize one representative scenario below. Figure 2 illustrates how the scenario is instantiated and adjudicated for an individual patient case. The complete operational specification for this scenario, clinical query, four-way adjudication criteria, required evi-

dence, grounding requirements, and adversarial controls, together with three additional fully annotated anchor scenarios, is provided in Appendix B. These per-case specifications are a central contribution of this work.

Post-CT Acute Kidney Injury. This scenario asks whether an ICU patient developed acute kidney injury within 48 hours after a CT scan, applying the published KDIGO criteria. Baseline creatinine is set by a three-tier hierarchy: the lowest value 3–12 months before admission, else the admission value, else an MDRD estimate at eGFR 75 mL/min/1.73 m^2 where CKD is documented.

- **Clinical Query:** "For ICU patient `<subject_id, hadm_id>` who underwent a CT this admission, did the patient develop AKI within 48 h after the CT? [...] Apply the KDIGO serum-creatinine criteria and frame the result as AKI following CT, not as contrast-caused injury. Where serum creatinine does not reflect the patient's own kidney function, say the question does not resolve rather than applying the thresholds." The elided passage states the baseline hierarchy above; the full prompt appears in Appendix B.
- **Adjudication Criteria:** A *qualifying rise* is any post-CT creatinine at least 0.3 mg/dL above baseline or at least 1.5× baseline; every value in the window is searched, not only the first.
  - *Yes:* a baseline is set, at least one creatinine falls within 48 h post-CT, there is a qualifying rise, and the patient is not on chronic dialysis and has no ESRD.
  - *No:* a baseline is set, at least two creatinines fall within 48 h post-CT, there is no qualifying rise, and the patient is not on chronic renal replacement therapy and has no ESRD or CKD stage ≥ 4.
  - *Indeterminate: Lack of Data:* no tier of the baseline hierarchy can be set; or no creatinine falls within 48 h post-CT, or fewer than two, leaving the peak unconfirmed; or CT timing cannot be resolved (order versus actual scan time, or no unique first qualifying CT); or no CT can be identified in the admission, so the premise is unmet and the case is out of scope.
  - *Indeterminate: Medically Ambiguous:* chronic dialysis or ESRD, so the KDIGO acute thresholds do not apply; or a labile pre-admission baseline, so no clean reference can be set and a $+0.3$ mg/dL rise cannot be confidently attributed.
- **Required Evidence:** the creatinine timeline from the baseline source through 48 h post-CT, showing every value; ESRD and dialysis status; and the CT *scan* time distinguished from the order time, confirmed against the radiology report where possible.
- **Required Grounding:** The specific MIMIC-IV record or record span from which the decisive finding was derived.

4.4 From Scenarios to Patient Cases

To instantiate a scenario, we first construct a scenario-specific candidate cohort from MIMIC-IV. Broad eligibility criteria identify records containing the encounters, index events, and evidence sources relevant to the clinical question. Because a case may be anchored to a particular admission, ICU stay, procedure, or other clinical event, case eligibility is defined at the level appropriate to the scenario.

We then apply scenario-specific heuristics to assign each eligible record to a provisional sampling stratum. These heuristics are used only for cohort construction and do not determine the final reference verdict. We sample 30 patients per scenario using stratification across *Yes*, *No*, and *Indeterminate* candidate strata,

with the Indeterminate stratum including cases provisionally associated with either lack of data or medical ambiguity. This intentional balancing prevents aggregate performance from being dominated by straightforward positive or negative cases and increases coverage of scenario-specific challenges, including missing evidence streams, conflicting documentation, ambiguous temporal attribution, and potentially misleading administrative codes. As such, the resulting case distribution is designed for evaluation and should not be interpreted as an estimate of the clinical prevalence of these outcomes.

Case-level reference verdicts are then produced through a model-assisted labeling procedure. An ensemble of three frontier model harness configurations from separate families, namely `Claude Code` Opus 4.8, `Codex` GPT-5.5, and `Gemini CLI` Gemini 3.1 Pro, independently evaluate each case with access to the complete scenario specification, including the clinical query, operational evaluation criteria, and evidence and grounding requirements. Their outputs are aggregated to produce the reference verdict. Across the 750 cases the agents reported unanimously on 556, split 2-vs-1 on 183, and produced three distinct verdicts on 11, with only the last set being escalated to clinicians for direct adjudication. This process is validated through the Clinical Board procedure described in Section 4.5.

4.5 Clinical Review and Label Calibration

Clinician review through our Clinical Board was used to validate each step of the pipeline. All Clinical Board reviewers are credentialed PhysioNet users who completed the required training and executed the PhysioNet Credentialed Health Data Use Agreement before reviewing any MIMIC-derived evidence. This review proceeded in two stages: first reviewing the construction of the 25 scenarios, then validating a sample of the model-generated reference verdicts.

Scenario review: two clinicians independently vetted every scenario, confirming the query reflects a real chart-review workflow, estimating manual adjudication time, and revising the four-way criteria (thresholds, windows, exclusions, ambiguity conditions, and evidence requirements). Each decision rule was tagged by provenance (i.e., named guideline or policy, established clinical knowledge, or benchmark-specific operational choice), making explicit which parts of the adjudication are externally governed. These edits were applied before the ensemble was run.

Verdict calibration: After verdicts were generated, a sample of them was shown to clinicians to validate the generation process. In this review, clinicians were given the scenario specification and the three ensemble reports in randomized order as Alpha, Beta, and Gamma to remove model-identity bias, then asked to return their own four-way verdict with a confidence and rationale. We note that this is not truly independent, as the reviewer uses the model-surfaced evidence to make their judgment instead of interfacing with the patient record themselves. This design is intentional: we rely on the models for the SQL and data-extraction work while reserving for the Clinical Board the medical judgment of interpreting that evidence. The reviewer also rated each report as *sound*, *minor error*, or *major error*, tagging errors by category (missed evidence, misapplied criteria, temporal or attribution error, unsupported inference, or incorrect abstention).

We ran verdict calibration at two scales: a 75-case sample (10%; three cases per scenario) with one reviewer each for verdict-level calibration, and a 25-scenario subset (one case each) with two independent reviewers, with no reviewer seeing the same case twice, to measure reproducibility. On the 75-case sample, the reference verdict matched the clinician verdict on 91% of cases with a defined label (64/70; Cohen's $\kappa = 0.87$). The five no-majority cases were escalated by design. Agreement was perfect when the ensemble was unanimous (46/46), and fell to 75% (18/24) on 2-vs-1 cases. Every majority-case disagreement resolved to the dissenting model rather than an outside label, with expert judgment staying

within the ensemble's range.

The double-reviewed subset places these figures against the reproducibility of expert judgment itself. The two reviewers agreed with each other on 80% of verdicts ($\kappa = 0.69$), while the reference agreed more closely with each of them individually ($\kappa = 0.87$ and 0.81): the model-assisted label sits closer to each clinician than the two clinicians sit to one another. Agreement again tracked ensemble consensus, and reviewer-reviewer agreement fell fastest on the non-unanimous cases. Non-unanimity therefore marks the cases where expert judgment is itself least reproducible, supporting their escalation. Reviewers rated most reports *sound*: 73% of the 225 report ratings on the 75-case sample and 81% of the 150 on the double-reviewed subset, with the remainder split between minor and major errors. These ratings were consistent across reviewers (Gwet's AC1 ≈ 0.70; Cohen's κ understates agreement here because *sound* dominates the marginal). Among the errors reviewers did flag, misapplied criteria was the most common category (38% and 53% of tags) ahead of missed evidence (24% in both), locating the ensemble's failures in applying a scenario's decision rules rather than in surfacing the underlying evidence.

5 Evaluation Framework

CLINICARE-BENCH evaluates clinical agentic systems using both their final outputs and their observable interaction traces. Each run produces a final report, a logged sequence of tool calls, the retrieved patient and policy evidence, and any computations or intermediate artifacts generated during the investigation. These artifacts are evaluated against the case-level reference verdict and scenario-specific evaluation metadata across various complementary dimensions, including verdict correctness, abstention behavior, patient-evidence grounding, policy grounding, process adherence, and reliability and resource use.

5.1 Units of evaluation

A *run* is one execution of a system on one case and the unit at which metrics are scored. In our primary protocol each case is run once, giving a one-to-one correspondence between cases and runs, so per-run accuracy is computed over all 750 runs. Under k repeated attempts a case contributes k runs, indexed by an *attempt* number and used for reliability metrics such as avg@k and pass^k. A *system* is the complete evaluated configuration, including the model, agent harness, prompting or scaffolding, and available tools. It is the entity that a run executes and that the benchmark ultimately characterizes.

Not every metric applies to every run. Verdict accuracy retains the entire set of runs. A *report-present run* is one in which the system returns a parseable final report, and it forms the denominator for process adherence, defect-free accuracy, and the defect gap. A *grounding-scoreable run* is a report-present run in which the system retrieved at least one patient record, and the four patient-grounding metrics of Section 5.3 are averaged over this set.

5.2 Verdict and Abstention Metrics

Let N denote the number of evaluated runs, indexed by $i \in \{1, \ldots, N\}$. Let y_i denote the case-level reference verdict for run i, and let $\hat{y}_i$ denote the predicted verdict.

Four-way verdict performance. Verdict accuracy measures exact agreement between $\hat{y}_i$ and y_i across four possible verdicts: *Yes*, *No*, *Indeterminate: Lack of Data*, or *Indeterminate: Medically Ambiguous*. Verdicts are scored by exact four-class match, with no partial credit. Missing or unparseable verdicts are counted as incorrect rather than excluded from the denominator. We report four-class accuracy and macro-F1

across the four verdict classes, using Macro-F1 to give equal weight to the two less frequent, safety-critical *Indeterminate* classes and the definitive *Yes* and *No* classes.

Abstention behavior. We separately characterize whether the system defers appropriately. Let

$$\mathcal{D} = \{\textit{Yes}, \textit{No}\}$$

denote the set of definitive verdicts, and let

$$\mathcal{A} = \{\textit{Indeterminate: Lack of Data}, \textit{Indeterminate: Medically Ambiguous}\}$$

denote the set of abstention verdicts.

Let

$$N_{\mathcal{A}} = \sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{A}] \qquad \text{and} \qquad N_{\mathcal{D}} = \sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{D}]$$

denote the numbers of reference-abstention and reference-definitive runs, respectively.

The *over-commitment rate* is the fraction of reference-abstention runs for which the system returns a definitive verdict:

$$\text{OverCommit} = \frac{1}{N_{\mathcal{A}}} \sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{A}]\mathbf{1}[\hat{y}_i \in \mathcal{D}].$$

The *over-abstention rate* is the fraction of reference-definitive runs for which the system returns an abstention verdict:

$$\text{OverAbstain} = \frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{D}]\mathbf{1}[\hat{y}_i \in \mathcal{A}].$$

5.3 Patient-Evidence Grounding

Patient-evidence grounding evaluates whether the system's clinical claims are supported by the patient record and whether the investigation retrieves and reports the evidence required for adjudication. We evaluate grounding along two complementary axes: *citation precision*, which measures the faithfulness of the evidence cited in the report, and *evidence coverage*, which measures the completeness of evidence acquisition and reporting. Each axis pairs a referential check, decidable from the trace by rule, with a substantive one that requires a semantic judgment about the report's content.

Citation precision. We calculate two types of citation precision. *Record-identifier precision* P_{record} is the fraction of record-level identifiers (e.g. patient, admission, ICU-stay, and note identifiers) cited within the final report that occur in evidence retrieved during the same run. *Claim-level precision* P_{claim} then checks whether the evidence a claim cites substantively supports it. This distinction is necessary because a citation may refer to a real retrieved record yet still be irrelevant to, inconsistent with, or insufficient to support the associated claim.

Claim-level precision is computed by decomposing the report into atomic clinical claims, such as measured laboratory and vital-sign values, medication doses, imaging and procedure findings, temporal relationships, and the presence or absence of clinical events. Each claim is classified by an LLM against the evidence it cites as *supported*, *contradicted*, or *unverifiable*. P_{claim} is the supported fraction. A claim is supported only when the cited evidence substantiates it with the correct patient and encounter attribution and a consistent temporal context; claims that the cited evidence contradicts, that cannot be verified from it, or that carry no resolvable citation are not.

Evidence coverage. Similarly, coverage is assessed at two points where required evidence can be lost. *Retrieval coverage* $R_{\text{retrieval}}$ is the fraction of a case's required record sources that the agent queried with the correct patient and, where applicable, encounter and temporal scope. Here, required sources are the MIMIC-IV tables named in the evidence specification of the case's scenario (Section 4.2), three to nine per scenario. A source counts as consulted when a correctly scoped query is issued, whether it returns records or a valid empty result, because an empty result may itself establish that the evidence is unavailable, a conclusion several scenarios require.

Finding coverage R_{finding} is the fraction of a case's required findings that the report addresses, either by presenting the relevant evidence or by explicitly asserting its absence where that is what the case requires. Required findings are drawn from the same scenario specification, four to eleven per scenario, and are compared to the report using an LLM judge. These are, again, different failure modes: an agent may query every required source and still not establish the finding its verdict rests on.

5.4 Policy Grounding

Policy grounding is scored separately from patient-evidence grounding, and asks whether a report's use of the governing standard is verifiable: whether its policy citations resolve to real passages that support the claims they carry, and whether the report identifies the standards that actually govern the case. Every scenario carries at least one governing document, so these metrics are computed across the whole suite. Whether the agent then applies the policy's thresholds, timing requirements, exclusions, and exceptions correctly is not scored here; that is captured by verdict correctness and by the process rubric of Section 5.5. As in Section 5.3, each axis pairs a referential check with a substantive one.

Citation correctness. Agents cite policy inline as `[policy: document L_a-L_b]`, naming a corpus document and a line span within it. A citation *resolves* when that document is present in the corpus and the span is valid, and the *citation-resolution rate* is the fraction of a run's policy citations that resolve. Citations naming an absent document or an invalid span do not resolve; because a single run can emit many, we also report unresolved citations as raw counts. *Citation support* then asks, for each resolved citation, whether the cited passage substantiates the particular policy-dependent claim it is attached to, rather than merely discussing a related topic. Resolution is computed by rule, while support is LLM-judged.

Governing-document coverage. For run i, let $\mathcal{G}_i$ denote the scenario's governing-document set and $\hat{\mathcal{G}}_i$ the set of corpus documents that the report cites. *Document precision* P_{doc} is the fraction of $\hat{\mathcal{G}}_i$ lying in $\mathcal{G}_i$, and *document recall* R_{doc} is the fraction of $\mathcal{G}_i$ that the report cites; both are set comparisons against an expert-specified reference and require no judge. Precision asks whether the documents a report invokes govern the case at all, recall whether it found all of them, and the two can diverge sharply: a report resting on one applicable standard while omitting the others scores high precision at low recall. The two are averaged over different sets of runs, since precision is undefined for a run that cites no corpus document and that run is omitted from it, whereas recall is defined for every run and scores zero.

5.5 Process Adherence

Verdict correctness and grounding do not fully establish that an agent's investigation was defensible. An agent may reach the correct verdict through an inappropriate shortcut, by omitting a verification the case turns on, or by making an unsupported inference that happens to agree with the reference verdict.

Conversely, it may conduct a sound investigation and still reach the wrong verdict through a localized interpretation error. We therefore score the agent's *observable investigation process* separately from its outcome. Here, observable means we use the agent's final report together with its logged tool-call trajectory; we make no attempt to inspect or score the model's private reasoning.

The process rubric. Each scenario carries a *process rubric*, authored by the Clinical Board and instantiated unchanged for every case of that scenario. MUST-DO criteria state the actions and checks required for a defensible adjudication. MUST-NOT criteria state prohibited shortcuts and unsupported inferences, such as inferring an endpoint from an administrative code alone, treating an order time as a performed-procedure time, or asserting a causal relationship the record does not support. We call a violated MUST-NOT criterion a *process defect*. Rubrics carry four to nine MUST-DO and three to eight MUST-NOT criteria per scenario, each weighted ± 1 or ± 2 by the authoring clinicians.

Criteria are graded by an LLM judge against the final report and the logged trajectory together. A methodology claim in the report is credited only when the trajectory corroborates it; where the trajectory shows the agent did something other than what the report describes, the criterion is graded on what the agent actually did.

Process-adherence score. For run i, let $\mathcal{M}_i$ denote the set of applicable criteria, each carrying a signed weight w_m that is positive for MUST-DO and negative for MUST-NOT, and let $z_{im} = 1$ when the judge grades criterion m as met, meaning the required action was performed or the prohibited one committed. The *process-adherence score*, abbreviated *process score* in the results, is

$$S^{(i)}_{\text{process}} = \max\left(0,\ \frac{\sum_{m \in \mathcal{M}_i} w_m z_{im}}{\sum_{m \in \mathcal{M}_i : w_m > 0} w_m}\right),$$

so credit accrues against the MUST-DO budget while sprung MUST-NOT criteria subtract from it. The denominator is the sum of positive weights alone, between 7 and 16 points depending on the scenario. The floor at zero is not vacuous: in one scenario the available penalties exceed the entire positive budget. For a system we report the mean of $S^{(i)}_{\text{process}}$ over report-present runs.

Defect-free accuracy. A run may reach the correct verdict while springing a process defect. We therefore record for each run whether any applicable MUST-NOT criterion was violated,

$$\tau_i = \mathbf{1}\left[\sum_{m \in \mathcal{M}_i : w_m < 0} z_{im} > 0\right],$$

and report a *defect-free accuracy*, which credits a run only when its verdict is correct and $\tau_i = 0$, together with the *defect gap* this opens against accuracy on the same report-present denominator.

5.6 Repeated-Run Reliability

A single-attempt score conflates whether a system can reach the correct verdict with whether it does so consistently. For a cohort evaluated over k separately executed attempts, let $\hat{y}_{i,r}$ denote the verdict returned for case i on attempt r. Because pass^k is undefined for a case that lacks a gradeable verdict in some attempt, the reliability metrics are computed over the n cases gradeable in every attempt.

We report avg@k, the mean verdict accuracy across all nk runs, together with the standard deviation of the k per-attempt accuracies, which distinguishes run-to-run noise from systematic error. We additionally report pass^k, the proportion of cases answered correctly on every attempt:

$$\text{pass}^k = \frac{1}{n} \sum_{i=1}^{n} \prod_{r=1}^{k} \mathbf{1} \left[\hat{y}_{i,r} = y_i \right].$$

For comparison, pass@k is the proportion answered correctly on at least one attempt. The two bound reliability from either side: pass@k measures whether repeated attempts ever produce the correct verdict, pass^k whether they always do.

Neither separates a system that varies from one that is consistently wrong, so we also report *verdict agreement*, the proportion of cases on which all k attempts return the same verdict, correct or not. Agreement decomposes as pass^k plus the proportion of cases whose attempts agree on an *incorrect* verdict; it is this second term that identifies a system reproducing an error rather than resolving the case.

5.7 Metric Adjudication and Judge Reliability

The framework pairs deterministic checks with LLM adjudication, and each metric above states which it uses. The division is consistent: checks that can be settled referentially, against the trace or the policy corpus, are computed by rule, while judgments about whether evidence supports a claim, whether a required finding is addressed, or whether a process criterion holds are delegated to a judge. The production judge is GPT-5.5, which is itself among the evaluated systems, so Section 6.7 examines the resulting dependencies. Within each metric the judge model, prompt, and rubric are held fixed across evaluated systems, so a difference in score reflects the systems rather than the grader.

Each judged component receives only the evidence its judgment requires: the final report, together with the cited patient evidence, the cited policy passage, or the logged trajectory as applicable. Trajectories are rendered for judging rather than passed verbatim. Each tool result is truncated to 12,000 characters and the duplicate copies that some harness trajectory formats emit are dropped; where report and trajectory together still exceed roughly 500,000 characters, they are graded in sequential chunks and the per-criterion grades combined disjunctively, so a criterion counts as met if any chunk supports it. The report always occupies the first chunk, so criteria decided from the report alone are unaffected.

We assess the reliability of the judged metrics by regrading identical outputs and, for process and policy scoring, by repeated evaluation with a panel of judges. Section 6.7 reports criterion-level disagreement, within-judge variation, and chance-corrected agreement statistics. Criteria whose grades prove persistently unstable are returned for expert revision rather than treated as stable measurements.

5.8 Resource Use and Cost

For each run we record the resources the investigation consumed: the number of agent turns, the number of typed tool calls, token volume, and monetary cost, taken from the harness where it reports one and computed from measured tokens otherwise. Token volume is summarized as *work-tokens*, the prompt tokens not served from cache plus the completion tokens, which approximates the compute actually purchased for the run. We report per-run medians, and cost as a mean across runs excluding the spend on judging.

None of these quantities is comparable across harness families without qualification, so we treat each as an index of effort rather than of efficiency. An agent turn is not a common unit, because harnesses batch

differently: a single step may carry many tool calls in one harness and roughly one in another. Typed tool calls are undercounted for harnesses that reach the tool server through a shell or a code-execution step, since those calls are never typed. Work-tokens depend on what a serving path reports about cache reads, and a path reporting none inflates the figure; where the cache correction must be estimated we give a range rather than a point value. We therefore state the provenance of each figure and compare within a serving path rather than across accounting regimes.

6 Experimental Results

We evaluate sixteen systems: ten production agent CLIs on their native vendor models, and six open-weight models on a single model-agnostic harness. Using the metrics of Section 5, the experiments target three dissociations the benchmark is built to expose. First, raw accuracy can mask process defects because a correct verdict may be reached through a prohibited shortcut. Second, systems systematically under-abstain, committing to a definitive answer where the record warrants deferral. Third, verdict accuracy and grounding quality dissociate: models reach the right verdict without citing the evidence that justifies it.

6.1 Experimental Setup

Harnesses and models. We run four agent harnesses. `Claude Code`, `Codex` and `Gemini CLI` are production agent CLIs, each locked to its own vendor model family and driving the benchmark tools through a native tool-calling loop, so they represent how each model is actually deployed. `opencode` is a model-agnostic open-source harness used to drive each of the six open-weight models. All four connect to the MIMIC-IV environment through an MCP interface and run with web search and fetch disabled, so every system answers from the in-environment patient record and policy corpus alone. MIMIC data is reached only through a credential-isolated tool server, so no system can bypass the governed tools to read raw records.

Table 2 lists the sixteen systems evaluated. Section 6.6 separately evaluates paired changes to the harness and model configuration.

Table 2. Systems evaluated in CliniCARE-Bench. Native leaderboard (each harness on its own frontier model) and the open-weight model comparison on the neutral `opencode` harness.

Harness	Models
Native leaderboard: each harness on its own frontier model	
`Claude Code`	Claude Opus 5, Claude Sonnet 5
`Codex`	GPT-5.6 Sol, GPT-5.6 Luna, GPT-5.5, GPT-5.4, GPT-5.4-mini
`Gemini CLI`	Gemini 3.1 Pro, Gemini 3.5 Flash, Gemini 3.6 Flash
Controlled comparison: open-weight models on one neutral harness	
`opencode`	DeepSeek V4 Pro, DeepSeek V4 Flash, Qwen 3.7 Plus, GLM 5.2, MiniMax M3, Kimi K2.7 Code

Metrics. Unless noted otherwise, all metrics follow Section 5. We abbreviate the two indeterminate classes, Lack of Data and Medical Ambiguity, as *lack* and *amb* respectively. Accuracy is reported per run (micro); for a system evaluated on all 750 cases, this equals the scenario-macro average, but the two

differ for partial runs. We additionally report expected calibration error (ECE), computed from the stated confidence included in each parseable report and defined in Eq. 1.

6.2 Results

Table 3 reports the primary leaderboard on all 750 cases. We analyze the three column groups in turn below, quality, abstention & calibration, and resource use & cost.

Table 3. Primary leaderboard. Each system is evaluated once per case. *Acc.* is four-class verdict accuracy. *Def-free* is defect-free accuracy and *Gap* the defect gap over report-present runs. *Proc* is the process rubric score, *Over-com*/*Over-abs* the directional abstention errors, and *ECE* expected calibration error. All values are percentages except *Mac-F1* and *ECE*, which are on a 0–1 scale, and Gap, which is in percentage points. Cost columns are per-run medians except USD/run (a mean, excluding judge spend). `opencode` work-token and USD figures are estimated ranges.

	Quality					Abstention & calibration			Resource use & cost			
System	**Acc**	**Mac-F1**	**Def-free**	**Gap**	**Proc**	**Over-com**	**Over-abs**	**ECE**	**Turns**	**Tools**	**Work-tok**	**USD/run**
`Claude Code`												
Opus 5	75.6	0.701	70.0	5.6	82.0	32.0	10.5	0.064	17	28	130.3k	2.05
Sonnet 5	73.9	0.654	65.6	8.3	70.8	30.7	10.8	0.046	13	22	90.7k	0.92
`Codex`												
GPT-5.6-Sol	73.2	0.664	67.5	5.7	80.0	31.3	12.0	0.194	22	16	88.2k	1.12
GPT-5.6-Luna	66.8	0.588	60.8	6.0	75.2	39.3	16.5	0.229	22	18	98.1k	0.21
GPT-5.5	76.1	0.699	71.3	4.8	78.6	38.7	7.7	0.090	18	48	121.9k	1.51
GPT-5.4	71.1	0.632	65.6	5.5	76.1	36.7	11.2	0.153	21	52	109.0k	0.73
GPT-5.4-mini	66.4	0.559	56.1	10.3	65.2	52.7	12.2	0.215	24	42	84.4k	0.20
`Gemini CLI`												
Gemini-3.6-Flash	72.9	0.652	59.1	13.9	72.6	38.0	7.3	0.251	62	37	389.1k	1.15
Gemini-3.5-Flash	71.9	0.622	60.0	11.9	71.1	42.0	7.5	0.262	62	37	384.9k	1.17
Gemini-3.1-Pro	72.7	0.664	57.9	14.8	56.4	21.3	11.2	0.256	42	23	190.2k	0.95
`opencode` (open-weight models on the neutral harness)												
DeepSeek-V4-Pro	68.7	0.612	58.7	10.0	69.4	40.7	9.7	0.192	14	40	63–582k	0.19–1.11
DeepSeek-V4-Flash	70.7	0.586	58.3	12.4	68.6	52.0	7.8	0.219	15	46	79–746k	0.03–0.12
GLM-5.2	73.6	0.643	64.8	8.8	76.4	42.0	7.5	0.130	14	31	66–605k	0.20–1.08
Qwen-3.7-Plus	65.3	0.579	53.1	12.3	65.9	46.0	14.0	0.187	11	28	47–434k	0.05–0.19
MiniMax-M3	65.9	0.555	54.7	11.2	72.5	55.3	10.5	0.191	21	45	107–995k	0.11–0.39
Kimi-K2.7-Code	66.5	0.558	54.1	12.4	69.6	52.7	10.5	0.156	20	55	130–1245k	0.91–3.27

No system exceeds 76.1% accuracy on the cohort. Accuracy ranges from 65.3% to 76.1% and macro-F1 ranges from 0.555 to 0.701. Macro-F1 is lower than accuracy for every system because it gives equal weight to the two less frequent *Indeterminate* classes, where performance is weakest. The systems form a continuum rather than distinct tiers: no adjacent systems differ by more than 2.0 percentage points across the 10.8-point accuracy range, and the harness families interleave throughout the ranking. GPT-5.4-mini (66.4%) and GPT-5.6-Luna (66.8%) fall below four and three open-weight systems, respectively. The strongest open-weight system, GLM-5.2 at 73.6%, is 2.5 points behind the leading system, GPT-5.5 at 76.1%. Because these differences are small, formal comparisons should use paired case-level tests rather than marginal standard errors for individual accuracies.

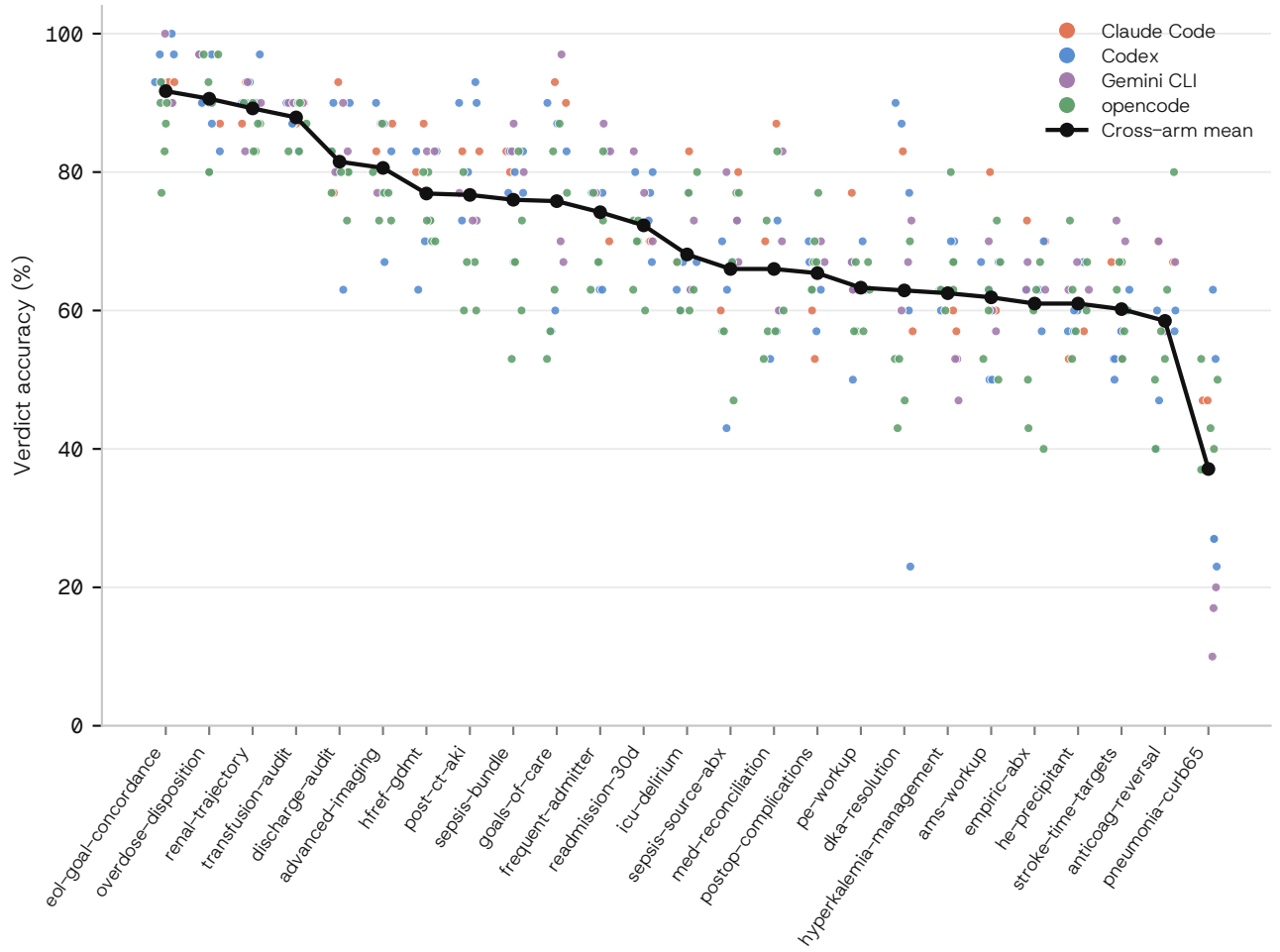

Figure 4. The leaderboard order does not hold scenario by scenario. Verdict accuracy of every system on each of the 25 scenarios, one dot per system colored by harness family, with the cross-system mean (dark line). Scenarios are sorted by that mean. Numbers reported in Table 15.

Scenario by scenario, ranking is unstable and difficulty is concentrated. Resolved per scenario (Figure 4), cohort ordering is reversed in 30.2% of the 3,000 system-pair × scenario comparisons and tied in a further 12.7%, where 30 cases per scenario cannot resolve the difference. Fourteen of the sixteen systems are the best system on at least one scenario, including Kimi-K2.7-Code, fourth from last on the leaderboard, which leads `he-precipitant` outright and ties for the lead on `overdose-disposition`. The ordering of the averages is nevertheless stable: a system's cohort accuracy predicts its mean per-scenario rank almost perfectly (Spearman $\rho = -0.94$, negative because rank 1 is best). The leaderboard is thus a statement about a case mix more than about which system to trust on a given clinical question.

Furthermore, scenario difficulty is concentrated in a few cases: cross-system mean accuracy runs from 91.7% on `eol-goal-concordance` to 37.1% on `pneumonia-curb65`, and the five hardest scenarios carry 30% of all errors against the 20% that uniform difficulty implies. Still, systems range from 10% to 63% on the hardest scenario, a 53 point spread, and the widest spread in the grid belongs to `dka-resolution` (23% to 90%), a scenario of middling average difficulty, suggesting that difficulty and discrimination are not the same axis.

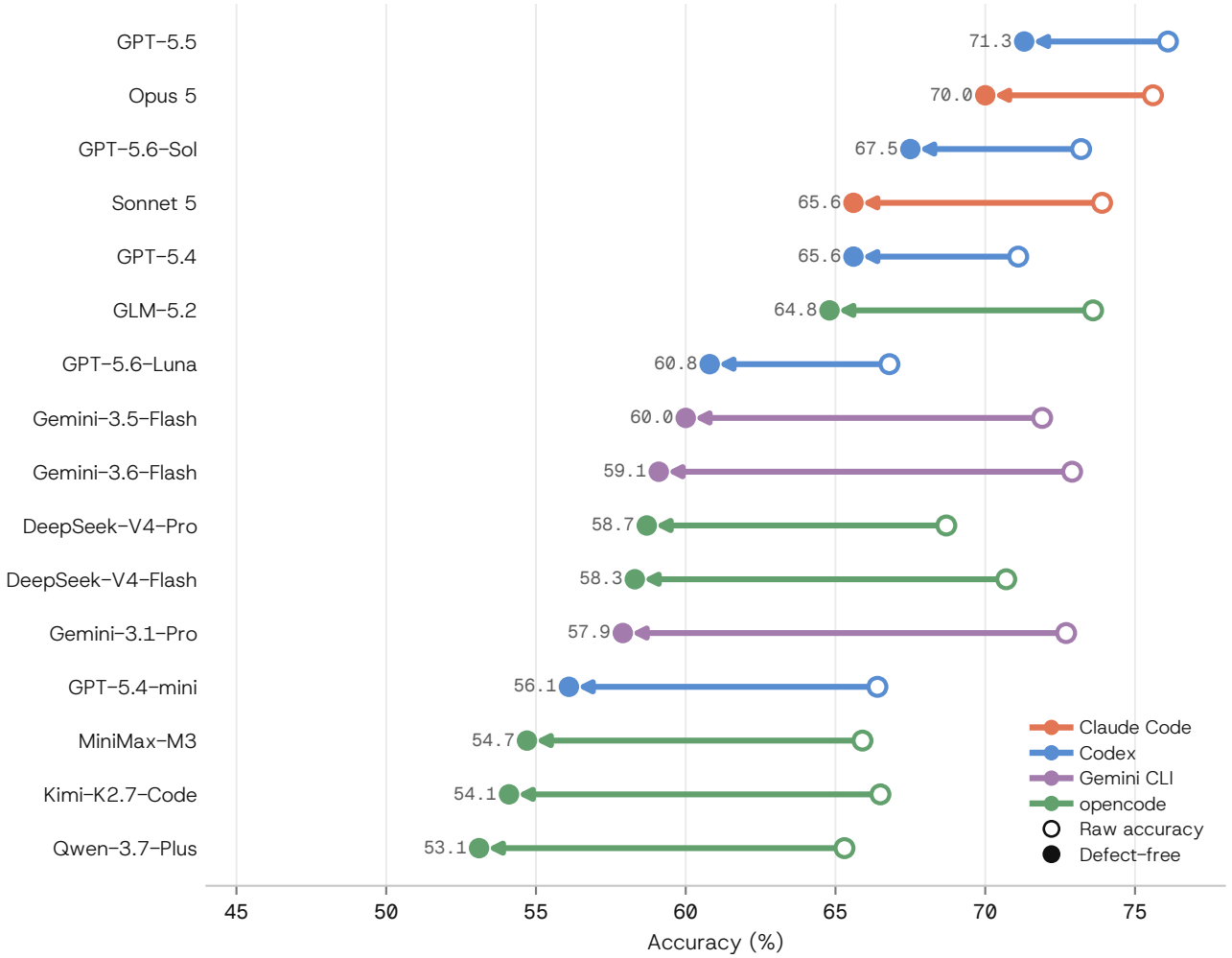

Figure 5. Defect-free accuracy reorders the leaderboard. Raw accuracy (open circles) and defect-free accuracy (filled markers) for all sixteen systems, ordered by defect-free accuracy; marker color denotes the harness family and connector length is the defect gap. Gemini-3.1-Pro falls from within noise of the leaders to below every Claude and Codex system bar GPT-5.4-mini once forbidden-route answers are removed, while GPT-5.5 loses least.

Defect-free accuracy reorders the field. Defect-free accuracy counts a report-present run as correct only when the verdict is right *and* no MUST-NOT criterion is violated. Compared to raw accuracy on report-present runs, every system loses between 4.8 and 14.8 points, corresponding to roughly one in sixteen to one in five correct verdicts failing the process audit.

The spread is the finding, not the magnitude (Figure 5). Were raw accuracy merely optimistic by a fixed margin, the gap would be roughly constant and the ranking would survive the correction. It does not. Gemini-3.1-Pro reaches 72.7% raw accuracy, within noise of the leading group, then falls to 57.9% defect-free, below every Claude system and every Codex system except GPT-5.4-mini, which it clears by 1.8 points. The reordering also promotes: GPT-5.6-Sol trails Sonnet 5 by 0.7 points on raw accuracy but leads it by 1.9 points once forbidden routes are removed. This operationalizes, on real longitudinal records, the outcome–process gap documented for medical LLMs [20, 21]: a correct label is not evidence of correct reasoning, and only a trace-level audit of the investigation distinguishes the two.

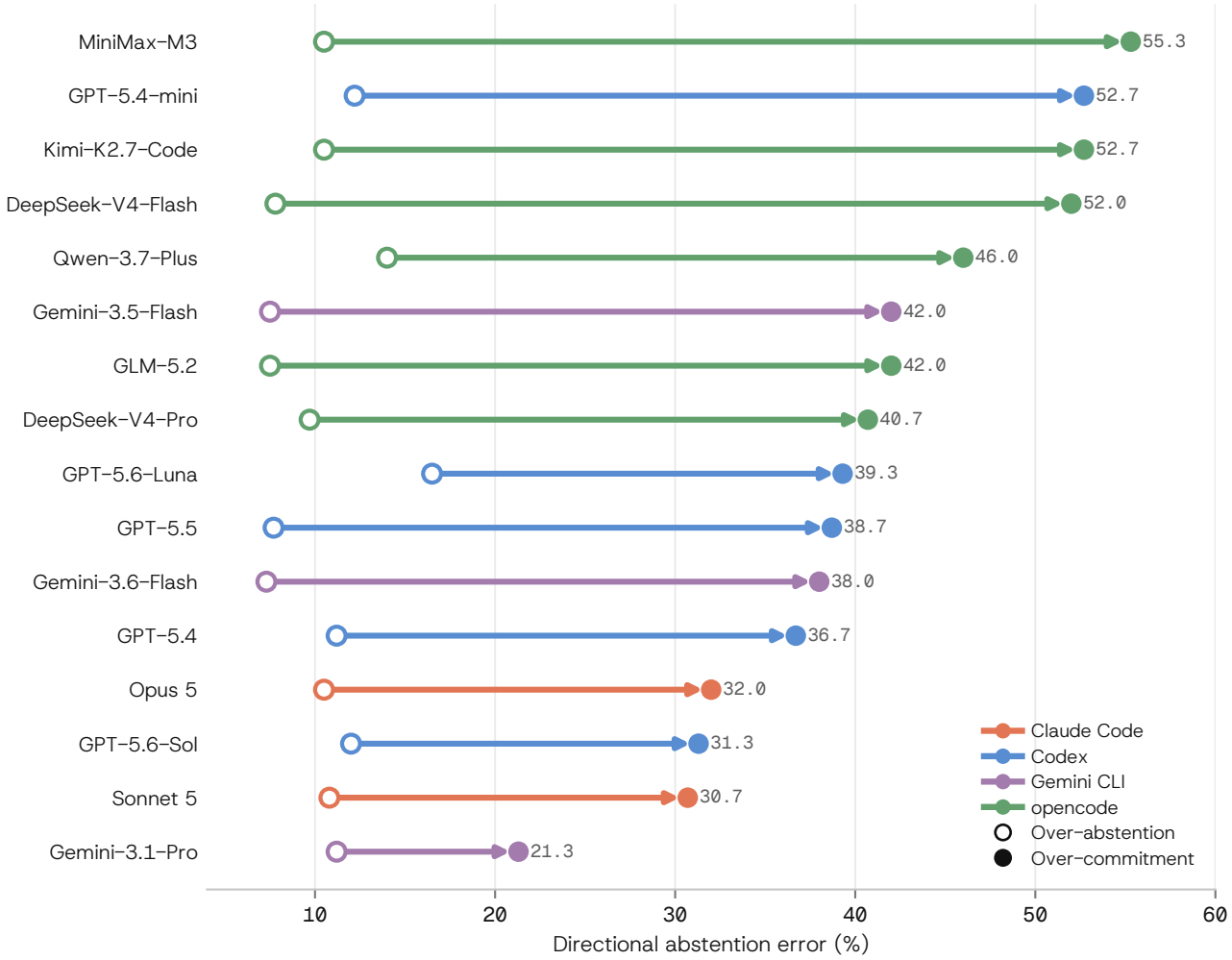

Figure 6. Every system under-abstains. Directional abstention errors for all sixteen systems: over-abstention (open circles), computed over reference-definitive runs, and over-commitment (filled markers), computed over reference-abstention runs; marker color denotes the harness family. Over-commitment exceeds over-abstention in every system. GPT-5.4-mini, DeepSeek-V4-Flash, MiniMax-M3 and Kimi-K2.7-Code over-commit on at least half of all deferral cases, whereas Gemini-3.1-Pro shows the smallest directional gap.

The process score measures something the verdict does not. Process adherence separates systems that accuracy places together. Opus 5 scores 82.0% on process against 75.6% accuracy, whereas Gemini-3.1-Pro reaches 72.7% accuracy on a process score of 56.4%, the same verdict quality reached through a materially less defensible investigation. This is the defect gap seen from the other side. For a clinical deliverable this matters more than the headline: a reader can audit a documented investigation but not a bare verdict.

Systems under-abstain: over-commitment exceeds over-abstention in every system. Measured directionally (Figure 6), over-commitment runs 21.3–55.3% against over-abstention's 7.3–16.5%, and *all sixteen* systems over-commit at the higher rate. That the asymmetry holds system by system rather than only in aggregate makes it a property of the systems rather than of the label distribution, and it reproduces on the development subset (Section 6.6). The magnitude, not the direction, is what varies: four systems over-commit on at least half of all deferral cases (MiniMax-M3 55.3%, DeepSeek-V4-Flash 52.0%,

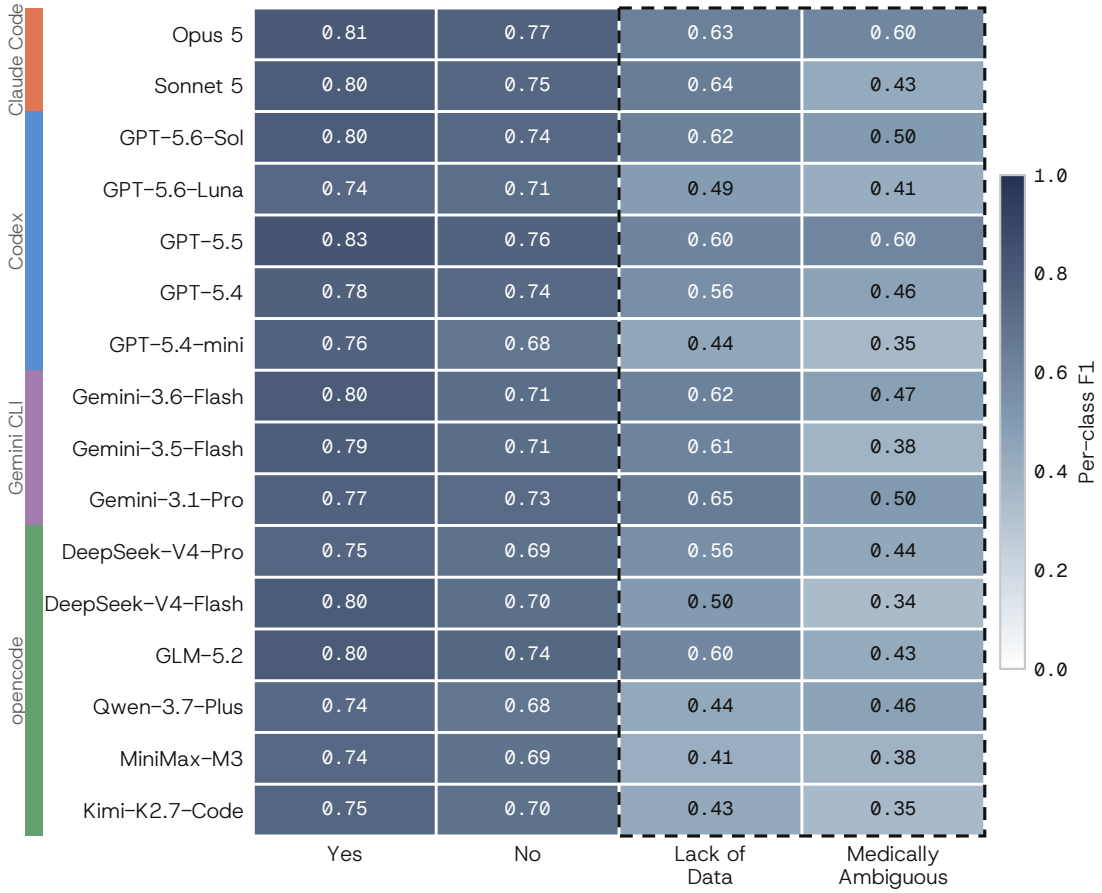

Figure 7. Performance collapses on the two safety-critical *Indeterminate* classes. Per-class F1 for all sixteen systems on the complete cohort (rows in leaderboard order, grouped by harness via the color bar at left). *Yes* and *No* stay strong while both deferral classes (dashed box) fall away on the weaker systems. Precision on the deferral classes is low throughout (*lack* 46–67%, *amb* 26–65%), as seen in Table 14.

GPT-5.4-mini and Kimi-K2.7-Code 52.7%) while Gemini-3.1-Pro does so on 21.3%. Over-abstention, by contrast, is tightly banded at 7.3–16.5%, and its highest value belongs to a vendor system (GPT-5.6-Luna), so the deferral side of the error budget varies far less across systems than the commitment side. Breaking accuracy down by class shows that even when models do abstain, they often do so on records that in fact settle the question (Figure 7): *Yes* and *No* F1 stay strong across the roster while both *Indeterminate* classes fall away, and precision on the deferral classes is low throughout (*lack* 46–67%, *amb* 26–65%), as seen in Table 14.

Calibration separates the systems more cleanly than accuracy does. Every parseable report includes a stated confidence in its class prediction, so we can ask whether that number carries information. Expected

calibration error is the bin-weighted deviation between stated confidence and realized accuracy,

$$\mathrm{ECE} \;=\; \sum_{b=1}^{B} \frac{|B_b|}{n} \,\big|\, \mathrm{conf}(B_b) - \mathrm{acc}(B_b) \,\big|, \tag{1}$$

over $B = 10$ equal-width bins of stated confidence, where n is the number of report-present runs with a valid confidence and $\mathrm{conf}(B_b)$ and $\mathrm{acc}(B_b)$ are the mean stated confidence and realized accuracy in bin b. ECE spans 0.046–0.262 across the sixteen systems, a nearly sixfold range on an axis accuracy cannot see, and it does *not* order them the way accuracy does: the three best-calibrated are Sonnet 5 (0.046), Opus 5 (0.064) and GPT-5.5 (0.090), while the three worst are all Gemini CLI systems (0.251–0.262) which sit mid-field on accuracy. In particular, the Gemini systems all emit only a few distinct confidence values, and their ECE reflects a lack of graded uncertainty rather than a poorly-shaped calibration curve. For a report addressed to a clinician who cannot cheaply re-derive the answer, a confidence that does not track correctness is a more consequential defect than a few points of accuracy.

Resource use and cost diverge from accuracy, and from each other. Across the vendor systems, a tenfold spread in per-run cost corresponds to only an approximately 10 point accuracy range (Figure 8). GPT-5.4-mini and GPT-5.6-Luna cost nearly the same ($0.20 and $0.21 per run) and differ by only 0.4 points in accuracy, yet GPT-5.6-Luna achieves 10.0 points higher process adherence and 4.7 points higher defect-free accuracy. Because the systems are similarly priced, these differences cannot be attributed to additional spend. Instead, they show that systems with comparable cost and verdict accuracy can differ substantially in the defensibility of their investigations. Their resource profiles also differ: GPT-5.4-mini issues a median of 42 typed tool calls compared with GPT-5.6-Luna's 18, roughly 2.3× as many, while consuming fewer work-tokens (84.4k versus 98.1k). The system making more typed tool calls nevertheless has the higher defect gap and lower process score, showing that greater retrieval activity alone does not ensure sound adjudication of the retrieved evidence.

6.3 Qualitative Analysis

We present an analysis of one case in detail to illustrate some of the failure signatures measured above. A cirrhotic patient is admitted with hepatic encephalopathy and moderate ascites, and no diagnostic paracentesis appears anywhere in the admission, so spontaneous bacterial peritonitis can be neither confirmed nor excluded and the reference verdict is *Indeterminate: Lack of Data* (`he-precipitant-278ac1`). Thirteen of the sixteen systems commit to a definitive label, twelve of them *Yes*, at stated confidences from 78% to 100%. Seven reach that label through the same prohibited shortcut, asserting peritonitis excluded with no ascitic-fluid result, the MUST-NOT criterion N4 (Table 4). The shortcut is not confined to one family: it is sprung by one Claude system, all three Gemini systems, and three of the six open-weight systems, while no Codex system encounters it. Not one of the sixteen earns D2, the criterion requiring confirmation that the major candidates, paracentesis especially, were actually evaluated.

Two systems make the separation of process from verdict concrete, in opposite directions. GPT-5.6-Sol is the only system in the roster to earn D5, which rewards escalating a key candidate that was never worked up despite indication; it springs no defect and carries the strongest MUST-DO profile of any committing system, and it still emits a definitive *No*. Recognizing an evidentiary gap is therefore not sufficient: the failure occurs at label assignment, after the reasoning that should have prevented it. Running the other way, the three systems that return the correct *Indeterminate* verdict — DeepSeek-V4-Pro, GLM-5.2 and Qwen-3.7-Plus — earn *none* of the nine process criteria between them, and one of the three states 95%

Figure 8. Accuracy rises only weakly with cost. Per-run cost (USD, log scale, harness-reported) against verdict accuracy for the ten vendor systems, colored by harness family, with ± 1 binomial standard error bars at $n = 750$. The six `opencode` systems are omitted because their per-run cost is a range rather than a point.

confidence in it. Their label is right and their investigation is undocumented. A single case thus exhibits both failure modes the two axes are designed to distinguish: sound reasoning that does not produce the correct label, and a correct label that reflects no defensible process. Scoring either axis alone would record one of these systems as exemplary and the other as broken, when the trace shows neither is.

6.4 Patient-Evidence Grounding

Grounding is where verdict accuracy and defensible reasoning most clearly diverge: a system can reach the correct verdict without citing the evidence that justifies it. Table 5 reports the four patient-evidence grounding metrics of Section 5.3 — P_{record} and P_{claim} for citation precision, $R_{\text{retrieval}}$ and R_{finding} for evidence coverage — for all sixteen model–harness combinations over the same 750 cases. All four are computed against the agent's own investigation rather than against the reference verdict, so they capture whether a determination is auditable, not whether it is correct. All four are stable under repetition: P_{record} and $R_{\text{retrieval}}$ are rule-based and return the same value on every computation, and both P_{claim} and R_{finding} reproduce reliably when the same reports are graded again (Section 6.7).

Record-identifier precision saturates and does not separate the systems. P_{record} spans 0.985–1.000 across the systems. No system fabricates record identifiers at a rate this check resolves, so the column establishes a lower bound on citation validity rather than discriminating between systems.

Table 4. Qualitative failure profile (`he-precipitant-278ac1`, all sixteen systems). Thirteen of sixteen commit to a definitive label; seven reach it through the same MUST-NOT criterion N4, "Asserted SBP excluded without an ascitic-fluid result". MUST-DO criteria are D1–D5 (D5: "Escalated when a key candidate was not worked up despite indication"). GPT-5.6-Sol is the only system to earn D5 and still commits; the three systems that return the correct verdict earn none of the nine criteria.

System	Verdict	Stated conf.	MUST-DO met	MUST-NOT sprung
Gold	*Ind.: Lack of Data*	–	–	–
Claude Opus 5 (Claude Code)	Yes	85%	D3,D4	–
Claude Sonnet 5 (Claude Code)	Yes	90%	D1,D3,D4	N4
GPT-5.6-Sol (Codex)	No	88%	D3,D4,**D5**	–
GPT-5.6-Luna (Codex)	Yes	88%	D3,D4	–
GPT-5.5 (Codex)	Yes	82%	D1,D3,D4	–
GPT-5.4 (Codex)	Yes	78%	D3,D4	–
GPT-5.4-mini (Codex)	Yes	92%	D3,D4	–
Gemini-3.6-Flash (Gemini CLI)	Yes	95%	D3,D4	N4
Gemini-3.5-Flash (Gemini CLI)	Yes	100%	D3,D4	N4
Gemini-3.1-Pro (Gemini CLI)	Yes	95%	D3	N4
DeepSeek-V4-Pro (opencode)	*Ind.: Lack of Data*	0%	–	–
DeepSeek-V4-Flash (opencode)	Yes	95%	D3	N4
GLM-5.2 (opencode)	*Ind.: Lack of Data*	95%	–	–
Kimi-K2.7-Code (opencode)	Yes	85%	D3	N4
MiniMax-M3 (opencode)	Yes	85%	D3,D4	N4
Qwen-3.7-Plus (opencode)	*Ind.: Lack of Data*	50%	–	–

Table 5. Patient-evidence grounding on all 750 cases. Citation precision and evidence coverage follow Section 5.3. Values are averaged over grounding-scoreable runs, defined as report-present runs in which the system retrieved at least one patient record.

	Citation precision		Evidence coverage	
System	P_{record}	P_{claim}	$R_{\text{retrieval}}$	R_{finding}
`Claude Code`				
Opus 5	0.985	0.884	0.852	0.843
Sonnet 5	0.988	0.869	0.818	0.744
`Codex`				
GPT-5.6-Sol	1.000	0.971	0.848	0.768
GPT-5.6-Luna	1.000	0.965	0.871	0.737
GPT-5.5	0.998	0.964	0.878	0.763
GPT-5.4	0.999	0.961	0.864	0.738
GPT-5.4-mini	0.997	0.964	0.813	0.673
`Gemini CLI`				
Gemini-3.6-Flash	0.995	0.899	0.859	0.789
Gemini-3.5-Flash	0.996	0.893	0.863	0.790
Gemini-3.1-Pro	0.999	0.903	0.689	0.604
`opencode` (open-weight models on the neutral harness)				
DeepSeek-V4-Pro	0.997	0.908	0.873	0.764
DeepSeek-V4-Flash	0.989	0.917	0.877	0.757
GLM-5.2	0.993	0.925	0.842	0.785
Qwen-3.7-Plus	0.994	0.930	0.821	0.732
MiniMax-M3	0.992	0.912	0.855	0.778
Kimi-K2.7-Code	0.997	0.940	0.900	0.767

Claim-level precision tracks the harness, not model capability. P_{claim} separates along harness families — Codex 0.965, `opencode` 0.922, Gemini CLI 0.898, Claude Code 0.877 — while within-family variation is much smaller. Each vendor harness holds its models tightly together despite spanning that vendor's range from small to frontier, so capability differences large enough to reorder verdict accuracy leave claim-level precision essentially unchanged. The six unrelated open-weight models on `opencode` are the one exception: their spread is 0.03 compared to within-family spreads of 0.01–0.015, but this remains a fraction of the gap separating the strongest and weakest family. Across the sixteen systems the P_{claim} ranking shows no positive association with the verdict-accuracy ranking of Table 3 (Spearman $\rho = -0.28$). This is the mirror image of the harness ablation (Table 10), where varying the harness at a fixed model left verdict accuracy inside noise: whether a determination is traceable to retrieved evidence is set largely by the scaffold that structures retrieval and report generation, whereas whether it is correct is set by the model.

Citation precision and evidence coverage are weakly anti-correlated. Across the 16 systems, P_{claim} correlates negatively with R_{finding} ($r = -0.22$) and weakly with $R_{\text{retrieval}}$ ($r = +0.23$), while the two coverage metrics correlate strongly ($r = +0.77$). This precision-coverage trade-off characterizes the field broadly but is most pronounced at the family level. For example, Codex and Claude Code occupy opposite extremes, where the most precise family is the least complete, and vice versa. However, this constraint does not dictate individual system performance; `opencode`, for instance, ranks second on both axes. Claim classifications reveal the mechanism behind this trade-off: across all families, ungrounded claims are predominantly *unverifiable* rather than *contradicted* (ranging from 3:1 to 5:1). Systems achieving higher coverage typically over-extrapolate beyond their retrieved evidence rather than misread it.

6.5 Policy Grounding

Policy grounding is scored separately from patient-evidence grounding (Section 5.4). We report its two axes: whether policy citations resolve and support the claims they carry, and whether the report identifies the governing documents, together with the effect of corpus access itself (Tables 6 and 7). Comparing the policy-on and policy-off conditions isolates whether corpus access changes system behavior.

Systems cite the governing policy correctly but incompletely. Policy grounding on the cohort (Table 6) separates two failure modes a single grounding score would merge. Take GPT-5.4-mini and GPT-5.6-Luna, two systems complete at 750 and within 0.4 accuracy points of each other. Citations almost always resolve (0.995 and 0.999), and precision of the cited document set against the clinical board's governing list is high (0.800 and 0.825). Recall of that set is roughly half (0.479 and 0.524): systems cite documents that genuinely govern, then omit about half of the others that also govern. Since a determination resting on part of the applicable standard can be right for the wrong reason, this is the policy-axis analogue of the defect gap, and a combined score would hide it because high precision offsets low recall. That failure shape is the one finding here that reproduces across all sixteen systems, because these are rates rather than counts and each is computed within a system: resolution runs 0.954–1.000 and document precision 0.764–0.899, while document recall never exceeds 0.703 and falls as low as 0.407. Every system in the roster predominantly cites real, applicable policy documents while omitting a substantial share of the other governing documents.

Unresolved policy citations dissociate sharply between systems of equal accuracy. The same two systems emit 20 and 3 unresolved policy citations, respectively—references that resolve to no corpus document—a nearly sevenfold difference between systems scoring within half a point of each other on accuracy. Citation support, the LLM-judged check that a cited passage entails the claim attached to it, moves the same way (0.629 against 0.814). The dissociation is far starker elsewhere in the roster: Gemini-3.5-Flash emits 303 unresolved citations and MiniMax-M3 159, against zero for GPT-5.5, GPT-5.4 and Gemini-3.1-Pro. With the defect gap and the confidence margins, this is the third axis on which these two systems are clearly distinguishable while their headline accuracy is not, and it is the one a clinician auditing the report would notice first.

Table 6. Policy grounding on all 750 cases. Resolution checks whether a citation names a corpus document and valid line span; Support is LLM-judged; Doc-P and Doc-R compare cited documents with the expert-specified governing set. Resolution and Support are averaged over runs with at least one policy citation, Doc-P over runs citing at least one corpus document, and Doc-R over all runs, with no-citation runs scoring zero. Unresolved citation counts are reported in the text.

System	Resolve	Support	Doc-P	Doc-R
`Claude Code`				
Opus 5	1.000	0.564	0.764	0.703
Sonnet 5	0.993	0.579	0.823	0.443
`Codex`				
GPT-5.6-Sol	1.000	0.735	0.814	0.581
GPT-5.6-Luna	0.999	0.814	0.825	0.524
GPT-5.5	1.000	0.747	0.786	0.586
GPT-5.4	1.000	0.624	0.782	0.549
GPT-5.4-mini	0.995	0.629	0.800	0.479
`Gemini CLI`				
Gemini-3.6-Flash	0.985	0.467	0.771	0.623
Gemini-3.5-Flash	0.984	0.467	0.769	0.614
Gemini-3.1-Pro	1.000	0.674	0.785	0.407
`opencode` (open-weight models on the neutral harness)				
DeepSeek-V4-Pro	0.997	0.621	0.870	0.525
DeepSeek-V4-Flash	0.954	0.566	0.849	0.452
GLM-5.2	0.995	0.584	0.820	0.576
Qwen-3.7-Plus	0.998	0.559	0.899	0.443
MiniMax-M3	0.976	0.471	0.817	0.530
Kimi-K2.7-Code	0.998	0.613	0.818	0.531

6.6 Ablations

The ablations use the frozen 148-case development subset described in Appendix E, which oversamples the two Indeterminate classes. We therefore use it only for paired contrasts and repeated-run analyses; its absolute scores are not comparable with Table 3. Paired ablations reuse Opus 4.8, whereas Appendix E.1 slices the full-cohort runs and therefore reports Opus 5 across per-pair tests and per-class breakdowns. All sixteen systems are complete at $n = 148$: trials lost to the execution environment were re-run rather than dropped, so every contrast below is paired over the same tasks. The subset is deliberately *not* used to separate systems from one another — of the $\binom{16}{2} = 120$ pairs, only 11 reach $p < 0.05$ under a two-sided exact McNemar test and none survives Holm correction. It does, however, confirm shared behavior: the

abstention asymmetry of Section 6.2 reproduces here on a deliberately different label distribution, with over-commitment exceeding over-abstention in *all sixteen* systems (26.4–60.4% against 5.3–16.8%), which is what makes it a property of the systems rather than of the cohort's label mix.

Withholding the policy corpus shows no reliable effect on either accuracy or ambiguity recognition. With the corpus withheld and nothing else changed, accuracy moves *inconsistently*: two of five systems fall, two rise, and one does not move, and a two-sided exact McNemar test on the paired cases separates *no* system's shift from zero ($p = 0.21$–1.00). Corpus access is therefore not worth a fixed number of points, and these deltas do not order the systems. Recall on *amb* is no more consistent: two systems fall, two rise and one is unchanged. At a support of 14 those movements are one to three cases apiece (± 7.1 points per case), so the column carries no direction either. We had expected withholding the corpus to cost *amb* recall specifically — the corpus is what tells an agent that a governing standard exists without settling the case — and it does not, which is a negative result rather than a measurement we can report as directional. The policy-enabled condition remains canonical. This is an ablation of the environment, not an alternative configuration (Table 7).

Table 7. Policy-corpus ablation. Conditions are paired by case with web search disabled. Δ is off minus on, so negative values indicate that withholding the corpus reduced performance. Exact two-sided McNemar tests find no accuracy effect ($p = 0.21$-1.00). *amb* recall has support 14, where one case changes recall by 7.1 points; its observed differences are also not significant ($p = 0.25$-1.00) and support neither a direction nor a ranking.

	Accuracy				Recall *amb* %		
Model (harness)	on	off	Δ	p	on	off	Δ
GPT-5.6-Sol (Codex)	70.3	70.9	+0.7	1.00	50.0	50.0	±0.0
GPT-5.5 (Codex)	73.0	70.9	−2.0	0.66	64.3	42.9	−21.4
Gemini-3.1-Pro (Gemini CLI)	77.0	72.3	−4.7	0.21	35.7	42.9	+7.1
GPT-5.6-Luna (Codex)	59.5	62.2	+2.7	0.56	71.4	50.0	−21.4
Opus 4.8 (Claude Code)	70.3	70.3	±0.0	1.00	50.0	57.1	+7.1

Single-attempt accuracy overstates consistent correctness, and the inconsistency is systematic. Across three separately executed attempts, pass^3 falls 8.3–12.6 points below avg@3, and the residual agreement is on the *same wrong* verdict, 13.5–20.9 points, and for GPT-5.6-Luna 30% of the cases on which it agreed with itself. Run-to-run variation is small next to that shortfall, with the standard deviation of accuracy across the three attempts at 0.7–4.7 points, so the gap is not attempt noise. High self-consistency is therefore not evidence of reliability: these systems are mostly *stably* wrong rather than randomly wrong, so repeating a query will not surface the error. Reliability is also worst exactly where the benchmark is hardest, as *every* system's lowest pass^3 falls on an *Indeterminate* class (Table 9).

At a fixed model, the harness is a cost choice, not a capability choice. Holding the model fixed and varying the harness across three model-agnostic harnesses moves accuracy little: the spread is 4.1 points, smaller than the across-model spread within a single harness, and no pairwise difference approaches significance on a paired test (Table 10). What separates the harnesses is cost, as median completion tokens differ 4.3-fold for accuracy differences inside noise, so for this model the harness is primarily a resource-use choice rather than a clearly separable capability difference.

Table 8. Repeated-run reliability, development subset. Each system has three independent attempts on all 148 cases. We report the reliability metrics defined in Section 5.6. *Spread* is pass@3 − pass^3, and *Overst.* is avg@3 − pass^3. *Agree-wr* is agreement on an incorrect verdict, so *Agree* equals pass^3 + Agree-wr up to rounding. Sampling error supports the aggregate pattern, not a ranking among systems.

System	n_{com}	avg@3	SD	pass^3	pass@3	Spread	Overst.	Agree	Agree-wr
Opus 4.8 (Claude Code)	148	**72.7**	3.7	**64.2**	**82.4**	18.2	8.6	**79.7**	15.5
GPT-5.5 (Codex)	148	71.8	1.4	63.5	80.4	**16.9**	**8.3**	77.0	13.5
GPT-5.6-Sol (Codex)	148	70.3	0.7	59.5	79.7	20.3	10.8	78.4	18.9
GPT-5.6-Luna (Codex)	148	60.6	1.0	48.0	72.3	24.3	12.6	68.9	20.9
Gemini-3.1-Pro (Gemini CLI)	148	72.3	4.7	60.8	**82.4**	21.6	11.5	76.4	15.5

Table 9. Per-class repeated-run reliability, development subset. Values use three attempts; class supports are YES 50, NO 45, LACK 39, AMB 14. pass^3 and pass@3 follow Section 5.6. At an AMB support of 14, one case changes a cell by 7.1 points, so the amb values indicate direction only and should not be used to rank systems.

	YES				NO				LACK				AMB			
Model (harness)	n	^3	@3	SD	n	^3	@3	SD	n	^3	@3	SD	n	^3	@3	SD
Opus 4.8 (Claude Code)	50	**78.0**	**96.0**	7.2	45	60.0	**86.7**	3.4	39	56.4	64.1	1.5	14	**50.0**	71.4	7.1
GPT-5.5 (Codex)	50	70.0	88.0	2.0	45	**73.3**	84.4	1.3	39	51.3	71.8	3.0	14	42.9	64.3	10.9
GPT-5.6-Sol (Codex)	50	68.0	88.0	1.2	45	64.4	82.2	4.4	39	48.7	71.8	2.6	14	42.9	64.3	4.1
GPT-5.6-Luna (Codex)	50	52.0	72.0	5.3	45	48.9	73.3	1.3	39	43.6	69.2	2.6	14	42.9	**78.6**	12.4
Gemini-3.1-Pro (Gemini CLI)	50	64.0	86.0	5.3	45	64.4	**86.7**	8.0	39	**69.2**	**84.6**	3.0	14	14.3	50.0	4.1

Test-time compute buys neither better verdicts nor better-documented ones. We compare three explicit levels on one model over the same 148 cases (Table 11). Compute rises substantially, 38% from `low` to `medium` and 72% end to end, as median within-case ratios rather than ratios of medians. Quality does not follow. Accuracy spans 0.7 points with every pairwise difference inside 0.1 standard errors and both exact McNemar tests at $p = 1.00$. Process adherence does rise monotonically across the three levels, by 4.7 points end to end, but at $1.5\,SE_d$ that too is inside noise — the ordering is suggestive and the separation is not established. Additional test-time compute buys neither better verdicts nor a better-documented method, which is what one expects if the binding constraint is retrieval elicitation and label discrimination rather than reasoning depth. One limit is that the trajectories expose no reasoning-token count, so completion tokens are the only available compute proxy and we cannot say *where* the extra compute was spent.

Table 10. Harness comparison at a fixed model. DeepSeek-V4-Pro is evaluated with three model-agnostic harnesses on the same 148 tasks. Accuracy differs by at most 4.1 points (exact McNemar $p \geq 0.46$), while median completion tokens differ 4.3-fold. The three builds carry identical task-content digests.

Harness	Acc %	Δ	p	Proc%	Turns	Comp. tok
`opencode`	66.2 ± 3.9	—	—	66.6 ± 2.2	14	5,051
`qwen-coder`	62.8 ± 4.0	−3.4	0.46	66.9 ± 2.2	14	9,524
`mini-swe-agent`	66.9 ± 3.9	+0.7	1.00	66.8 ± 2.2	52	21,876

Table 11. Test-time compute buys compute, not accuracy. Opus 4.8 on Claude Code, three explicit reasoning-effort levels over the same 148 development tasks. Completion tokens rise 38% to `medium` and 72% end to end, but no pairwise accuracy difference exceeds 0.1 SE_d (exact McNemar $p \geq 1.00$) and no process difference exceeds 1.5 SE_d. Reasoning depth is not the binding constraint on this benchmark.

Effort	Acc %	Δ	p	Proc%	Comp. tok	Δ tok
`low`	71.6 ± 3.7	—	—	65.0 ± 2.2	7,332	—
`medium`	72.3 ± 3.7	+0.7	1.00	68.3 ± 2.1	10,268	+38%
`high`	72.3 ± 3.7	+0.7	1.00	69.7 ± 2.2	12,420	+72%

6.7 Benchmark Validation: Judge Reliability

The process score and the reference labels both feed the rankings, so we validate each. The reference labels are validated against independent clinician judgment in Section 4.5 (91% agreement on the calibration sample, $\kappa = 0.87$); here we validate the deterministic and LLM-judged components of patient-evidence grounding, process adherence, and policy grounding.

Judge reliability and self-consistency. The process rubric is LLM-adjudicated, so we measure whether the grade is stable across judges rather than assuming it. A five-judge panel grades every trajectory over the full rubric with three replicates each. We report inter-judge agreement primarily as the mean pairwise disagreement. It flags 16% of criteria as contested, with the judges self-consistent across replicates and statistically interchangeable on the aggregate score (companion Fleiss' $\kappa \approx 0.65$). The most-contested criteria (for example, a hidden antibiotic de-escalation window and an era-dependent heart-failure therapy) are exactly the cases an independent expert audit flagged as under-specified, so judge disagreement acts as an automatic detector of ambiguous criteria, which are then routed back for expert revision.

Grounding-metric determinism and judge reliability. The four patient-evidence grounding metrics divide by construction, and each is validated according to what it produces. P_{record} and $R_{\text{retrieval}}$ are computed by rule from the trace and involve no judge at all: recomputing them over a complete 750-task run returns every per-trial value unchanged. The other two are LLM-judged, so we grade the same reports twice and compare the two passes. R_{finding} asks the judge one covered/not-covered question per rubric category. Because those categories are fixed in advance, the two passes can be compared question by question: over 200 reports, 50 per harness family, they agree on 97.9% of 1,331 category decisions (Cohen's $\kappa = 0.95$), 175 of the 200 reports receive an identical score, and no family average moves by more than 0.015. P_{claim} gives no such fixed question list, because the judge must first break the report into atomic claims and it does not divide it the same way twice; there is nothing to align, so we report how far the score itself moves, which is at most 0.02 and in no consistent direction. We therefore treat the two rule-based metrics as exact, R_{finding} as reproducible, and differences of 0.02 or less on P_{claim} as noise.

Policy-judge reliability and self-consistency. Policy grounding is mostly deterministic: citation resolution and governing-document matching are computed by rule, while only citation support requires LLM judgment. Accordingly, the two rule-based components reproduce exactly, and the LLM-judged support component is stable across judges (companion Fleiss' $\kappa \approx 0.86$, within-judge noise 0.023–0.025 on a $0, 1$ scale). The production judge is therefore representative rather than unusually strict or lenient. The residual variation is systematic: judges sometimes under-credit near-verbatim citations, making

citation-support estimates conservative lower bounds, and we find no evidence of own-family favoritism. Defining governing-document identification through expert-specified document sets rather than LLM judgment removes a major source of ambiguity at the rubric-design stage. This supports treating the policy-grounding metrics as meaningful measurements rather than judge noise.

Judge and label independence. GPT-5.5 serves three roles: one of three systems in the reference-verdict ensemble, an evaluated system through Codex, and the production judge for the semantically adjudicated metrics. Verdict accuracy involves no judge, being a deterministic four-class match against the reference labels. The labels are bounded by the ensemble and by clinician review: GPT-5.5 casts one of three votes, redundant on the 556 of 750 cases the ensemble decided unanimously, and the reference agrees with each clinician more closely than the two clinicians agree with each other (Section 4.5). Process adherence and policy support are graded by multi-judge panels (Section 6.7). The remaining exposure is P_{claim} and R_{finding}, which the production judge grades alone and which regrading validates for stability rather than bias. Codex leads P_{claim}, but GPT-5.5 ranks fourth of its five systems, so the gap does not track the judge's own model. Replication with an independent judge remains the direct test, left for future work.

7 Conclusion

CLINICARE-BENCH reframes clinical-agent evaluation around the capabilities required for defensible retrospective clinical audit: planning evidence retrieval across a longitudinal EHR, grounding clinical claims in traceable patient evidence, applying governing standards, following an observable and clinically defensible investigation process, and deferring when the available record cannot support a unique conclusion. Building on the trace-level clinical auditing introduced by Hager et al. [19], CLINICARE-BENCH brings these dimensions together within a common patient-level adjudication framework.

The benchmark comprises 25 clinician-authored and validated scenarios instantiated as 750 patient cases over MIMIC-IV. Each run is evaluated against a case-level reference verdict produced through independent multi-model adjudication and calibrated against blinded Clinical Board review on a stratified subset. Beyond final-verdict correctness, the benchmark evaluates calibrated abstention, patient-evidence grounding, process adherence, policy grounding, repeated-run reliability, and resources.

The evaluation exposes two deployment-relevant failure modes that raw accuracy obscures. First, every evaluated system under-abstains, returning a definitive verdict on cases for which the reference standard requires deferral because evidence is missing or the record remains medically ambiguous. Second, defect-free accuracy falls by 4.8–14.8 percentage points relative to report-present accuracy and changes the system ranking. For the most affected system, roughly one in five report-present correct verdicts is reached through a prohibited shortcut. These results show that a correct final answer does not by itself establish that an agent performed a defensible clinical investigation.

Scope and limitations. CLINICARE-BENCH trades per-scenario case volume for breadth of workflow coverage and depth of annotation. It spans 25 scenarios across 14 medical specialties and ten reasoning capabilities, but contains only 30 cases per scenario and 750 cases in total, compared with the 2,400 cases concentrated on four abdominal pathologies in MIMIC-CDM [19]. Its contribution is therefore not greater sample depth within a single diagnostic family, but the density of its scenario and case specifications: explicit four-way adjudication criteria, required-evidence and grounding requirements,

weighted MUST-DO/MUST-NOT process rubrics, and clinician-calibrated reference verdicts. Expanding the number of cases per scenario while preserving this annotation depth is an important direction for future releases.

The current study also has several important limitations. The benchmark is derived from a single de-identified academic medical-center dataset, and its deliberately stratified case distribution should not be interpreted as clinical prevalence. Reference verdicts are produced by a model-assisted adjudication procedure and calibrated against clinician review on a subset, rather than established through exhaustive independent clinician review of all 750 cases. Clinical Board review in this study calibrates the scenario specifications and reference verdicts and should not be interpreted as a direct human-versus-agent performance comparison.

Although developed for clinical audit, the benchmark's central design principles, including governed and logged tool access, explicit evidentiary requirements, calibrated abstention, and separate scoring of outcome and observable process, apply more broadly to other high-stakes domains in which decisions must be transparent, reproducible, and auditable, such as finance, cybersecurity, and law.

Open-source release and MedHELM integration. We will release CliniCARE-Bench in two tiers. The non-patient-specific artifacts (scenario specifications, evaluation framework, harness and metric code, and judge prompts) will be released openly under a permissive license. The patient-linked artifacts, which are derived from MIMIC-IV, will be made available on PhysioNet under its credentialed access policy and data use agreement. We are integrating the benchmark into the MedHELM evaluation harness [6] and, with Pacific AI, upstreaming it alongside related efforts such as PhysicianBench [31] and HealthAdminBench [39], so that its process-aware, grounding- and abstention-scored cases are usable across a broader open medical-evaluation ecosystem.

Author Contributions

Y. Xue conceived the project, defined the vision and benchmark direction, led the manuscript effort, and supervised the work. V. Chatrath led benchmark execution, including the Clinical Board network and experiments, and served as primary manuscript lead. B. Zhu served as healthcare lead and technical lead and was a primary developer. G. Pu served as technical lead and was a primary developer. J. Fan led evidence-grounding and metrics ideation and implementation. A. Shanker led policy-grounding ideation, and V. Ursekar led process rubric implementation and judge calibration. A. Sharma contributed to evidence grounding and the Clinical Board network and created the MIMIC environment. J. Qin provided domain expertise and experimental ideation. K. Han contributed medical domain expertise, and related work. S. D. Tiwari and S. Dan contributed to medical case release and environment setup. Y. Li and V. Kalmath contributed to manuscript iteration. D. Y. Zhang oversaw the environment release. Z. Doctor contributed medical domain expertise. Z. Yin served as a research advisor in machine learning and AI for healthcare, guiding clinical agent evaluation design and contributing to the drafting and review of the related-work section. C. Wang served as research advisor for benchmark design, and N. Shah provided senior clinical and research guidance, including benchmark positioning and connections to the broader medical-AI evaluation community. All authors contributed to the writing, reviewed and approved the final manuscript.

References

[1] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

[2] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375, 2023. URL https://api.semanticscholar.org/CorpusID:257687695.

[3] Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K. Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chio-Kin Tong, Kavin Karthik, Arnav Dugar, Akshay Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. Healthbench professional: Evaluating large language models on real clinician chats. *ArXiv*, abs/2604.27470, 2026. URL https://api.semanticscholar.org/CorpusID:287767203.

[4] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *ArXiv*, abs/2405.07960, 2024. URL https://api.semanticscholar.org/CorpusID:269757778.

[5] Luyang Luo, Sung Eun Kim, Xiaoman Zhang, Julius Kernbach, Roshan Kenia, Julián Nicolás Acosta, Larry A Nathanson, Adrian D Haimovich, Adam Rodman, Ethan Goh, Jonathan H. Chen, Nigam H. Shah, David A. Kim, James Zou, Faisal Mahmood, Jakob Nikolas Kather, Matthew P. Lungren, Vivek Natarajan, Eric J. Topol, and Pranav Rajpurkar. A clinical environment simulator for dynamic ai evaluation. *Nature Medicine*, 2026. doi: 10.1038/s41591-026-04252-6. URL https://doi.org/10.1038/s41591-026-04252-6.

[6] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya J. Shah, Ethan Goh, Dong han Yao, Brian T. Soetikno, Eduardo Pontes Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachnanjali L Saralkar, Chia-Chun Chiang, Jenelle A. Jindal, Tho D. Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Colin Hong, Mohana Roy, Michael Francis Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, François Grolleau, Kameron Collin Black, Bethel R Mieso, Aydin Zahedivash, Wen wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Y Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, A. Ju. Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan H. Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay S. Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. Medhelm: Holistic evaluation of large language models for medical tasks. *ArXiv*, abs/2505.23802, 2025. URL https://api.semanticscholar.org/CorpusID:279070977.

[7] Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher D. Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. Medalign: A clinician-generated dataset for instruction

following with electronic medical records. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/AAAI.V38I20.30205. URL https://doi.org/10.1609/aaai.v38i20.30205.

[8] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Alan Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=CsXC6IcdwI.

[9] Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, Mark Alexander Burbridge, James C. McAvoy, Roya Saffary, Stephen P. Ma, Dev Dash, James Xie, Ellen Y. Wang, Clifford A. Schmiesing, Nigam H. Shah, and Nima Aghaeepour. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1), 2026. doi: 10.1056/AIdbp2500418. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500418.

[10] Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle A Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi Omiye, Mehr Kashyap, and Nigam Shah. FactEHR: A dataset for evaluating factuality in clinical notes using LLMs. In Monica Agrawal, Kaivalya Deshpande, Matthew Engelhard, Shalmali Joshi, Shengpu Tang, and Iñigo Urteaga, editors, *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR, 15–16 Aug 2025. URL https://proceedings.mlr.press/v298/munnangi25a.html.

[11] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Oluwasanmi Koyejo, and Nigam H. Shah. Timer: temporal instruction modeling and evaluation for longitudinal clinical records. *NPJ Digital Medicine*, 8, 2025. doi: 10.1038/s41746-025-01965-9. URL https://api.semanticscholar.org/CorpusID:276813703.

[12] Yusuke Watanabe, Yohei Kobashi, Takeshi Kojima, Yusuke Iwasawa, Yasushi Okuno, and Yutaka Matsuo. Clindet-bench: Beyond abstention, evaluating judgment determinability of llms in clinical decision-making, 2026. URL https://arxiv.org/abs/2602.22771.

[13] Sravanthi Machcha, Sushrita Yerra, Sahil Gupta, Aishwarya Sahoo, Sharmin Sultana, Hong Yu, and Zonghai Yao. Knowing when to abstain: Medical llms under clinical uncertainty, 2026. URL https://arxiv.org/abs/2601.12471.

[14] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

[15] Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2012.

[16] Center for Medicare & Medicaid Services. Severe Sepsis and Septic Shock: Management Bundle Measure, 2020. URL https://www.cms.gov/priorities/innovation/media/document/bpci-advanced-alt-fs-my4-sepsis.

[17] Jeffrey L. Carson, Simon J. Stanworth, Gordon Guyatt, Stacey Valentine, Jane Dennis, Sara Bakhtary, Claudia S. Cohn, Allan Dubon, Brenda J. Grossman, Gaurav K. Gupta, Aaron S. Hess, Jessica L.

Jacobson, Lewis J. Kaplan, Yulia Lin, Ryan A. Metcalf, Colin H. Murphy, Katerina Pavenski, Micah T. Prochaska, Jay S. Raval, Eric Salazar, Nabiha H. Saifee, Aaron A. R. Tobian, Cynthia So-Osman, Jonathan Waters, Erica M. Wood, Nicole D. Zantek, and Monica B. Pagano. Red blood cell transfusion: 2023 AABB international guidelines. *JAMA*, 330(19), 2023. doi: 10.1001/jama.2023.12914.

[18] Paul A. Heidenreich, Biykem Bozkurt, David Aguilar, Larry A. Allen, Joni J. Byun, Monica M. Colvin, Anita Deswal, Mark H. Drazner, Shannon M. Dunlay, Linda R. Evers, James C. Fang, Savitri E. Fedson, Gregg C. Fonarow, Salim S. Hayek, Adrian F. Hernandez, Prateeti Khazanie, Michelle M. Kittleson, Christopher S. Lee, Mark S. Link, Carmelo A. Milano, Lorraine C. Nnacheta, Alexander T. Sandhu, Lynne Warner Stevenson, Orly Vardeny, Amanda R. Vest, and Clyde W. Yancy. 2022 aha/acc/hfsa guideline for the management of heart failure: A report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145(18):e895–e1032, 2022. doi: 10.1161/CIR.0000000000001063.

[19] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[20] Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021, 2025.

[21] Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel C. F. Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, HoHin Lee, Praneeth Sanapathi, Sarah Hilado, Tristan Naumann, Javier Alvarez-Valle, Jiang Bian, Mu Wei, Khalil Malik, Lidong Zhou, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Doug Burger, Eric Topol, Hoifung Poon, and Paul Vozila. Evaluating the robustness and readiness of large frontier models in health ai applications. *Nature Medicine*, 2026. doi: 10.1038/s41591-026-04501-8.

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

[23] Jacob T Rosenthal, Ashley Beecy, and Mert R Sabuncu. Rethinking clinical trials for medical ai with dynamic deployments of adaptive systems. *npj Digital Medicine*, 8(1):252, 2025.

[24] Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, et al. Towards autonomous medical artificial intelligence agents. *Nature*, 2026. doi: 10.1038/s41586-026-10675-5. URL https://www.nature.com/articles/s41586-026-10675-5. Advance online publication.

[25] Yuxing Lu, Yushuhong Lin, Wenqi Shi, J. Ben Tamo, Xukai Zhao, Jinzhuo Wang, and May Dongmei Wang. Clinenv: An interactive multi-stage long horizon ehr environment for agents, 2026. URL https://arxiv.org/abs/2606.02568.

[26] Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. Ehrsql: A practical text-to-sql benchmark for electronic health records. *Advances in Neural Information Processing Systems*, 35:15589–15601, 2022.

[27] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1245. URL https://aclanthology.org/2024.emnlp-main.1245/.

[28] Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, page AIdbp2500144, 2025.

[29] Gyubok Lee, Elea Bach, Eric Yang, Tom J. Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *ArXiv*, abs/2509.19319, 2025. URL https://api.semanticscholar.org/CorpusID:281505607.

[30] Yitong Qiao, Lei Liu, Yue Shen, Jian Wang, Jinjie Gu, Zhixuan Chu, and Kui Ren. Ehr-complex: Benchmarking medical agents for complex clinical reasoning, 2026. URL https://arxiv.org/abs/2606.23301.

[31] Ruoqi Liu, Imran Q. Mohiuddin, Austin J. Schoeffler, Kavita Renduchintala, Ashwin Nayak, Prasantha L. Vemu, Shivam C. Vedak, Kameron C. Black, John L. Havlik, Isaac Ogunmola, Stephen P. Ma, Roopa Dhatt, and Jonathan H. Chen. Physicianbench: Evaluating llm agents in real-world ehr environments, 2026. URL https://arxiv.org/abs/2605.02240.

[32] Yanzhen Chen, Zihan Xu, Xiaocheng Zhang, Zhiting Fan, Weiqi Zhai, Hongxia Xu, and Zuozhu Liu. Longmedbench: Benchmarking medical agents for long-horizon clinical decision-making, 2026. URL https://arxiv.org/abs/2607.09322.

[33] Sarvesh Soni and Dina Demner-Fushman. A dataset for addressing patient's information needs related to clinical course of hospitalization. *Scientific Data*, 13:523, 2026. doi: 10.1038/s41597-026-06639-z. URL https://doi.org/10.1038/s41597-026-06639-z.

[34] Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. Clicare: Grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records, 2026. URL https://arxiv.org/abs/2507.22533.

[35] Gyubok Lee, Sunjun Kweon, Seongsu Bae, and Edward Choi. Overview of the EHRSQL 2024 shared task on reliable text-to-SQL modeling on electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 644–654. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.clinicalnlp-1.62. URL https://aclanthology.org/2024.clinicalnlp-1.62/.

[36] Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026.

[37] Suhana Bedi, Iddah Mlauzi, Daniel Shin, Sanmi Koyejo, and Nigam H. Shah. The optimization paradox in clinical AI multi-agent systems. In *Agentic & GenAI Evaluation Workshop, KDD 2025 (Evaluation and Trustworthiness of Agentic and Generative AI Models)*, 2025. URL https://openreview.net/forum?id=EJDg2nLVHd.

[38] Vincent Siu, Jingxuan He, Kyle Montgomery, Zhun Wang, Neil Gong, Chenguang Wang, and Dawn Song. A framework for formalizing LLM agent security, 2026. URL https://arxiv.org/abs/2603.19469.

[39] Suhana Bedi, Ryan Welch, Ethan H. Steinberg, Michael Wornow, Taeil Kim, Haroun Zakaria Ahmed, Peter V. Sterling, Bravim K. Purohit, Qurat-Ul-Ain Akram, Angelic Acosta, Esther Nubla, Priti Sharma, Mike Pfeffer, Oluwasanmi Koyejo, and Nigam H. Shah. Healthadminbench: Evaluating computer-use agents on healthcare administration tasks. *ArXiv*, abs/2604.09937, 2026. URL https://api.semanticscholar.org/CorpusID:287425604.

[40] Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, January 2023. doi: 10.13026/1n74-ne17. URL https://doi.org/10.13026/1n74-ne17. Version 2.2.

[41] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, January 2023. doi: 10.13026/5ntk-km72. URL https://doi.org/10.13026/5ntk-km72. Version 2.2.

[42] Gordon F. Tomaselli, Kenneth W. Mahaffey, Adam Cuker, Paul P. Dobesh, John U. Doherty, John W. Eikelboom, Roberta Florido, Ty J. Gluckman, William J. Hucker, Roxana Mehran, Steven R. Messé, Alexander C. Perino, Fatima Rodriguez, Ravindra Sarode, Deborah M. Siegal, and Barbara S. Wiggins. 2020 ACC Expert Consensus Decision Pathway on Management of Bleeding in Patients on Oral Anticoagulants. *JACC*, 76(5):594–622, August 2020. doi: 10.1016/j.jacc.2020.04.053.

A Terminology and Units of Evaluation

To distinguish the benchmark's clinical specifications, patient-level instances, and experimental executions, we use the following terminology throughout the paper.

Term	Definition
Scenario	A reusable, clinician-authored evaluation specification comprising a clinical-query template, operational adjudication criteria, evidence and policy-grounding requirements, and a process rubric. The benchmark contains 25 scenarios.
Case	A patient-specific instantiation of a scenario over a particular MIMIC-IV record and, where applicable, a specific encounter or clinical event. Each scenario is instantiated on 30 cases, yielding 750 cases in total.
Run	One execution of an evaluated system on one case. A run produces a tool-call trajectory, retrieved evidence, computations or artifacts, and a final report containing the system's verdict and supporting citations.
Attempt	The repetition index when the same system is independently run more than once on the same case. Repeated attempts are used to measure consistency, including pass^k and avg@k.
Cohort	A defined collection of cases used for analysis, such as the 30 cases associated with one scenario, the 148-case development subset, or the full 750-case benchmark.
System	The complete evaluated configuration, including the model, agent harness, prompting or scaffolding, and available tools.

Table 12. Terminology and units of evaluation used in CLINICARE-BENCH.

B Additional CliniCARE-Bench scenarios

We detail four representative clinical scenarios below. Each states its clinical query, the four-way adjudication criteria (*Yes / No / Indeterminate: Lack of Data / Indeterminate: Medically Ambiguous*), the evidence a grounded answer must surface, and the grounding requirement.

Post-CT Acute Kidney Injury. This scenario asks whether an ICU patient developed acute kidney injury within 48 hours after a CT scan, applying the published KDIGO criteria while explicitly avoiding any inference that the CT or contrast caused the injury.

- **Clinical Query:** "For ICU patient `<subject_id, hadm_id>` who underwent a CT this admission, did the patient develop AKI within 48 h after the CT? Set the baseline creatinine hierarchically—the lowest value in the 3–12 months before admission; else the admission creatinine (the patient may already present in AKI); else, if CKD is documented, an MDRD estimate assuming eGFR 75 mL/min/1.73 m^2. Apply the KDIGO serum-creatinine criteria and frame the result as AKI following CT, not as contrast-caused injury; where serum creatinine does not reflect the patient's own kidney function, say the question does not resolve rather than applying the thresholds."
- **Adjudication Criteria:**
  - *Yes:* A baseline is set by the hierarchy (the lowest creatinine in the 3–12 months pre-admission; else the admission creatinine; else an MDRD estimate where CKD is documented); at least one creatinine value falls $\leq$ 48 h post-CT, with all post-CT values searched for the peak (not only the first); a post-CT value meets a KDIGO threshold (absolute rise $\geq$ 0.3 mg/dL above baseline, or $\geq 1.5\times$ baseline); and the patient is not on chronic dialysis or ESRD.
  - *No:* A baseline can be established by the hierarchy and at least two creatinine values fall within 48 h post-CT (adequate sampling); no post-CT value meets either KDIGO threshold (absolute rise $\geq$ 0.3 mg/dL or $\geq 1.5\times$ baseline); and the patient is not on chronic renal replacement therapy and has no ESRD or CKD stage ≥ 4.
  - *Indeterminate: Lack of Data:* No creatinine is available to set any tier of the baseline hierarchy (no 3–12 month prior value, no admission creatinine, and no CKD documentation supporting an MDRD estimate), or none falls within 48 h post-CT; CT timing cannot be resolved (order time versus actual scan time, or multiple CTs prevent a unique "first qualifying" CT); no CT can be identified during the admission, so the premise is unmet and the case is out of scope; or post-CT sampling is insufficient ($<$ 2 creatinine values) to confirm the peak.
  - *Indeterminate: Medically Ambiguous:* Chronic dialysis or ESRD (ICD N18.6 / Z99.2 / 585.6 / V45.11, or any hemodialysis order), so the KDIGO acute thresholds do not apply; or a labile baseline—the candidate prior creatinines (3–12 months before admission) fluctuate materially, so a clean reference cannot be set and a $+0.3$ mg/dL rise cannot be confidently attributed.
- **Required Evidence:** Creatinine timeline from the baseline source (the lowest value in the 3–12 months pre-admission; else the admission value; else the CKD/MDRD basis) through 48 h post-CT, showing every value; ESRD/dialysis status; and the CT *scan* time distinguished from the order time (delays), confirmed via the radiology report where possible.
- **Required Grounding:** The specific MIMIC-IV record or record span from which the decisive finding was derived.

Advanced Cross-Sectional Imaging Receipt. This scenario asks whether an ICU patient received advanced cross-sectional imaging—a CT or MRI—at any point during a specific ICU stay, resolving the question from the order stream confirmed against the radiology report rather than from an order alone.

- **Clinical Query:** "For ICU stay <subject_id, hadm_id, stay_id>, did the patient receive advanced cross-sectional imaging (a CT or MRI) at any point during this ICU stay window [intime, outtime]? Decide how to identify imaging events and cite the evidence; an order alone does not settle the question where a confirming radiology report is available, and for a multi-stay admission attribute each order to a single stay before answering."
- **Adjudication Criteria:**
  - *Yes:* At least one POE radiology order (order_type="Radiology") of a CT/MRI-class subtype ("CT Scan", "MRI", or "Cross-Sectional Interventional Radiology"), with ordertime inside the ICU stay window and a non-discontinued status, together with a radiology report confirming the study was actually performed. An order with no confirming report is not sufficient for *Yes*.
  - *No:* Within the ICU stay window the patient received only bedside or non-cross-sectional imaging ("General Xray", "Ultrasound", "Noninvasive Vascular", or a TEE echo order rather than a radiology CT/MRI), with no CT or MRI; or no radiology order of any kind falls within the window while the order stream for the stay is present and complete, so the absence is a true negative rather than missing data.
  - *Indeterminate: Lack of Data:* A CT/MRI order was cancelled or discontinued with no confirming report; a multi-ICU-stay admission where the order cannot be attributed to this specific stay; or no ICU stay window is recorded for the patient, so the premise cannot be evaluated.
  - *Indeterminate: Medically Ambiguous:* The order subtype is genuinely ambiguous as to whether it denotes a cross-sectional study.
- **Required Evidence:** The POE radiology orders (subtype, ordertime, order status); the ICU stay window (icustays.intime / outtime / stay_id); the confirming radiology report text and charttime; and, for a multi-stay admission, the attribution of each order to a single stay.
- **Required Grounding:** The specific MIMIC-IV record or record span from which the decisive finding was derived.

Medication Reconciliation Discrepancy. This scenario asks whether an admission contains a clinically significant medication-reconciliation discrepancy across the three medication sources—home medications at arrival, the inpatient record, and the discharge list—distinguishing an unintended discrepancy from a documented intentional change.

- **Clinical Query:** "For admission <subject_id, hadm_id> (with ED stay <stay_id>), reconcile the discharge medication list against the home medications recorded at arrival and the medications administered in the last 24 h. Are there clinically significant reconciliation discrepancies? List each discrepancy with its type and significance; a difference whose intent—a deliberate change versus an unintended omission—cannot be established from the record should be escalated rather than called a discrepancy."
- **Adjudication Criteria:**

  - *Yes:* At least one high-significance unintended discrepancy—an omission of a chronic medication, a home medication held and never restarted, a time-limited course with a missing duration, or an undocumented dose, frequency, or route change—for which no deliberate clinical rationale is documented.
  - *No:* All three sources reconcile, and every difference is either trivial or a documented intentional change.
  - *Indeterminate: Lack of Data:* There is no ED encounter (so no home-medication reconciliation list) or no structured discharge list, so the three-way reconciliation cannot be performed.
  - *Indeterminate: Medically Ambiguous:* A difference exists but its intent—a deliberate change versus an unintended omission—cannot be determined from the record, so that item is escalated rather than adjudicated as a discrepancy.
- **Required Evidence:** Home medications at ED arrival (MIMIC-IV-ED `medrecon`); active inpatient orders and the last-24 h administrations (`prescriptions`, `emar`, `pharmacy`); and the discharge medication list from both the discharge note and `prescriptions`—compared home $\rightarrow$ inpatient $\rightarrow$ discharge, with the last-24 h window measured to `dischtime`.
- **Required Grounding:** The specific MIMIC-IV record or record span from which the decisive finding was derived.

Longitudinal Renal-Function Trajectory. This scenario asks whether a patient's renal function is on a worsening trajectory across all recorded admissions, assessed by estimated GFR rather than serum creatinine alone, since creatinine understates GFR decline as muscle mass falls.

- **Clinical Query:** "For patient <subject_id> with multiple admissions, is renal function on a worsening trajectory (versus stable or improving) across their course? Assess by estimated GFR (via a validated equation such as the 2021 CKD-EPI, or MDRD / Cockcroft–Gault) rather than serum creatinine alone, since creatinine understates GFR decline as muscle mass falls; cite the specific admissions, the values relied on, and the inflection points."
- **Adjudication Criteria:**
  - *Yes:* A sustained decline in per-admission baseline eGFR across ≥ 2 admissions beyond a defined delta, or the onset of kidney replacement therapy (hemodialysis, peritoneal dialysis, or transplantation), or documented CKD-stage progression with concordant labs (or, where available across ≥ 2 admissions, a sustained shift in cystatin-C beyond the delta). A declining-eGFR trajectory is robust to the muscle-mass confound.
  - *No:* Per-admission baseline eGFR is flat or improving across admissions; a discrete, reversible AKI that recovers to baseline does not count as worsening. Where both this criterion and the medical-ambiguity clause below could fit—flat creatinine in an older, frail patient—this *No* criterion governs.
  - *Indeterminate: Lack of Data:* Fewer than two admissions, or creatinine sampling too sparse to construct per-admission baselines; or the trajectory is dominated by a single acute, reversible event with no clear baseline.
  - *Indeterminate: Medically Ambiguous:* Flat or mildly rising creatinine in an older, frail, or long-course patient, where muscle loss means stable creatinine cannot exclude real GFR decline—especially with coded CKD and no cystatin-C or muscle-corrected eGFR available (the usual case in MIMIC-IV,

2008–2019)—provided the per-admission eGFR trajectory is not itself flat or improving; or lab-versus-ICD discordance, where codes assert CKD or progression but creatinine is mild or flat and a human read is needed.

- **Required Evidence:** Per-admission baseline, peak, and discharge creatinine with dates; reno-active medication changes; discharge-summary spans for the inflection points; and any weight/BMI trend and any cystatin-C or eGFR values—with an explicit note when these are absent.
- **Required Grounding:** The specific MIMIC-IV record or record span from which the decisive finding was derived.

B.1 Representative Scenario Selection

Table 13 highlights a representative subset to convey the breadth of the suite. Scenario numbers are consistent with the index in Figure 3.

Table 13. Representative Scenario Selection (Suite Overview)

Scenario	Short Title	Brief Core Objective
8	Antibiotic Stewardship	Evaluate whether empiric therapy covered the cultured organism's susceptibilities and was narrowed within 48 h once a susceptible narrower-spectrum option was available.
10	DKA Resolution Timeline	Reconstruct the diabetic-ketoacidosis resolution timeline and flag protocol deviations (e.g., dextrose not started when glucose dropped below 200 mg/dL, or no SC-insulin overlap before drip discontinuation).
12	Hyperacute Stroke Audit	Compute door-to-needle and door-to-puncture intervals against AHA targets and extract documented justifications for withholding thrombolysis.
13	Hepatic Encephalopathy	Audit narrative and structured data to confirm or rule out the standard HE precipitants (e.g., active infection, GI hemorrhage, electrolyte derangement, lactulose non-adherence).
15	ICU Delirium Screening	Confirm CAM-ICU (or equivalent) screening was documented per shift and, where delirium was present, that an ABCDEF bundle response was applied and deliriogenic agents (benzodiazepines, anticholinergics, meperidine) were avoided or clinically justified.
17	Anticoagulation Reversal	Audit whether reversal matched current guidelines for the specific anticoagulant and bleed (e.g., 4-factor PCC over FFP for warfarin major bleed; agent-specific reversal for DOACs).
19	Hyperkalemia Sequence	Verify adherence to the safety sequence (calcium stabilization before intracellular shifting and elimination) and measure time-to-calcium for critical potassium values ($\geq$ 6.5 mEq/L).
22	End-of-Life Concordance	Audit the final 72 hours of life to determine whether intensity of intervention (ventilation, pressors, dialysis, procedures) was concordant with documented goals of care.
24	Frequent-Admitter Synthesis	Synthesize a single patient's recurrent admissions to extract recurring diagnoses and $\geq$ 1 documented intervenable driver (e.g., medication non-adherence, substance use, housing instability).

C Per-class precision and recall

Table 14. Per-class F1, with precision and recall for the two *Indeterminate* classes. The two rightmost cells read P / R. Denominators follow the cohort's natural label distribution, not the development subset's balanced one, so cells are not comparable with Table 17.

		F1				Precision / Recall %	
Model (harness)	**n**	Yes	No	lack	amb	lack	amb
Opus 5 (Claude Code)	750	0.809	0.771	0.628	0.596	60.0 / 65.9	56.7 / 63.0
Sonnet 5 (Claude Code)	750	0.796	0.749	0.643	0.426	61.5 / 67.5	38.2 / 48.1
GPT-5.6-Sol (Codex)	750	0.795	0.740	0.619	0.500	57.7 / 66.7	45.5 / 55.6
GPT-5.6-Luna (Codex)	750	0.739	0.708	0.494	0.411	45.8 / 53.7	32.6 / 55.6
GPT-5.5 (Codex)	750	0.830	0.762	0.605	0.600	62.6 / 58.5	65.2 / 55.6
GPT-5.4 (Codex)	750	0.775	0.736	0.555	0.464	53.4 / 57.7	44.8 / 48.1
GPT-5.4-mini (Codex)	750	0.764	0.678	0.442	0.353	48.5 / 40.7	29.3 / 44.4
Gemini-3.6-Flash (Gemini CLI)	750	0.801	0.714	0.621	0.473	66.1 / 58.5	46.4 / 48.1
Gemini-3.5-Flash (Gemini CLI)	750	0.786	0.714	0.611	0.377	66.0 / 56.9	38.5 / 37.0
Gemini-3.1-Pro (Gemini CLI)	750	0.771	0.733	0.653	0.500	56.5 / 77.2	64.7 / 40.7
DeepSeek-V4-Pro (opencode)	750	0.752	0.695	0.556	0.444	58.6 / 52.8	38.9 / 51.9
DeepSeek-V4-Flash (opencode)	750	0.800	0.705	0.500	0.340	58.1 / 43.9	34.6 / 33.3
GLM-5.2 (opencode)	750	0.804	0.743	0.597	0.426	67.3 / 53.7	38.2 / 48.1
Qwen-3.7-Plus (opencode)	750	0.737	0.678	0.438	0.463	46.4 / 41.5	34.5 / 70.4
MiniMax-M3 (opencode)	750	0.742	0.690	0.408	0.380	52.6 / 33.3	28.8 / 55.6
Kimi-K2.7-Code (opencode)	750	0.751	0.699	0.431	0.350	54.3 / 35.8	26.4 / 51.9

D Verdict accuracy per scenario

Table 15. Verdict accuracy (%) per scenario, all 16 systems. Rows are the 25 clinical scenarios, sorted by the unweighted mean over systems given in the last column; the best cell in each row is bold. A cell's denominator is that system's graded runs for that scenario, exactly 30 for every cell in the grid, so a single cell is a weaker estimate than a row or a column. Column key: O5 = Opus 5, S5 = Sonnet 5, Sol = GPT-5.6-Sol, Luna = GPT-5.6-Luna, 5.5 = GPT-5.5, 5.4 = GPT-5.4, 5.4m = GPT-5.4-mini, G3.6 = Gemini-3.6-Flash, G3.5 = Gemini-3.5-Flash, G3.1 = Gemini-3.1-Pro, D4P = DeepSeek-V4-Pro, D4F = DeepSeek-V4-Flash, GLM = GLM-5.2, Qwn = Qwen-3.7-Plus, MM3 = MiniMax-M3, K2.7 = Kimi-K2.7-Code.

Scenario	O5	S5	Sol	Luna	5.5	5.4	5.4m	G3.6	G3.5	G3.1	D4P	D4F	GLM	Qwn	MM3	K2.7	Mean
eol-goal-concordance	93	93	**100**	97	93	97	93	90	90	**100**	77	93	90	83	87	90	91.7
overdose-disposition	87	**97**	87	83	90	90	**97**	**97**	**97**	90	93	**97**	90	80	80	**97**	90.6
renal-trajectory	87	93	90	90	93	90	**97**	83	93	90	90	87	83	87	90	83	89.2
transfusion-audit	87	**90**	87	83	**90**	**90**	87	**90**	**90**	**90**	87	**90**	**90**	83	**90**	83	87.9
discharge-audit	**93**	77	80	83	90	90	63	80	83	90	77	73	80	80	80	83	81.5
advanced-imaging	83	87	87	**90**	83	77	67	87	87	77	80	73	87	77	73	77	80.6
hfref-gdmt	80	**87**	83	63	83	73	70	83	83	77	80	73	80	73	70	70	76.9
post-ct-aki	83	83	90	**93**	90	73	80	73	77	73	60	77	67	60	80	67	76.7
sepsis-bundle	83	80	83	80	83	77	77	83	80	**87**	67	73	83	67	53	60	76.0
goals-of-care	93	90	90	60	87	83	57	67	70	**97**	87	77	83	63	57	53	75.8
frequent-admitter	77	70	83	77	77	63	63	**87**	77	83	67	73	83	77	63	67	74.2
readmission-30d	70	70	73	80	80	77	67	77	70	**83**	73	63	73	70	70	60	72.3
icu-delirium	67	**83**	60	63	63	67	67	77	73	63	60	60	77	67	63	80	68.1
sepsis-source-abx	**80**	60	63	43	70	67	73	**80**	67	73	57	77	77	57	67	47	66.0
med-reconciliation	**87**	70	73	57	60	60	53	60	70	83	73	53	83	57	57	60	66.0
postop-complications	53	60	57	63	70	67	67	70	67	67	70	**77**	63	63	67	67	65.4
pe-workup	**77**	63	70	63	67	63	50	63	63	67	63	67	57	67	57	57	63.3
dka-resolution	83	57	87	60	**90**	77	23	73	67	60	53	47	70	43	53	63	62.9
hyperkalemia-management	60	57	63	60	70	70	67	47	53	53	67	60	**80**	63	63	67	62.5
ams-workup	60	**80**	63	50	67	67	50	70	60	57	60	50	67	63	53	73	61.9
empiric-abx	70	**73**	63	57	70	63	63	63	63	67	63	60	67	50	40	43	61.0
he-precipitant	57	53	60	57	60	57	67	67	63	63	60	67	63	57	53	**73**	61.0
stroke-time-targets	67	57	53	50	63	57	53	70	**73**	53	63	67	57	53	67	60	60.2
anticoag-reversal	67	70	60	40	60	57	47	67	70	57	53	**80**	50	57	63	40	58.5
pneumonia-curb65	47	47	23	27	53	27	**63**	20	10	17	37	53	40	37	50	43	37.1

E Development-subset experiments

The main text reports the 750-case cohort. This appendix carries the experiments measured on the 148-case *development subset*. The *dev subset* is a frozen subset of 148 of the 750 cases, built by taking up to two patients per verdict class from each scenario. It over-samples the two rare *Indeterminate* classes (gold supports *Yes* 50, *No* 45, *lack* 39, *amb* 14), so an intervention aimed at deferral can move a measurable number of cases. Its case ids are a verified subset of the 750.

E.1 Sixteen-system comparison and per-class detail

Sixteen systems, and the dissociations rather than the ranking. Accuracy on the subset spans 59.5–71.6 and does not resolve at the top (Table 16): Opus 5 leads at 71.6 with GPT-5.5 0.7 behind and Gemini-3.1-Pro 0.7 behind that, three systems from three harness families inside 1.4 points — well under the $\pm$3.7-point standard error at this n. The subset is too small to separate them: of the $\binom{16}{2} = 120$ pairs, only 11 reach $p < 0.05$ under two-sided exact McNemar on paired per-case outcomes and *none* survives Holm correction.

Table 16. Sixteen-system comparison, development subset. *148 cases, single attempt, clinician-calibrated reference verdicts; all arms complete.* Columns as in Table 3. ECE is expected calibration error (Eq. 1); it is not comparable across all arms, because stated confidence is near-degenerate for the Gemini CLI arms — one emits three distinct values across 148 reports — so those figures indicate an absence of graded uncertainty rather than a finer ranking. Rows are obtained by *slicing* each arm's completed 750-case run to the 148 subset ids, so the Claude arm here is Opus 5 — unlike the paired ablations of Section 6.6, which reuse the earlier Opus 4.8 runs.

Model (harness)	n	Acc	Macro-F1	Def-free	Gap	Proc%	ECE	Over-com	Over-abs
Opus 5 (Claude Code)	148	**71.6**	**0.703**	66.9	4.7	**80.7**	**0.096**	35.8	9.5
Sonnet 5 (Claude Code)	148	66.2	0.626	56.8	9.5	68.4	0.119	35.8	12.6
GPT-5.6-Sol (Codex)	148	68.2	0.666	62.8	5.4	77.4	0.239	37.7	11.6
GPT-5.6-Luna (Codex)	148	61.5	0.594	55.4	6.1	73.1	0.278	39.6	16.8
GPT-5.5 (Codex)	148	70.9	0.687	**68.2**	**2.7**	74.5	0.135	47.2	7.4
GPT-5.4 (Codex)	148	68.2	0.643	62.8	5.4	72.7	0.174	35.8	8.4
GPT-5.4-mini (Codex)	148	61.5	0.549	54.1	7.4	63.3	0.253	56.6	11.6
Gemini-3.6-Flash (Gemini CLI)	148	67.6	0.649	54.7	12.8	70.2	0.302	43.4	**5.3**
Gemini-3.5-Flash (Gemini CLI)	148	67.6	0.626	56.8	10.8	71.4	0.301	47.2	6.3
Gemini-3.1-Pro (Gemini CLI)	148	70.3	0.668	60.1	10.1	56.2	0.273	**26.4**	7.4
DeepSeek-V4-Pro (opencode)	148	64.9	0.613	59.5	5.4	67.4	0.228	41.5	8.4
DeepSeek-V4-Flash (opencode)	148	63.5	0.557	55.4	8.1	66.3	0.298	56.6	6.3
GLM-5.2 (opencode)	148	66.9	0.631	58.8	8.1	74.6	0.188	45.3	9.5
Qwen-3.7-Plus (opencode)	148	59.5	0.565	54.1	5.4	66.1	0.240	45.3	15.8
MiniMax-M3 (opencode)	148	60.1	0.552	52.0	8.1	72.1	0.244	56.6	11.6
Kimi-K2.7-Code (opencode)	148	60.8	0.557	48.0	12.8	66.3	0.212	60.4	9.5

Over-commitment exceeds over-abstention in every system without exception. Measured directionally (last two columns of Table 16), over-commitment runs 26.4–60.4% against 53 deferral cases while over-abstention runs 5.3–16.8% against 95 definitive ones, and *all sixteen* systems over-commit at the higher rate — reproducing Section 6.2 on a deliberately different label distribution. The two directions are not inverses: Gemini-3.6-Flash posts the lowest over-abstention (5.3%) alongside a high over-commitment

(43.4%), whereas Gemini-3.1-Pro achieves the lowest over-commitment (26.4%) without a corresponding penalty (7.4%), so a system can improve on one axis without paying on the other.

Ambiguity elicitation responds to intervention; benchmark difficulty does not. Revising the ambiguity guidance in the case prompts raised pooled recall on the *amb* class roughly threefold, improving in every arm measured with none regressing, while accuracy over the same change stayed flat and no per-arm shift survived correction. That measurement was taken on the pre-revision prompt build, before the label revision described in Section 6.6, so we report its direction and withhold the figures rather than invite comparison with the tables here. Elicitation improved; the benchmark's difficulty did not.

Per-class detail. Table 17 gives per-class F1 across the sixteen arms, alongside the macro average that the leaderboard reports. *lack* F1 spreads the field more widely than accuracy does — 0.62–0.73 for the strongest arms against 0.35–0.47 for the weakest — on a safety-critical class. It does *not*, however, separate the tiers: GPT-5.4-mini reaches only 0.441, below both GLM-5.2 (0.603) and DeepSeek-V4-Pro (0.597), so the vendor and open-weight ranges overlap here as they do on accuracy. At an *amb* support of 14 one case moves a cell by 7.1 points, so that column supports no ranking.

Table 17. Per-class F1, development subset. All arms $n = 148$; gold supports *Yes* 50, *No* 45, *lack* 39, *amb* 14. Macro is the unweighted mean of the four, so the two rare *Indeterminate* classes carry equal weight with the common ones. Ordered as in Table 16.

Model (harness)	Yes	No	lack	amb	Macro
Opus 5 (Claude Code)	0.762	**0.716**	0.667	**0.667**	**0.703**
Sonnet 5 (Claude Code)	0.729	0.644	0.649	0.480	0.626
GPT-5.6-Sol (Codex)	0.772	0.633	0.649	0.609	0.666
GPT-5.6-Luna (Codex)	0.688	0.627	0.541	0.519	0.594
GPT-5.5 (Codex)	**0.808**	0.712	0.562	**0.667**	0.687
GPT-5.4 (Codex)	0.752	0.700	0.620	0.500	0.643
GPT-5.4-mini (Codex)	0.707	0.691	0.441	0.357	0.549
Gemini-3.6-Flash (Gemini CLI)	0.766	0.653	0.594	0.583	0.649
Gemini-3.5-Flash (Gemini CLI)	0.729	0.706	0.615	0.455	0.626
Gemini-3.1-Pro (Gemini CLI)	0.738	0.681	**0.725**	0.526	0.668
DeepSeek-V4-Pro (opencode)	0.722	0.653	0.597	0.480	0.613
DeepSeek-V4-Flash (opencode)	0.765	0.646	0.483	0.333	0.557
GLM-5.2 (opencode)	0.734	0.688	0.603	0.500	0.631
Qwen-3.7-Plus (opencode)	0.713	0.592	0.469	0.485	0.565
MiniMax-M3 (opencode)	0.692	0.667	0.377	0.471	0.552
Kimi-K2.7-Code (opencode)	0.738	0.636	0.353	0.500	0.557

E.2 Policy grounding on the development subset

Table 18 reports the policy axes on the development subset, where every arm is complete at $n = 148$. It is separated from the cohort result for the reasons above; Table 6 carries the figures the main text uses.

Table 18. Policy grounding, development subset. Document-set precision and recall are deterministic set membership against the expert-specified governing set; support is the LLM-judged check that a cited passage entails its claim. *Fab.* counts citations resolving to no corpus document. Citation resolution (0.95–1.00 for every arm) and F1 (determined by the two columns beside it) are omitted as non-discriminating.

Model (harness)	Support	Doc-P	Doc-R	Fab.
Opus 5 (Claude Code)	0.551	0.753	**0.686**	0
Sonnet 5 (Claude Code)	0.566	0.802	0.421	6
GPT-5.6-Sol (Codex)	0.732	0.820	0.575	0
GPT-5.6-Luna (Codex)	**0.834**	0.804	0.498	3
GPT-5.5 (Codex)	0.744	0.764	0.562	0
GPT-5.4 (Codex)	0.632	0.779	0.531	0
GPT-5.4-mini (Codex)	0.634	0.773	0.461	0
Gemini-3.6-Flash (Gemini CLI)	0.454	0.771	0.635	16
Gemini-3.5-Flash (Gemini CLI)	0.474	0.737	0.578	71
Gemini-3.1-Pro (Gemini CLI)	0.652	0.792	0.396	0
DeepSeek-V4-Pro (opencode)	0.605	0.861	0.517	3
DeepSeek-V4-Flash (opencode)	0.585	0.842	0.418	25
GLM-5.2 (opencode)	0.569	0.812	0.562	16
Qwen-3.7-Plus (opencode)	0.582	**0.870**	0.401	1
MiniMax-M3 (opencode)	0.453	0.799	0.543	29
Kimi-K2.7-Code (opencode)	0.651	0.806	0.532	0