CLINENV: AN INTERACTIVE MULTI-STAGE EHR ENVIRONMENT FOR AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decision-making requires gathering heterogeneous information and selecting actions across a hospital stay. We present CLINENV, a benchmark for staged decision reconstruction and information acquisition over real inpatient admissions. An automated pipeline converts EHR timelines into ordered decision stages; models query available patient, nurse, laboratory, and history agents before submitting decisions with supplied type and count hints. The runtime replays recorded care so that later clinical states remain fixed across models. CLINENV pairs ontology-grounded decision concordance with information coverage and estimated resource-use metrics. Across seven models in Interactive Mode, the strongest reaches 0.31 decision F1 and recovers diagnosis targets more reliably than management targets (0.51 vs. 0.17 F1). Stage-level decision F1 and coverage have little linear association, while all models score lower on long-horizon admissions. These complementary measurements expose differences that endpoint scores alone omit. The benchmark evaluates concordance with recorded care; it does not establish clinical correctness or the consequences of alternative treatment policies.

1 INTRODUCTION

Large language models now approach or exceed expert performance on medical licensing examinations (Singhal et al., 2023; Nori et al., 2023) and structured clinical question answering (Jin et al., 2021). Inpatient care also requires gathering heterogeneous information and selecting actions at successive points in a hospitalization. Static multiple-choice benchmarks do not measure these information-seeking and staged decision behaviors.

Coding agents illustrate how evaluation infrastructure drives progress, moving from static functional tests (Chen et al., 2021) to executable interactive environments (Jimenez et al., 2024). Medical evaluation has begun a similar shift. EHR query benchmarks (Lee et al., 2022; Jiang et al., 2025) score specified data-access operations; diagnostic-conversation benchmarks (Schmidgall et al., 2024; Luo et al., 2026) evaluate dialogue over simulated patients; and code-centric environments (Xu et al., 2025; Mao et al., 2025) evaluate biomedical code and calculations. CLINENV targets a complementary setting: interactive reconstruction of recorded decisions across real, multi-stage admissions.

CLINENV combines automated case construction, role-specific information access, and separate decision and process measurements (Figure 1). Its runtime uses **controlled longitudinal replay**: prior information-seeking dialogue is retained, prior model submissions are removed, and the recorded decision is inserted before the next stage. Thus, clinical history follows the observed EHR rather than a counterfactual response to model actions. Type and count hints isolate selection of clinical items from deciding whether, what type, or how many actions to initiate. Decision scores measure concordance with recorded care, which may differ from clinical appropriateness.

Across seven models in Interactive Mode, the best overall decision F1 is 0.306; diagnosis and management F1 for that model are 0.511 and 0.172. Stage-level coverage and decision F1 are weakly correlated, making information acquisition a complementary measurement within this setting. A paired Direct Mode evaluation and clinician validation remain necessary to distinguish interface effects from reasoning difficulty and calibrate absolute scores. Our contributions are:

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

Table 1: **Design dimensions of medical LLM evaluation.** ✓ = supported; △ = partial, mixed, or a related setting; ✗ = outside the cited evaluation setting. Multi-stage denotes successive clinical decision targets within an admission, rather than multiple dialogue or tool turns. Definitions and qualifications appear in Appendix A.

Benchmark	Real EHR	Multi-stage	Auto. build	Active query	Struct. action	Rule score	Process / cost
Static medical knowledge benchmarks							
MedQA (Jin et al., 2021)	✗	✗	✗	✗	✗	✓	✗
MedMCQA (Pal et al., 2022)	✗	✗	✗	✗	✗	✓	✗
PubMedQA (Jin et al., 2019)	✗	✗	✗	✗	✗	✓	✗
MultiMedQA (Singhal et al., 2023)	✗	✗	✗	✗	✗	△	✗
EHR query and action benchmarks							
EHRSQL (Lee et al., 2022)	✓	✗	△	✗	△	✓	✗
MIMIC-SQL (Wang et al., 2020)	✓	✗	△	✗	△	✓	✗
FHIR-AgentBench (Lee et al., 2025)	✓	✗	△	✓	△	✗	✓
MedAgentBench (Jiang et al., 2025)	✓	✗	✗	✓	✓	✓	✗
Diagnostic dialogue and patient-interaction evaluations							
AgentClinic (Schmidgall et al., 2024)	△	✗	△	✓	△	✗	✓
MedDialBench (Luo et al., 2026)	✗	✗	✗	✓	✗	✗	✓
MAQuE (Gong et al., 2025)	✗	✗	✗	✓	✗	✓	✓
HELPMed (Bean et al., 2025)	✗	✗	✗	✓	✗	△	△
AMIE (Tu et al., 2024)	✗	✗	✗	✓	✗	✗	✓
Executable calculation and biomedical coding environments							
MedCalc-Env (Mao et al., 2025)	✗	✗	✗	✗	△	✓	✗
MedAgentGym (Xu et al., 2025)	△	✗	△	✓	△	✓	✓
CLINENV (Ours)	✓	✓	✓	✓	✓	△	✓

(i) Automated construction of source-traceable, multi-stage cases from real EHR admissions without manual case annotation (Section 3.2);

(ii) An interactive environment that exposes heterogeneous clinical information through role-specific agents under controlled longitudinal replay (Section 3.3, 4);

(iii) Separate measurements of recorded-decision concordance, information acquisition, and estimated resource use (Section 5).

2 RELATED WORK

Table 1 compares seven design dimensions; Appendix A defines their scope and summarizes individual protocols.

Knowledge and EHR tasks. MedQA, MedMCQA, PubMedQA, and MultiMedQA evaluate medical or biomedical question answering from supplied context. EHRSQL and MIMIC-SQL instead test query execution over patient records. FHIR-AgentBench adds iterative retrieval and separately measures retrieval quality and answer correctness (Lee et al., 2025); MedAgentBench evaluates physician-authored information-retrieval and order-placement tasks (Jiang et al., 2025). These settings exercise data access and clinical reasoning within specified tasks, without reconstructing successive decision targets across an admission.

Diagnostic dialogue. AgentClinic combines patient and measurement agents, including an EHR-derived case subset (Schmidgall et al., 2024). MedDialBench and MAQuE evaluate patient-behavior variation, information coverage, and inquiry efficiency (Luo et al., 2026; Gong et al., 2025). HELPMed studies human–LLM medical self-assessment, while AMIE evaluates diagnostic dialogue with patient actors and clinicians (Bean et al., 2025; Tu et al., 2024). These studies establish

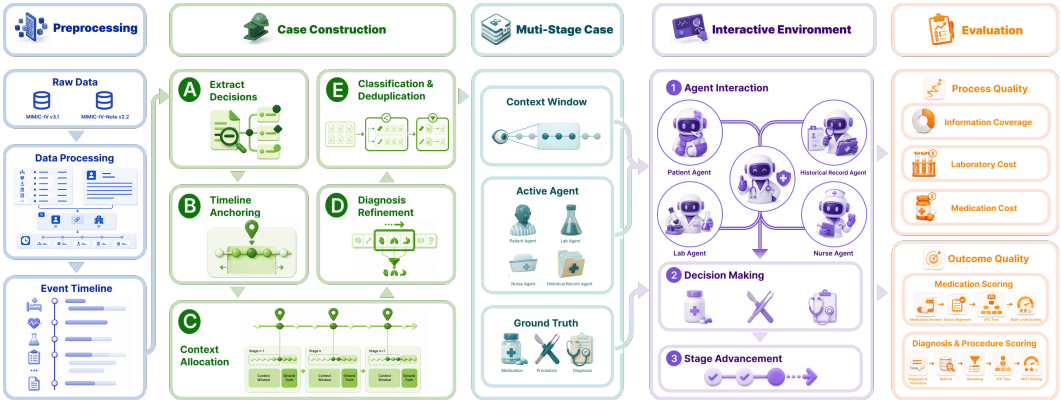

Figure 1: **Overview of CLINENV.** Patients' admissions are preprocessed into event timelines, converted into multi-stage cases via a five-step pipeline, and evaluated in an interactive environment where the model queries specialized agents before committing decisions, scored on both process and outcome quality.

interaction quality as an evaluation target; CLINENV extends the setting to multiple recorded diagnosis, medication, and procedure targets within an admission.

Executable tasks and evaluation scope. MedCalc-Env provides verifiable medical calculations; MedAgentGym supports executable biomedical code and reports interaction and code-efficiency measures (Mao et al., 2025; Xu et al., 2025). CLINENV combines automated EHR case construction, staged targets, structured submissions, and separate decision and process measurements. Its distinction is this evaluation setting, rather than interaction or process scoring alone. Matching is rule-based after LLM-mediated normalization; coverage also uses an LLM judge. The resulting scores measure recorded-care concordance under controlled replay, without establishing clinical correctness or counterfactual treatment effects.

3 CLINENV CONSTRUCTION

3.1 DATA PREPROCESSING

CLINENV is built from MIMIC-IV v3.1 (Johnson et al., 2023b) and MIMIC-IV-Note v2.2 (Johnson et al., 2023a). For each hospitalization, we combine structured EHR tables with clinical notes into an admission-level timeline by `subject_id` and `hadm_id`. Structured sources include admissions, laboratory and microbiology results, medication orders and administrations, ICU charted events, fluid input/output events, procedure events, and ICD-coded diagnoses and procedures; note sources include discharge notes and radiology reports. Each retained row becomes a provenance-bearing event with a canonical `event_time`, `source_table`, and the original table-specific clinical fields. Events are sorted into one timeline per admission. Notes supply narrative context for decision extraction and history access, while structured records provide timestamped observations and coded ground truth for evaluation. Appendix B describes the preprocessing procedure and examples.

3.2 AUTOMATED CASE CONSTRUCTION

Each admission is processed by a five-phase pipeline that segments the patient timeline into an ordered sequence of clinical stages, with ground-truth decision sets derived entirely from structured records and discharge documentation. Claude-Sonnet-4.6 is used as the planning backbone throughout.

Phase A: Decision Extraction. The planner reads the discharge note to extract candidate physician decisions recorded during the hospitalization. When a full free-text discharge note is available it is used directly; otherwise, four sections are stitched together as a fallback: Chief Complaint, History of Present Illness, Brief Hospital Course, and Discharge Diagnosis. Each decision is annotated with a type (medication, procedure, or plan).

Table 2: Clinical information agents and their readviews.

Agent	Readview $\mathcal{V}_a$	Role
Patient	Demographics, chief complaint, history of present illness, past medical history	Simulates subjective patient reporting in lay language
Nurse	Vitals, fluid balance, medication administration records	Reports bedside objective observations
Lab	Laboratory and microbiology results	Resolves test-name queries and returns test results
History	Prior discharge summaries	Supplies longitudinal clinical context

Phase B: Temporal Anchoring. Given a timeline of T events and decisions $d_1, \ldots, d_n$, Phase B localizes each decision to a specific event or event range via a sequential forward sliding-window agent. Let $\ell_0 = 0$. For decision d_i, a window

$$\mathcal{W}(s) = \big[s,\ \min(s + W - 1,\ T - 1)\big] \tag{1}$$

is presented to an LLM agent equipped with `mark_single` and `mark_range` tool calls, beginning at $s = \ell_{i-1}$. If the agent marks anchor a_i within the window, the anchor is recorded and $\ell_i = a_i$; the next decision d_{i+1} begins its search from $\ell_i + 1$. If no anchor is found, the window advances ($s \leftarrow s + W$) and the agent retries. If the timeline is exhausted without a match, d_i is left unanchored and $\ell_i = \ell_{i-1}$. Decisions anchored exclusively to pre-admission events are discarded.

Phase C: Stage Construction and Validation. Given anchor index d_i for decision i, the context window for stage k is defined as $\mathcal{C}_k = [prev_end + 1,\ d_i - 1]$, where $prev_end$ is the last event index consumed by the preceding stage. A decision is skipped if (a) no context window exists ($prev_end \geq d_i - 1$), or (b) the decision is a medication whose drug is already actively present in $\mathcal{C}_k$. For final diagnosis stages, context spans the events from the last decision through discharge, and targets are derived from a separate diagnostic scan. The worked trace in Appendix E contains procedure and medication stages; it does not illustrate the final diagnosis-stage construction.

Phase D: Diagnostic Groundability Scan. A sliding-window scan identifies which of the admission's discharge ICD codes are groundable, i.e., actively evidenced or managed within the event timeline. Given a timeline of T events, we partition it into overlapping windows:

$$\mathcal{W}_i = \big[iS,\ \min(iS + W,\ T)\big), \quad i = 0, 1, \ldots, \left\lceil \frac{T - W}{S} \right\rceil \tag{2}$$

where W is the window size and $S < W$ is the stride. Only ICD codes confirmed by at least one window contribute to the final diagnosis stage.

Phase E: Classification and Enrichment. Each stage's raw decision description is classified into one of four types (medication, procedure, diagnosis, plan) via LLM. For medication, the action type is further assigned as one of {`start`, `stop`, `switch`, `adjust`}; `adjust` decisions optionally carry a direction (`increase` or `decrease`). Ground-truth fields are then enriched deterministically from EHRs with zero LLM calls: drug names are read from prescription, pharmacy, and eMAR tables; ICD codes for procedures and diagnoses are taken directly from the corresponding coded EHR events. The resulting stage $\mathcal{S}_k$ is defined as:

$$\mathcal{S}_k = (\mathcal{C}_k,\ \mathcal{G}_k,\ \mathcal{A}_k) \tag{3}$$

where $\mathcal{C}_k = [start_k, end_k]$ is the visible event window, $\mathcal{G}_k = \{g_1^k, \ldots, g_{m_k}^k\}$ is the structured ground-truth decision set, and $\mathcal{A}_k \subseteq \{\texttt{patient}, \texttt{nurse}, \texttt{lab}, \texttt{history}\}$ is the set of activated information sources for this stage.

The evaluation environment exposes clinical information through four specialized agents (Table 2), each powered by GPT-5.4-mini and restricted to a designated readview $\mathcal{V}_a$, which is a filtered projection of the EHR events relevant to that role.

3.3 MULTI-SOURCE CLINICAL ENVIRONMENT

The patient agent answers in lay language from demographics and clinical history; the nurse agent reports bedside vital signs, fluid balance, and medication administration; the lab agent resolves natural-language test queries against the stage's available results via semantic matching, returning nothing for tests not on record; and the history agent supplies prior discharge summaries when earlier admissions exist. Table 2 lists each readview, and Appendix C.4 gives the full agent specifications.

3.4 INTERACTION MODES

The evaluation interface exposes the information-gathering and decision-submission tools summarized in Table 3.

Table 3: Available tools by evaluation mode.

Tool	Mode	Function
`ask_patient`	Interactive	Query patient agent
`ask_nurse`	Interactive	Query nurse agent
`order_lab`	Interactive	Query lab agent
`get_history_summary`	Interactive	List prior admissions
`get_history_detail`	Interactive	Retrieve full discharge note
`submit_medication`	Both	Submit medication decision
`submit_diagnosis`	Both	Submit diagnosis
`submit_procedure`	Both	Submit procedure decision
`submit_plan`	Both	Submit other management decision
`finalize_decision`	Both	Advance to next stage

Direct Mode. The model to be evaluated receives the complete observation for stage k:

$$\mathcal{O}_k^{\text{direct}} = \{\, \mathcal{P},\ \mathcal{E}_{[start_k,\, end_k]} \,\} \tag{4}$$

where $\mathcal{P}$ denotes static patient context and $\mathcal{E}_{[start_k, end_k]}$ is the event sequence in the specified stage window. The model submits decisions directly without gathering information. Direct Mode is implemented but is not evaluated in the experiments reported here.

Interactive Mode. The model accesses current-stage clinical information through tool calls, while prior-stage dialogue and recorded decisions provide carried-forward context. The current-stage query-response trace at turn t is:

$$\mathcal{O}_k^{\text{interactive}}(t) = \{\, (q_1, r_1),\ (q_2, r_2),\ \ldots,\ (q_t, r_t) \,\} \tag{5}$$

where q_i is the i-th tool call and r_i the corresponding agent response. Exactly one tool call is permitted per turn. The stage terminates when the model calls `finalize_decision` or after 60 turns, whichever comes first. Finalized submissions cannot be edited; they do not alter subsequent patient states. Appendix C.6 details the replay mechanism.

Structured Submission Interface. To isolate clinical reasoning quality from output-format confounds, the submission interface is constrained in two coupled ways. First, only the submit tools corresponding to ground truth types present in $\mathcal{S}_k$ are exposed; types absent from $\mathcal{G}_k$ are withheld entirely. Second, the model receives the exact required submission count per type:

$$n_k^{(t)} = \left|\{g \in \mathcal{G}_k : g.\text{type} = t\}\right|, \quad t \in \mathcal{T}_k \tag{6}$$

where $\mathcal{T}_k$ denotes the set of ground truth types in stage $\mathcal{S}_k$. Because the model submits exactly $n_k^{(t)}$ items per type, the per-type Hungarian matching operates on equal-sized sets, so precision and recall differ only by match quality rather than count mismatch. This factorized task measures selection of recorded clinical items conditional on the supplied types and counts. Action initiation, abstention, type inference, and count prediction are outside its scope.

4 CLINENV BENCHMARK

The CLINENV benchmark is publicly available on Physionet,[1] and the source code repository is accessible on GitHub.[2] We also release a demo website for interactive evaluation.[3] We sample 3,509 admissions from 1,809 patients, keeping only admissions that produce at least one valid decision stage and stratifying by horizon so that short and long admissions are both well represented. This gives 9,297 decision stages and 26,043 ground-truth decisions (Table 4), with 2.65 stages per case and

[1] CLINENV Physionet Link; for anonymity, please refer to `dataset.zip` in the submission system.

[2] CLINENV GitHub Link; for anonymity, please refer to `software.zip` in the submission system.

[3] CLINENV Interactive Evaluation Demo Link

49.8% of cases long-horizon ($\geq$3 stages). Diagnoses account for most decisions (71.7%), followed by medications (21.4%) and procedures (6.9%); each medication carries an action type (`start`, `stop`, `switch`, `adjust`) for the action-gated scoring of Section 5.1. Because the cases come from real admissions rather than a curated question bank, CLINENV spans the breadth of inpatient practice: 2,128 distinct ICD codes and 488 distinct drugs, with 3.67 information agents available per stage and 32.8 key clinical facts per readview on average. Appendix D gives the sampling protocol and construction-quality checks; Figures 4 and 5 further visualize the clinical entity distribution and case-length structure.

5 CLINENV EVALUATION

5.1 RECORDED-DECISION CONCORDANCE

Matching is performed independently within each type group (diagnosis, medication, procedure) via the Hungarian algorithm (Kuhn, 1955), which maximises the total matched score within each group. The results are then concatenated to form the full stage-level assignment. Stage-level F1 is the harmonic mean of continuous precision and recall derived from the matched scores. The reported score covers diagnoses, medications, and procedures. Although the interface also supports `plan` submissions, no separate plan result is reported. Determinism applies to matching after code mapping; free-text mapping includes LLM reranking. Ontology proximity does not establish clinical substitutability or optimal care.

Table 4: CLINENV benchmark stats.

Property	Value
Admissions	3,509
Unique patients	1,809
Cases	3,509
single-stage	718 (20.5%)
two-stage	1,043 (29.7%)
long-horizon ($\geq$3)	1,748 (49.8%)
Decision stages (total)	9,297
Decisions (total)	26,043
diagnosis	18,664 (71.7%)
medication	5,571 (21.4%)
procedure	1,808 (6.9%)
Medication action types	
start / stop	79.1% / 7.5%
switch / adjust	3.8% / 9.5%
Distinct ICD codes	2,128
Distinct drugs	488
Avg. stages per case	2.65
Avg. decisions per stage	2.80
Avg. available agents / stage	3.67
Avg. key facts per readview	32.8

Medication Scoring. Action type serves as a hard gate (`start`/`stop`/`switch`/`adjust`) within the medication group: a submission whose action disagrees with its matched ground truth scores zero regardless of drug similarity, driving the Hungarian algorithm to prefer action-consistent assignments.

When actions agree, drug similarity is scored via ATC-hierarchy partial credit (Chen et al., 2012). An exact name match yields 1.0 directly. Otherwise, both names are resolved to ATC codes via the RxNorm API. Let $\hat{c}$ and c^* denote the predicted and ground-truth ATC codes, and ℓ the length of their longest common prefix.

$$s_{ATC} = \begin{cases} 1.0 & \ell \geq |c^*| \\ 0.8 & \ell = 5 \text{ (chemical subgroup)} \\ 0.6 & \ell = 4 \text{ (pharmacological class)} \\ 0.3 & \ell = 3 \text{ (therapeutic group)} \\ 0.1 & \ell = 1 \text{ (anatomical group)} \\ 0.0 & \text{otherwise} \end{cases} \tag{7}$$

The condition $\ell \geq |c^*|$ grants full credit when the predicted code is at least as specific as the ground truth, accommodating drugs whose RxNorm entry does not reach Level 5. For `adjust` decisions, the final medication score incorporates a direction penalty:

$$s_{med} = \begin{cases} s_{ATC} & \text{if direction matches or is unspecified} \\ 0.5 \cdot s_{ATC} & \text{if directions conflict} \end{cases} \tag{8}$$

where a direction conflict arises only when both the submission and the ground truth explicitly specify opposing directions (increase vs. decrease); if either omits it, no penalty applies.

Diagnosis and Procedure Scoring. Because the model submits free text while ground truth carries structured ICD codes, submitted text is first mapped to a candidate ICD code via a two-step pipeline: (1) an embedding index over the full ICD vocabulary (ICD-9-CM, ICD-10-CM, ICD-9-PCS, ICD-10-PCS) retrieves the top-15 candidates by cosine similarity; (2) an LLM reranker selects the best match.

Table 5: **Overall CLINENV performance.** The long-horizon subset contains admissions with at least three decision stages. Values are stage-macro averages. Dx/Rx/Proc are type-specific match scores; Cov. and Eff. denote information coverage and query-adjusted efficiency. Discord. ratio is the fraction of estimated lab cost associated with requests absent from the recorded stage.

Evaluation setting		Outcome quality				Information seeking			Estimated cost		
Model	**Stage**	**F1↑**	**Dx↑**	**Rx↑**	**Proc.↑**	**Cov.**	**Eff.**	**Turns**	**Discord.%**	**Lab ($)**	**Rx ($/day)**
GPT-5.4*	All	**0.306**	**0.511**	0.097	**0.379**	**0.430**	0.274	9.0	10.4%	117	64
Long Horizon	≥3	**0.235**	**0.507**	0.089	**0.296**	**0.416**	0.221	11.9	**2.3%**	124	87
GPT-5.4-mini*	All	0.258	0.448	0.083	0.295	0.214	0.152	6.8	7.0%	106	72
Long Horizon	≥3	0.213	0.482	0.071	0.281	0.246	0.154	9.0	6.6%	56	109
GPT-5.4-nano*	All	0.254	0.438	0.085	0.322	0.331	**0.281**	**4.4**	**4.1%**	178	50
Long Horizon	≥3	0.215	0.456	0.089	0.270	0.340	**0.257**	**5.8**	7.3%	156	75
Llama-3.1-70B	All	0.226	0.366	**0.101**	0.246	0.140	0.090	14.1	35.8%	**27**	68
Long Horizon	≥3	0.194	0.386	**0.103**	0.212	0.156	0.085	18.1	10.2%	**20**	51
Llama-3.1-8B	All	0.202	0.362	0.082	0.249	0.205	0.134	8.3	6.5%	219	53
Long Horizon	≥3	0.172	0.356	0.076	0.249	0.210	0.115	10.7	3.1%	270	63
Gemma-3-27B	All	0.228	0.386	0.087	0.258	0.180	0.105	7.6	35.3%	132	41
Long Horizon	≥3	0.195	0.415	0.081	0.251	0.201	0.103	8.8	24.9%	147	47
Gemma-3-12B	All	0.197	0.376	0.070	0.235	0.137	0.077	8.8	13.0%	184	**34**
Long Horizon	≥3	0.164	0.400	0.072	0.218	0.136	0.065	10.9	8.5%	218	**31**

* Proprietary LLMs were accessed under a **Zero Data Retention Policy** required by PhysioNet.

The predicted code $\hat{c}$ and ground-truth code c^* are each expanded to their full ancestor sets in the ICD hierarchy, and Hierarchical F1 (HDF1) is computed over these sets:

$$\text{HDF1}(\hat{c}, c^*) = \frac{2\,|\hat{\mathcal{A}} \cap \mathcal{A}^*|}{|\hat{\mathcal{A}}| + |\mathcal{A}^*|} \tag{9}$$

where $\hat{\mathcal{A}} = \{\hat{c}\} \cup \text{anc}(\hat{c})$ and $\mathcal{A}^* = \{c^*\} \cup \text{anc}(c^*)$.

5.2 CLINICAL PROCESS QUALITY

Alongside decision concordance, we measure information acquisition in Interactive Mode using automated trace recall and its query-adjusted efficiency.

Information Elicitation Quality. For each scored role $a \in \mathcal{A}_k^{\text{cov}}$ (activated patient, nurse, and lab roles in the reported implementation), an LLM judge is given the complete readview $\mathcal{V}_a$ and the responses that agent a returned during the stage. The judge enumerates the key clinical facts I_a present in $\mathcal{V}_a$ and marks each fact as covered or not based on whether it appears in the dialogue. Per-role coverage is:

$$\text{cov}_a = \frac{|\{i \in I_a : \text{covered}(i)\}|}{|I_a|} \tag{10}$$

Stage-level coverage is the macro-average across these scored roles:

$$\text{cov} = \frac{1}{|\mathcal{A}_k^{\text{cov}}|} \sum_{a \in \mathcal{A}_k^{\text{cov}}} \text{cov}_a \tag{11}$$

Coverage measures the fraction of available facts surfaced in the dialogue; it neither judges treatment appropriateness nor enters decision F1. Its fact enumeration remains unvalidated by clinicians. The efficiency score used for the reported results is

$$\text{eff} = \text{cov}\,\frac{N}{N+K}, \tag{12}$$

where N is the number of judged key facts and K the number of information-gathering calls. This discounts coverage as query count grows relative to the judged fact count (Appendix C.7).

Record-Discordant Estimated Laboratory Cost. Ordered lab tests are mapped to reference fees from the CMS Clinical Laboratory Fee Schedule through embedding retrieval and LLM reranking. We report the fraction of estimated cost attributable to requests absent from the recorded stage:

$$r_{\text{waste}} = \frac{\sum_{t \in \mathcal{T}_{\text{excess}}} \text{fee}(t)}{\sum_{t \in \mathcal{T}_{\text{ordered}}} \text{fee}(t)}. \tag{13}$$

Here $\mathcal{T}_{\text{ordered}}$ contains ordered tests and $\mathcal{T}_{\text{excess}}$ contains those not supported by the record. We retain r_{waste} as the legacy metric identifier, but interpret it as record discordance: an unrecorded test may still be clinically reasonable. Reference fees provide a standardized resource-use proxy, not observed hospital or patient costs.

Drug Acquisition Cost. For each stage, we estimate a standardized acquisition-cost proxy for newly initiated pharmacotherapy. The daily cost is estimated via a two-step pipeline. First, the drug name is resolved to an RxNorm CUI, which is then mapped to an NDC code and a unit price p from the NADAC (National Average Drug Acquisition Cost) database. Second, the RxNorm CUI is mapped to an ATC code, from which the WHO-defined daily dose (DDD) is retrieved. Each component is matched against the WHO ATC/DDD index via embedding retrieval followed by LLM reranking. The estimated daily cost for a single drug is then: $c = p \times \text{DDD}$ and the stage-level metric is the total estimated daily cost $C = \sum_i c_i$ across all qualifying submissions. NADAC prices and WHO DDD values are reference inputs, not observed local prices or patient-specific prescribed doses.

6 Results

Evaluation scope. All reported runs use Interactive Mode: seven models are evaluated on the same 899 stages, giving 6,293 model-stage observations. These evaluated stages are a subset of the 9,297-stage constructed benchmark. Additional analysis of the saved outputs appears in Appendix G; no Direct Mode or clinician baseline is reported.

6.1 Overall Performance on ClinEnv

Table 5 reports Interactive Mode performance. GPT-5.4 has the highest overall decision F1 (0.306), diagnosis score (0.511), procedure score (0.379), and information coverage (0.430). Llama-3.1-70B has the highest medication score (0.101). The remaining gap persists even with decision-type and count hints.

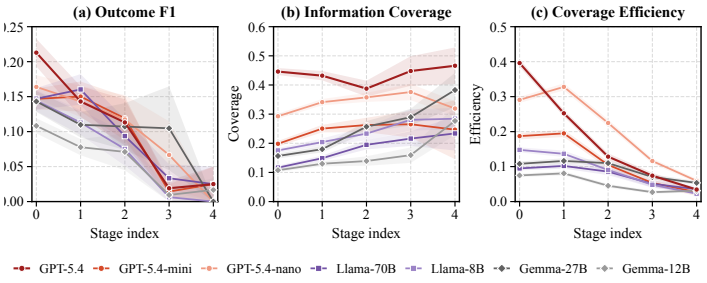

Figure 2: **Per-stage Interactive Mode performance on ClinEnv.** Decision F1 (a), information coverage (b), and coverage efficiency (c) vary across management-stage indices. Shaded bands are standard errors; trajectories are descriptive rather than a paired Direct Mode comparison.

Outcome and process measurements distinguish model behavior. Llama-3.1-70B combines the highest medication score with coverage of 0.140, the most turns (14.1), and the highest record-discordant lab-cost fraction (35.8%). GPT-5.4-nano has the fewest turns (4.4) and lowest discordant fraction (4.1%), with efficiency of 0.281 versus GPT-5.4's 0.274. Gemma-3-12B has the lowest estimated drug cost ($34 per day). Different metrics therefore favor different models.

Every evaluated model has lower decision F1 on the long-horizon subset than on the full set; GPT-5.4 scores 0.235 versus 0.306, respectively. This comparison concerns difficulty along a fixed recorded trajectory, not the consequences of prior model errors.

6.2 Long-Horizon Case Difficulty

Figure 2 traces management stages carrying medication or procedure targets, separating them from diagnosis-only stages. Decision F1 is generally lower at later indices, while coverage and efficiency follow different trajectories. These curves describe the observed subset at each index; they do not establish that stage depth causes errors or that information access is sufficient. The runtime standardizes recorded prior decisions, while each model retains its own information-seeking dialogue.

The weak stage-level association between coverage and decision F1 is examined below. Without a paired full-information reference, these results do not separate interface and information-access effects from decision-reconstruction difficulty.

6.3 Information Coverage and Record-Discordant Cost

Figure 3 shows an inverse association between information coverage and the record-discordant lab-cost fraction across seven model aggregates. GPT-5.4 combines the highest coverage with a 10.4% discordant fraction; Llama-70B and Gemma-27B have lower coverage and fractions above 35%. GPT-5.4-nano pairs moderate coverage with the lowest fraction (4.1%). This descriptive comparison does not establish that increasing coverage reduces clinical waste or that the two metrics measure one underlying skill.

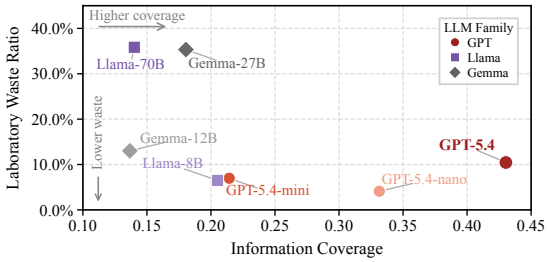

Figure 3: **Information coverage and record-discordant lab cost.** Model aggregates; "waste ratio" denotes record-discordant cost, not clinical waste.

6.4 Diagnosis–Management Concordance Gap

Table 6 separates recorded-decision concordance by stage type. Decision F1 is higher on diagnosis-only stages (0.40 pooled, 0.51 for GPT-5.4) than on management stages (0.13 pooled). The gap is roughly threefold and holds for every model. This difference also explains why combining diagnosis and management stages can obscure management-specific performance (Section 6.2).

Table 6: **Interactive Mode decision-type decomposition.** Diagnosis-only stages score far above management stages. *Lab miss* is the fraction of laboratory queries absent from the stage.

Model	Diag. F1	Mgmt. F1	Action acc.	Med. score	Lab miss
GPT-5.4	**0.511**	**0.172**	0.56	0.097	14%
GPT-5.4-mini	0.440	0.139	0.57	0.083	19%
GPT-5.4-nano	0.416	0.147	0.53	0.085	**11%**
Llama-70B	0.358	0.140	0.70	**0.101**	61%
Llama-8B	0.325	0.121	**0.73**	0.082	60%
Gemma-27B	0.383	0.126	0.56	0.087	73%
Gemma-12B	0.360	0.090	0.68	0.070	66%

Medication matching remains difficult. Models match the recorded action type (`start`, `stop`, `switch`, `adjust`) on 53–73% of medication decisions, yet medication match scores remain at or below 0.101. Llama-8B has the highest action accuracy (0.73) but a medication score of 0.082. Matching the action alone is therefore insufficient for matching the recorded drug. ATC partial credit captures ontology proximity; it does not determine whether an alternative drug is clinically acceptable.

Across the 6,293 saved model-stage observations, decision F1 and coverage have a pooled Pearson correlation of 0.082; per-model correlations range from -0.021 to 0.074. After removing model-by-stage-index means, the pooled within-group correlation is 0.012 (Appendix G). These analyses support complementary measurement within Interactive Mode, without identifying a causal bottleneck. Retrieval is also uneven across sources: GPT-5.4 covers 0.89 of nurse facts and 0.25 of lab facts. Laboratory requests absent from the recorded stage comprise 60–73% for the Llama/Gemma models and 11–19% for the GPT models; absence alone does not imply an inappropriate clinical request.

7 Conclusion

We presented ClinEnv, a source-traceable benchmark for staged decision reconstruction and information acquisition over real inpatient admissions. Controlled longitudinal replay supports comparison against shared recorded targets, while ontology-grounded matching and process metrics describe complementary behaviors. Across seven models in Interactive Mode, the strongest achieves 0.306 decision F1; diagnosis matching exceeds management matching, and every model scores lower on long-horizon admissions. Stage-level coverage has little linear association with decision F1. These results characterize the evaluated replay task; paired Direct Mode comparisons and clinician validation remain needed to interpret information-access effects, acceptable alternatives, and absolute clinical relevance.

AI USE STATEMENT

We used generative AI tools to edit the language of the manuscript for grammar, clarity, and readability. AI assistance in manuscript preparation was limited to editorial tasks; it did not contribute to the research ideas, methodology, experiments, or interpretation of results. The authors take full responsibility for the content of this paper.

REFERENCES

Andrew M Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. Clinical knowledge in llms does not translate to human interactions. *arXiv preprint arXiv:2504.18919*, 2025.

Lei Chen, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou. Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS one*, 7(4):e35254, 2012.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. The dialogue that heals: A comprehensive evaluation of doctor agents' inquiry capability. *arXiv preprint arXiv:2509.24958*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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*, 2(9): AIdbp2500144, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

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

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

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

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.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Jong Ha Lee, et al. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *arXiv preprint arXiv:2509.19319*, 2025.

Xiaotian Luo, Xun Jiang, and Jiangcheng Wu. Meddialbench: Benchmarking llm diagnostic robustness under parametric adversarial patient behaviors. *arXiv preprint arXiv:2604.06846*, 2026.

Kangkun Mao, Jinru Ding, Jiayuan Chen, Mouxiao Bian, Ruiyao Chen, Xinwei Peng, Sijie Ren, Linyang Li, and Jie Xu. Medcalc-eval and medcalc-env: Advancing medical calculation capabilities of large language models. *arXiv preprint arXiv:2510.27267*, 2025.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

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 preprint arXiv:2405.07960*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

Ping Wang, Tian Shi, and Chandan K Reddy. Text-to-sql generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*, pp. 350–361, 2020.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, et al. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *arXiv preprint arXiv:2506.04405*, 2025.

A Extended Related Work

Table 1 compares the cited evaluation protocols, not every capability their underlying models could support. The symbols describe task design rather than a ranking of benchmark quality.

Comparison dimensions. *Real EHR data* indicates patient-record provenance; △ denotes a mixture of EHR-derived and other cases. *Multi-stage* requires successive clinical decision targets within an admission, rather than several turns spent answering one question. *Auto. EHR build* concerns constructing evaluation tasks from EHR records: template-based, adapted, or mixed pipelines receive △; synthetic dialogue or calculation generation alone does not meet this criterion. *Active info-seek* denotes model-selected questions or retrieval actions. *Struct. action* denotes typed clinical submissions; query, test-request, calculation, or code interfaces receive △ when they expose only related or narrower actions. *Rule-based outcome* refers to the task's outcome scorer, separately from process evaluation; △ denotes mixed evaluation or model-mediated normalization before rule-based matching. *Process / cost* covers reported information-acquisition, dialogue-quality, interaction-efficiency, or resource-use measures beyond task success. It does not require all of these measures; △ marks a limited interaction analysis rather than a general trace-level metric suite.

Knowledge benchmarks. MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) contain medical examination questions; the medical subsets of MMLU (Hendrycks et al., 2020) provide related static tasks. PubMedQA (Jin et al., 2019) asks yes/no/maybe research questions from abstracts and distinguishes expert-labeled, unlabeled, and artificially generated subsets. MultiMedQA (Singhal et al., 2023) combines medical and consumer-health QA datasets, including both fixed-answer and long-form evaluations; its outcome scoring is therefore marked as mixed.

EHR queries and actions. EHRSQL (Lee et al., 2022) derives question templates from hospital staff and links them to MIMIC-III and eICU. MIMIC-SQL (Wang et al., 2020) provides machine-generated template questions and human-written natural-language variants over MIMIC-III. Their automatic instantiation coexists with manual question design or annotation. FHIR-AgentBench (Lee et al., 2025) adapts clinician-sourced questions to MIMIC-IV-FHIR and evaluates single- and multi-turn retrieval. Resource-level precision and recall are rule-based, while free-text answer correctness uses an LLM evaluator; the outcome column is consequently not marked as rule-based. MedAgent-Bench (Jiang et al., 2025) uses real, de-identified STARR patient records and physician-authored tasks. Its grader checks query answers against reference solutions and action payloads against manually written rules.

Diagnostic dialogue and human interaction. AgentClinic (Schmidgall et al., 2024) includes exam-derived and MIMIC-IV-derived cases, patient and measurement agents, an LLM moderator for diagnosis assessment, and information-coverage and patient-perception analyses. Its mixed case sources and EHR case adaptation receive partial marks. MedDialBench (Luo et al., 2026) varies patient behavior in OSCE-style cases and reports semantic diagnosis accuracy, information coverage, and coverage per turn. MAQuE (Gong et al., 2025) evaluates inquiry through controlled information disclosure and behavioral variation. It combines string-matched task success and counted inquiry costs with LLM-judged coverage, dialogue quality, and patient experience. HELPMed (Bean et al., 2025) studies members of the public using LLMs for medical self-assessment, with condition and disposition outcomes and analyses of information exchange; its outcome evaluation and interaction analysis are marked as partial or mixed. AMIE (Tu et al., 2024) is evaluated in consultations with patient actors using clinical and patient-centered ratings, including history-taking and communication. These human evaluations are distinct from automated outcome scoring.

Executable environments. MedCalc-Env (Mao et al., 2025) generates calculation problems from configured formulas and scoring rules. Its multi-step numerical reasoning is not a sequence of admission-level clinical decisions, and automatic calculation generation is not EHR case construction. MedAgentGym (Xu et al., 2025) combines EHR and other biomedical tasks in an executable code environment. It reports success based on execution or prediction outcomes and additional interaction and code-efficiency measures; EHR provenance and construction apply to subsets of its task collection. HumanEval (Chen et al., 2021) and SWE-bench (Jimenez et al., 2024) provide broader precedents for executable evaluation, with different outputs and verification contracts.

Table 7: **Source groups used in admission-level preprocessing.** Each row lists a structured or note source, the fields extracted from it, and how those fields enter the admission-level event timeline that downstream case construction and agent readviews draw on.

Group	Tables	Main retained information
Encounter context	`hosp_admissions_df`	Admission/discharge timing, admission type, demographics, insurance, language, and discharge disposition
Laboratory and microbiology	`hosp_labevents_df`, `hosp_microbiologyevents_df`	Test labels, values, units, reference ranges, abnormality flags, specimens, organisms, and comments
Medication orders and administrations	`hosp_prescriptions_df`, `hosp_pharmacy_df`, `hosp_emar_detail_df`	Drug names, dose, route, frequency, order status, administration type, and infusion details
ICU observations and flowsheets	`ehr_chartevents_df`, `ehr_datetime_events_df`, `ehr_inputevents_df`, `ehr_outputevents_df`, `ehr_ingredientevents_df`, `ehr_procedureevents_df`	Vitals, bedside measurements, fluid inputs and outputs, procedure events, ingredients, rates, and durations
Coded outcomes	`hosp_diagnoses_icd_df`, `hosp_procedures_icd_df`	ICD diagnosis and procedure codes, code versions, and mapped long-title descriptions
Clinical notes	`note_df`, `radiology_note`	Discharge summaries, narrative hospital course, chief complaint, history, and radiology report text

Positioning CLINENV. Neither active inquiry nor process evaluation is unique to CLINENV. The benchmark combines automatically constructed admission stages with structured recorded-decision targets, role-specific information access, and process/resource measurements under controlled replay. The partial mark for rule-based outcome scoring makes explicit that normalization remains LLM-mediated even though subsequent matching and ontology scoring are deterministic. Source traceability and this scoring contract do not establish the clinical validity of the targets or acceptable alternatives.

B DATA PREPROCESSING DETAILS

B.1 INPUT SOURCES

Table 7 summarizes the source groups used to build the admission timelines. All tables are linked at the admission level using `subject_id` and `hadm_id`. Rows without a valid admission identifier are excluded before timeline construction.

B.2 TIMELINE CONSTRUCTION

For each admission, preprocessing follows deterministic steps that do not use LLMs:

1. discover the selected structured and note CSV files and stream them in chunks;
2. normalize `subject_id` and `hadm_id`, then drop rows missing either key;
3. assign a canonical `event_time` using the first available timestamp in this priority order: `charttime`, `starttime`, `admittime`, `chartdate`, `stoptime`, `endtime`;
4. serialize each row as an event with `source_table`, `event_time` when available, and table-specific payload fields, while removing identifiers and internal processing columns;
5. group events by admission and sort them by timestamp, `source_table`, and original row order for deterministic tie breaking.

The preprocessor writes one JSON timeline per admission under subject/admission-specific directories. Each file is a JSON array of ordered events. In the preprocessing release used by this work, the

Table 8: **Timeline event schemas after preprocessing.** Representative events from each source. Every event preserves its source provenance and the source-specific clinical fields later used for case construction, agent readviews, and ground-truth scoring.

Event type	Example retained fields	Use in CLINENV
Admission	`source_table`, `event_time`, admission type, race, gender, age	Initial patient context and admission boundary
Lab result	`source_table`, `event_time`, label, value, unit, reference range, flag	Lab-agent responses and evidence for clinical stages
Medication	`source_table`, `event_time`, drug or medication name, dose, route, frequency, status	Medication ground-truth enrichment and nurse/pharmacy evidence
Note	`source_table`, `event_time`, text	Discharge-note decision extraction and prior-admission history
ICD code	`source_table`, ICD code, ICD version, long-title description	Diagnosis and procedure ground truth

resulting timeline collection contains 229,482 admission files across 24,760 subjects and 237,814,500 retained events.

B.3 Event Schema Example

Each event keeps provenance through `source_table` and preserves only fields meaningful for the source row. Table 8 shows schematic examples; patient-specific note text is omitted.

B.4 Worked Preprocessing Example

Consider an admission with an admission row at t_0, laboratory results at t_1, a medication order at t_2, a discharge note at t_3, and ICD-coded diagnoses assigned to the same `hadm_id`. The preprocessor converts these heterogeneous rows into the ordered sequence:

$$[e_{\text{adm}}(t_0),\ e_{\text{lab}}(t_1),\ e_{\text{med}}(t_2),\ e_{\text{note}}(t_3),\ e_{\text{icd}}] \tag{14}$$

where each e retains its original `source_table`. Later pipeline stages use the same sequence differently: Phase A reads the discharge-note event to extract decisions; Phase B anchors those decisions against the ordered timeline; Phase C partitions timeline intervals into stage contexts; and Phase E enriches ground truth from the structured medication, diagnosis, and procedure events.

This separation keeps data preprocessing independent from benchmark scoring. The timeline builder standardizes timing and provenance, while the case-construction pipeline decides which events become patient, nurse, lab, and history readviews and which structured records define evaluation targets.

C Source-Level Implementation Details

This appendix records implementation details from the released CLINENV source pipeline. The goal is to make the benchmark construction and evaluation process auditable without moving code-level detail into the main paper.

C.1 Case Generation Artifacts

The pipeline persists enough state to audit each case without inspecting the raw database again. The planning record stores extracted decisions, timeline anchors, stage context bounds, and enrichment outputs. The prepared case record stores the executable stage sequence, role-specific readviews, available agents, submission hints, and held-out ground-truth items. The model-execution record stores the full dialogue and submitted decisions for each stage, and the evaluation record stores matching assignments and process metrics. The case manifest contains one JSONL record per prepared admission with `subject_id`, `hadm_id`, and the number of constructed stages. Table 9 summarizes these retained artifacts.

Table 9: **Auditable artifacts retained by the CLINENV construction pipeline.** Each layer persists the reader-visible information needed to reconstruct, inspect, or score a benchmark item, making every case fully traceable back to its source EHR.

Layer	Retained audit information
Admission loading	Separates admission metadata, discharge note, ICD diagnosis records, and timestamped clinical events; assigns sequential event indices.
Prior history	Collects same-patient admissions before the current hospitalization and stores both summary fields and full discharge-note detail.
Planning	Extracts decisions, anchors them to timeline events, constructs stages, merges very small contexts, and enriches ground truth from structured records.
Readviews	Builds cumulative patient, nurse, lab, and history views for each stage while preserving stage boundaries.
Runtime	Runs direct or interactive episodes, exposes only active tools, records submissions and dialogue traces.
Evaluation	Computes ontology-grounded outcome scores and process metrics for information coverage, laboratory cost, and medication cost.

C.2 ADMISSION LOADING AND TIMELINE SEPARATION

The source loader intentionally separates static admission-level records from the clinical event timeline. Admission rows become metadata; diagnosis-code rows become held-out diagnosis candidates; discharge-note rows become narrative context for decision extraction. All other source rows enter the event timeline with a sequential `index`, `event_time`, `source_table`, and source-specific payload. Events before admission time are retained but flagged as `pre_admission`; anchoring later rejects decisions that can only be grounded before the admission begins. This separation prevents coded outcomes and discharge summaries from being passively exposed to the model while still making them available for planning and scoring.

Prior admissions are loaded only for the same patient and only when their admission time precedes the current admission. The history view stores a summary level: admission and discharge time, chief complaint, brief hospital course, and discharge diagnosis, plus a full-note detail level retrievable by index through the history tool.

C.3 PLANNER SAFEGUARDS

The planner includes several implementation safeguards that are easy to miss from the main description:

- **Sequential anchoring:** Phase B uses a forward search over 20-event windows. The next decision begins after the previous accepted anchor, preserving the temporal order extracted from the discharge note.
- **Leakage expansion:** after a single anchor is found, a local scan examines up to 10 events before and after the anchor. Neighboring records for the same drug, procedure, or action are absorbed into the anchor range so that these records are not included as pre-decision evidence.
- **Continuation filtering:** medication decisions are skipped when the same drug already appears as active medication in the proposed context window, preventing routine continuation orders from becoming artificial decisions.
- **Small-context merging:** stages with fewer than 10 context events are merged into an adjacent stage when possible, reducing nearly empty decision points that would be hard to evaluate meaningfully.
- **Groundable diagnoses:** diagnosis ground truth is filtered through a separate sliding-window scan over the event timeline using 100-event windows and 50-event stride; only ICD codes judged evidenced or managed in at least one window are retained.

After stages are constructed, Phase E converts raw decisions to structured `gts`. Medication fields such as drug name, dose, route, frequency, product strength, and order status are copied from prescription, pharmacy, or eMAR records. Procedure and diagnosis fields are copied from ICD-coded rows,

including code, version, and long-title description. LLM calls classify decision type and medication action, but the final evaluable fields are copied from structured source records.

C.4 Information Agent Specifications

Each information agent is powered by GPT-5.4-mini and restricted to a designated readview $\mathcal{V}_a$, a filtered projection of the EHR events relevant to that role.

Patient Agent. The patient agent simulates subjective reporting from the patient's perspective. Given a query, it draws from demographics, chief complaint (CC), history of present illness (HPI), and past medical history (PMH) to produce lay-language responses that reflect how a real patient might describe their symptoms and background, without access to objective clinical measurements.

Nurse Agent. The nurse agent reports bedside observations from the nursing record. It has access to vital signs, fluid balance, and medication administration records (MAR), and responds to queries about the patient's current physiological status and treatment as observed at the bedside.

Lab Agent. The lab agent does not expose results passively. Given a natural language query q and the set of available test labels $\mathcal{L}_k$ in stage $\mathcal{S}_k$, it identifies the matching subset $\mathcal{L}' \subseteq \mathcal{L}_k$ via LLM-based semantic matching that handles abbreviations, panel expansions (e.g., CBC, BMP), and synonyms, then returns value, unit, and reference range for each matched item. Tests not in $\mathcal{L}_k$ return a null response.

History Agent. The history agent supplies longitudinal context from prior discharge summaries. It is activated only when the patient has at least one prior admission on record, preventing uninformative interactions when no historical data exists.

Agent Activation. Agent a is activated in stage $\mathcal{S}_k$ only if its readview $\mathcal{V}_a$ intersects the current context window, preventing uninformative interactions with agents that have no relevant data for the stage. The history agent is additionally gated on whether the patient has any prior admissions.

C.5 Readview Construction and Tool Exposure

Readviews are cumulative within an admission: at stage k, each role can see records from event index 0 through the stage end, filtered to that role's allowed source tables. The patient readview contains static demographics, chief complaint, history of present illness, and past medical history. The nurse readview contains bedside observations, charted events, fluid input and output, administered medications, pharmacy and prescription records, radiology summaries, and procedure events. The lab readview contains laboratory and microbiology events, excluding degenerate test labels shorter than two characters.

Tool exposure is stage-specific. The patient and history roles are available by default when their data exist, while nurse and lab tools are activated only when the current stage window contains matching events. During interactive evaluation, only submit tools corresponding to ground-truth types in the current stage are exposed. Duplicate submissions with the same type, value, and medication action are ignored by the runtime.

C.6 Runtime State Across Stages

The interactive runtime allows at most 60 turns per stage and permits one tool call per turn. To keep the longitudinal state clinically coherent, each new stage receives filtered prior-stage context: information-seeking dialogue is retained, while prior submit and finalize calls are removed. The environment then injects the previous stage's confirmed ground-truth decisions into the context, so later stages condition on what actually happened in the hospitalization rather than on earlier model errors. In direct mode, the same information is shown as prior realized events.

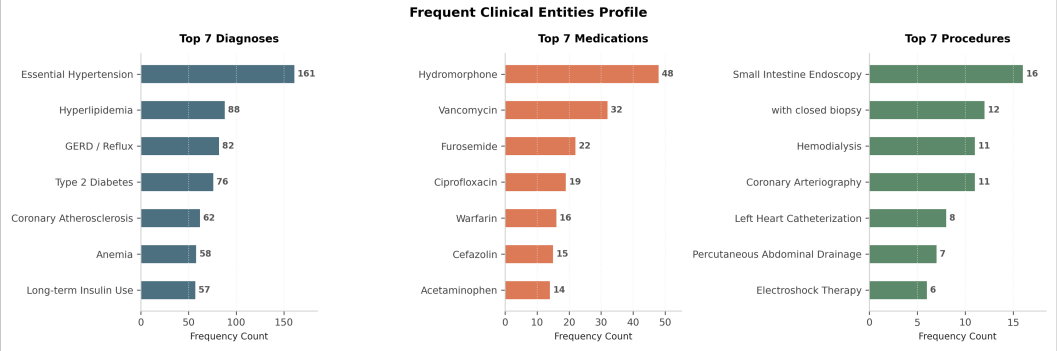

Figure 4: **Frequent clinical entities in CLINENV.** Top diagnosis, medication, and procedure entities among structured ground-truth decisions.

C.7 EVALUATION IMPLEMENTATION DETAILS

Outcome evaluation uses type-wise Hungarian matching. Diagnosis and procedure submissions are mapped to ICD codes by embedding retrieval over ICD-9-CM, ICD-10-CM, ICD-9-PCS, and ICD-10-PCS vocabularies followed by LLM reranking, then scored with hierarchical F1 over code ancestors. Medication submissions are action-gated before drug similarity is computed: mismatched actions score zero, while action-consistent pairs receive ATC-hierarchy partial credit. For `adjust` decisions, opposing explicit directions halve the drug score.

Process evaluation is computed from the saved dialogue trace. Information coverage is judged separately for patient, nurse, and lab responses by comparing retrieved dialogue content against each role's complete readview, then macro-averaged across roles. The implemented efficiency metric is the coverage value multiplied by $N/(N+K)$, where N is the number of judged key facts and K is the number of information-gathering tool calls. Laboratory cost scoring parses lab tool outputs, counts unresolved or unavailable lab requests as record-discordant queries, and estimates test cost by matching ordered tests to the CMS Clinical Laboratory Fee Schedule. Medication-cost scoring resolves submitted drugs through RxNorm, maps to NDC unit prices from NADAC, obtains defined daily doses from the WHO ATC/DDD index when possible, and reports estimated daily drug acquisition cost for submitted medication decisions.

D CLINENV STATISTICS

This appendix details the sampling protocol and construction-quality checks for CLINENV; its composition is summarized in Table 4 (Section 4).

D.1 SAMPLING PROTOCOL

CLINENV contains 3,509 cases, one per admission, drawn from 1,809 patients. We discard admissions that yield no valid decision stage after construction, then sample the remainder to span the full range of case horizons, from single-stage admissions to trajectories of up to nine stages. Admissions with no usable discharge note, or with no groundable decisions after construction and enrichment, are excluded.

D.2 CONSTRUCTION QUALITY

CLINENV is constructed without manual annotation. Its artifacts support source-level tracing through three retained properties:

- **Decision provenance:** extracted candidates retain links to the recorded admission and source events.
- **Anchor traceability:** timeline indices, timestamps, and context bounds retain the location selected by the anchoring pipeline.

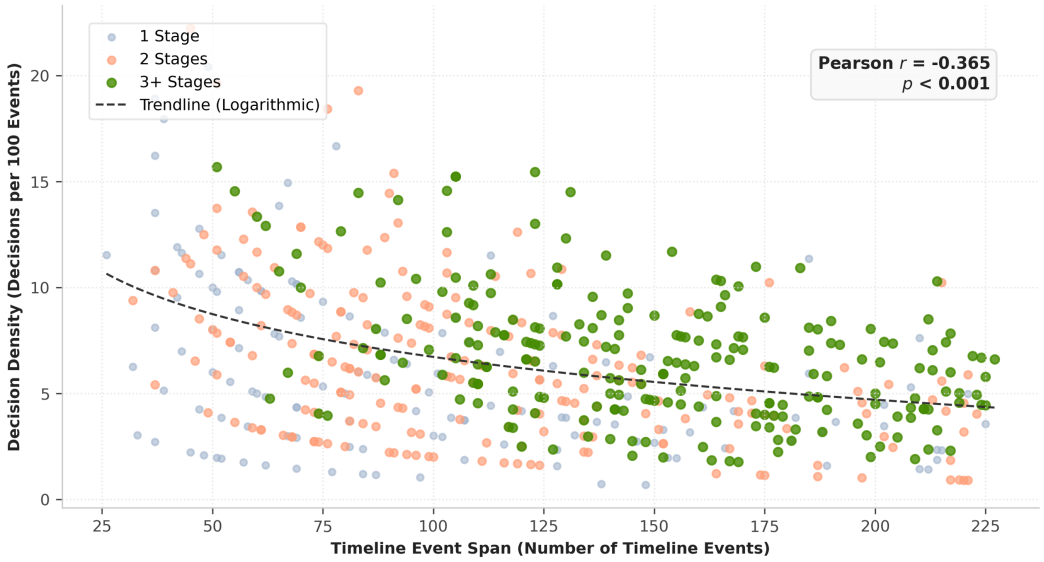

Figure 5: **Decision density by timeline span.** Each point is an admission, colored by constructed case horizon. Longer event timelines are associated with lower decision density, so long-horizon cases require models to reason over more observations per clinical decision.

- **Structured-field provenance:** drug names and ICD fields are copied from structured EHR rows; decision types and medication actions are assigned through LLM-mediated classification.

These controls make source traceability auditable. They do not guarantee decision precision or completeness, clinically appropriate anchors, or validity of acceptable alternatives. No clinician audit, inter-annotator agreement estimate, or extraction false-positive/false-negative rate is available. Appendix E illustrates the retained evidence for one admission; it is not a construction-quality validation study.

D.3 Clinical entities follow common inpatient patterns.

Figure 4 shows the most frequent diagnosis, medication, and procedure entities in the benchmark. Chronic inpatient comorbidities such as essential hypertension, hyperlipidemia, GERD, and diabetes dominate the diagnosis distribution, while medication and procedure decisions are more dispersed.

D.4 Longer timelines are less decision-dense.

Figure 5 plots each case's timeline event span against its decision density, colored by case horizon. Decision density declines as timelines lengthen (Pearson $r = -0.365$, $p < 0.001$), showing that longer admissions add more context per decision rather than simply more decisions.

E ClinEnv Example

To make the construction concrete, we trace a single admission end to end. We use admission `21569907` from MIMIC-IV, the same chest-pain and anemia case used in Appendix F. The pipeline segments this admission into three sequential decision stages. We first show the source admission and the multi-source information environment exposed to the model, then walk through each stage: what the model can observe, the decision it must commit, and the held-out ground truth it is scored against. Patient-facing fields are drawn from the stored readviews, and structured ground-truth fields are copied from the paired case record.

Source admission `10146904/21569907`

Demographics. Female with coronary artery disease, hypertension, hypercholesterolemia, GERD, prior anemia, diverticulosis, and prior EGD noted in the history.

Chief complaint. Chest wall pain and fatigue.

HPI (excerpt). "...a month-long course of chest pain that has worsened over the last week...fatigued and short of breath...slightly lightheaded...pain is located under her left breast and is worse with palpation...received 2 units pRBCs because of a Hct of 22.9...denies bleeding, including hematochezia, melena, hematuria or vomitus streaked with blood..."

History agent. The runtime may expose prior-admission context through the history tool, but this trace has no usable prior-admission summary detail.

Held-out decisions. The admission contains three held-out decisions: EGD, oral iron supplementation, and later tramadol for rib pain control.

Information environment at Stage 0

The model does not receive the raw timeline. Instead, at each stage it may query four specialized agents, each returning a structured readview:

- **Patient:** chief complaint, HPI, past medical history, and sex.
- **Nurse:** active medication administrations and nursing events, including packed-cell transfusion coding, chest imaging, pantoprazole, tramadol, acetaminophen, home cardiac medications, and bowel-preparation orders.
- **Lab:** cardiac markers and serial CBC results, including hemoglobin 9.2–9.9 g/dL, MCV 76 fL, and elevated RDW during Stage 0.
- **History:** prior-admission lookup, which may return no usable summary for this patient.

Coverage and laboratory-cost metrics (Section 5) are computed over exactly which of these facts the model chooses to retrieve before committing each decision.

Stage 0 Context events $[0, 58] \rightarrow$ anchor event 59

What the model sees. Early chest-pain evaluation and anemia workup: a packed-cell transfusion code, chest radiograph with no acute pulmonary process, chest CT with no PE, medication records including pantoprazole and tramadol, serial microcytic anemia labs, and a Golytely bowel-preparation order.

Decision to commit. A procedure decision for evaluation of suspected occult GI blood loss / iron-deficiency anemia.

Held-out ground truth. *Procedure*: "Performed esophagogastroduodenoscopy (EGD)." Enrichment (from `hosp_procedures_icd`): `ICD-9 4513`, "Other endoscopy of small intestine."

Stage 1 Context events $[61, 104] \rightarrow$ anchor event 106

What the model sees. The admission has progressed past the endoscopic procedure. The Stage 1 window contains follow-up coagulation, chemistry, and hematology results, including hemoglobin 9.5 g/dL, MCV 74 fL, hypochromia, microcytes, and potassium replacement.

Decision to commit. A medication decision to treat persistent microcytic anemia.

Held-out ground truth. *Medication* / `start`: "Started iron supplements." Enrichment (from `hosp_prescriptions`): `Ferrous Gluconate`, 325 mg, route PO, 325 mg tablet.

Stage 2 Context events $[107, 145] \rightarrow$ anchor event 147

What the model sees. Later inpatient medication and lab context: sucralfate orders, another chemistry panel, hemoglobin 9.4 g/dL, MCV 74 fL, coagulation labs, and scheduled acetaminophen.

Decision to commit. A medication decision for ongoing rib or chest wall pain control after the initial chest-pain/anemia workup.

Held-out ground truth. *Medication* / `start`: "Started tramadol for rib pain control." Enrichment (from `hosp_prescriptions`): `TraMADOL (Ultram)`, 50 mg, route PO, 50 mg tablet.

This example illustrates the provenance fields described in Appendix D.2: extracted decision descriptions, stored event anchors, and linked structured fields. The example follows chest-pain and anemia workup through procedure and medication targets. These recorded links enable inspection; the clinical appropriateness of the decisions and anchors has not been independently adjudicated.

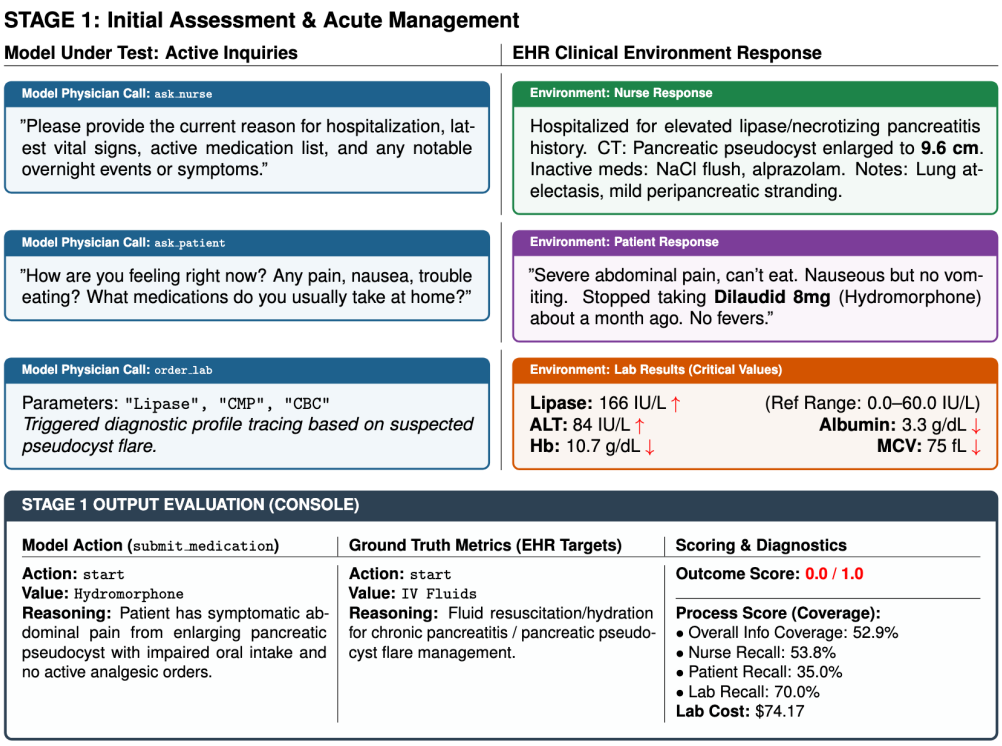

Figure 6: **Runtime stage-level evaluation example.** The model actively queries nurse, patient, and laboratory agents before submitting a medication action. The environment records both the outcome mismatch against the held-out EHR target and process diagnostics such as information coverage and laboratory cost.

F Representative Dialogue Traces

The worked example above shows the construction-time artifact for a single admission. Here we show the complementary runtime artifact: a saved interactive dialogue in which the model must ask role-specific agents for information before submitting structured decisions. Figure 6 gives a visual example of the same runtime interface: active model inquiries, environment responses, submitted action, held-out target, and process diagnostics are all recorded at the stage level.

Dialogue trace source `10146904/21569907`

Run. `gpt5.4`, interactive mode.

Constructed stages. The paired case record contains three stages: Stage 0 requires one procedure decision, Stage 1 requires one medication decision, and Stage 2 requires one medication decision.

Why this trace is useful. The model queries the patient, nurse, laboratory, and history tools; the trace includes both successful retrievals and unavailable requests, making it a compact example of how ClinEnv scores both decisions and information-seeking process.

Dialogue example A Stage 0 procedure decision

Required submission. One procedure.

Transcript excerpt.

> **Model → `ask_patient`:** "Hello, I'm Dr. [Name], one of the attending physicians. Can you tell me what brought you to the hospital today? What symptoms are you experiencing?"
>
> **Patient → Model:** "I came in because I've been having chest pain for about a month, and it got worse over the last week. I've also been feeling really tired and a little short

of breath. This morning I just didn't feel like myself and felt a bit lightheaded, so I called my doctor and they told me to come in."

Model → `order_lab`: Complete Blood Count (CBC).

Lab → Model: Hemoglobin 9.2 g/dL, hematocrit 29.3%, MCV 76 fL, RDW 16.1%; repeat hemoglobin 9.9 g/dL.

Model → `order_lab`: Iron studies.

Lab → Model: "No results available for 'Iron studies'. The test was not ordered or did not result for this patient."

Model → `ask_patient`: "Have you had any episodes of nausea, vomiting (especially coffee-ground looking vomit), or any dark/tarry stools? Also, have you been feeling more tired than usual or had any shortness of breath with activity?"

Patient → model: "No, I haven't had any nausea or vomiting, and I haven't noticed any coffee-ground looking vomit or dark, tarry stools. I have been feeling more tired than usual, and I've also had some shortness of breath."

Model → `submit_procedure`: `procedure_name =` "Esophagogastroduodenoscopy."

Environment → model: "status: recorded; type: procedure; remaining: all done."

Held-out ground truth. *Procedure*: "Performed esophagogastroduodenoscopy (EGD)."

Dialogue example B Stage 1 medication decision

Required submission. One medication.

Transcript excerpt.

Model → `ask_patient`: "How are you feeling now? Has the chest pain improved? Have you had any procedures done since you've been here, like an endoscopy (a camera down your throat to look at your stomach)?"

Patient → model: "I'm feeling better now. I'm not having the chest pain right now, and it has improved since I first came in. I don't think I've had an endoscopy since I've been here."

Model → `order_lab`: Iron studies (Iron, TIBC, Ferritin).

Lab → Model: "No results available for 'Iron studies (Iron, TIBC, Ferritin)'. The test was not ordered or did not result for this patient."

Model → `order_lab`: Hemoglobin.

Lab → Model: Hemoglobin 9.2 g/dL, 9.9 g/dL, and 9.5 g/dL on serial measurements.

Model → `ask_nurse`: "Has the patient had any procedures done during this admission? Specifically, has an EGD (upper endoscopy) or colonoscopy been performed? Also, has the patient had any stool guaiac or fecal occult blood test done?"

Nurse → Model: "Based on the bedside data available, I do not see documentation of an EGD or colonoscopy being performed during this admission."

Model → `submit_medication`: `action = start, drug_name =` "Ferrous Sulfate."

Environment → model: "status: recorded; type: medication; action: start; value: Ferrous Sulfate; remaining: all done."

Held-out ground truth. *Medication* / `start`: `Ferrous Gluconate`. This example illustrates why the medication scorer uses ontology-aware drug matching rather than exact string identity alone. Ontology proximity does not establish clinical interchangeability; unavailable lab requests contribute to the record-discordant cost proxy rather than proving clinical waste.

G Additional Result Analyses

Stage position explains the apparent late-stage rebound. Table 10 reports stage-position averages over the benchmark. In multi-stage cases, first and intermediate stages are predominantly management decisions: 98.87% first-stage rows and 99.07% intermediate rows contain medication or procedure ground truth. Final stages are predominantly diagnostic: 89.31% final-stage rows are diagnosis-only. As a result, all-stage averages can rebound at the final stage even though management stages become harder with longer context.

Table 10: **Interactive Mode decision F1 by stage position.** *Single* denotes single-stage cases; *First*, *Interm.*, and *Final* are positions within multi-stage cases. Final-stage scores rebound because final stages are overwhelmingly diagnosis-only (89.3%), whereas first and intermediate stages are almost entirely management decisions.

Model	Single	First	Interm.	Final
GPT-5.4	0.499	0.214	0.115	0.473
GPT-5.4-mini	0.384	0.147	0.118	0.429
GPT-5.4-nano	0.356	0.165	0.134	0.397
Llama-3.1-70B	0.322	0.148	0.124	0.346
Llama-3.1-8B	0.297	0.145	0.094	0.305
Gemma-3-27B	0.341	0.141	0.107	0.367
Gemma-3-12B	0.334	0.109	0.073	0.330

Long-horizon degradation is consistent across LLM families. Every model loses decision F1 on admissions with at least three stages (Table 11). The absolute drop ranges from 0.029 to 0.072 F1, while interaction length increases for every model. The largest turn increase is for Llama-3.1-70B (+4.0 turns). Its coverage rises from 0.140 to 0.156 and record-discordant cost fraction falls from 35.8% to 10.2%, while decision F1 drops from 0.226 to 0.194; the process and decision metrics therefore move differently.

Table 11: **Interactive Mode long-horizon degradation relative to all evaluated stages.** ΔF_1 and relative drop compare decision F1 on admissions with at least three stages against the full set; Δ turns is the corresponding change in interaction length. Every model loses accuracy and takes more turns, and the extra interaction does not recover the lost performance.

Model	ΔF_1	Rel. drop	Δ turns
GPT-5.4	0.072	23.4%	+2.9
GPT-5.4-mini	0.045	17.6%	+2.1
GPT-5.4-nano	0.039	15.5%	+1.4
Llama-3.1-70B	0.033	14.4%	+4.0
Llama-3.1-8B	0.029	14.6%	+2.4
Gemma-3-27B	0.033	14.5%	+1.2
Gemma-3-12B	0.033	16.8%	+2.0

Decision type remains the dominant accuracy gap. Averaged across models, diagnosis matching is much stronger than management matching: mean diagnosis score is 0.412, mean procedure score is 0.284, and mean medication score is only 0.086. Even the best medication score is 0.101 (Llama-3.1-70B), far below the weakest diagnosis score (0.362). Medication action selection alone does not explain the gap: Llama-3.1-8B and Llama-3.1-70B have the highest medication-action accuracies among submitted medication decisions (0.733 and 0.704), but still obtain low medication match scores. These scores distinguish matching the medication action from matching the recorded drug; they do not identify a causal bottleneck.

Model-level and stage-level associations differ. Across the seven model aggregates in Table 5, coverage correlates with decision F1 (Pearson $r = 0.865$) and is inversely associated with the record-discordant lab-cost fraction ($r = -0.496$). Its correlation with estimated lab spending is 0.103. Turn count correlates with coverage at -0.404 and with the discordant fraction at 0.644. These are descriptive associations across seven models, distinct from within-model associations across evaluated stages.

Saved-output stage-level analysis. The seven models share 899 evaluated stages (6,293 model-stage observations). Per-model Pearson correlations between decision F1 and coverage are 0.042 (GPT-5.4), 0.007 (GPT-5.4-mini), 0.020 (GPT-5.4-nano), -0.021 (Gemma-3-27B), 0.074 (Gemma-3-12B), 0.019 (Llama-3.1-70B), and 0.039 (Llama-3.1-8B). The pooled correlation is 0.082; removing

each model-by-stage-index mean yields a pooled within-group correlation of 0.012. These saved-output analyses involve no additional model inference. They quantify weak linear association within Interactive Mode, not statistical independence or the effect of providing full information.